Robot Transparency, Trust and Utility

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

Abstract. As robot reasoning becomes more complex, debugging becomes increasingly hard based solely on observable behaviour, even for robot designers and technical specialists. Similarly, non-specialist users find it hard to create useful mental models of robot reasoning solely from observed behaviour. The EPSRC Principles of Robotics mandate that our artefacts should be transparent, but what does this mean in practice, and how does transparency affect both trust and utility? We investigate this relationship in the literature and find it to be complex, particularly in non industrial environments where transparency may have a wider range of effects on trust and utility depending on the application and purpose of the robot. We outline our programme of research to support our assertion that it is nevertheless possible to create transparent agents that are emotionally engaging despite having a transparent machine nature.

1 INTRODUCTION

The EPSRC Principles of Robotics includes a specific reference to transparency: "Robots are manufactured artefacts. They should not be designed in a deceptive way to exploit vulnerable users; instead their machine nature should be transparent." see [1]. This initially appears to be a straightforward normative assertion, drawing on the commonly held idea that agents should not be deceptive, since deception generally leads to exploitation. This paper considers whether in fact transparency is really such a simple idea, and also whether making certain types of agents transparent reduces their utility. In considering this question, we must also address the relationship between transparency and trust.

In this paper, we use the terms robot and agent interchangeably and by these terms we mean an embodied, autonomous intelligent artefact.

What does it mean to trust a robot? We might initially simply assert that if an AI is more transparent, then we are able to trust it more, and therefore its utility increases. We could also argue that trust is only required when an agent is not fully transparent, and therefore that increased transparency reduces the need for trust [4]. If the utility of an artefact is measured by the degree to which it is trusted, then increasing transparency may reduce that utility. This might, for example, be the case for a robot that's primary function is to provide companionship.

So, we start to see that there is a complex relationship between the ideas of utility, transparency and trust. This relationship will depend on the purpose of the AI. In this paper we review the literature relating to transparency and trust, and we also describe ongoing practical research to investigate the proposal that it is indeed possible build an emotionally engaging yet transparent robot.

2 THEORY OF MIND, TRUST AND TRANSPARENCY

Although we may presuppose that communication between animals, and particularly between humans must be complex, in fact natural communication systems tend to exploit relatively simple and minimal signals, the meaning of which derives from extensive models [16]. In other words, evolution, or a shared phylogenetic history, provides adequate priors such that minimal data is required to communicate context. Although some would argue otherwise [8], it is generally agreed that effective interaction, whether coercion or cooperation, relies on each party having some theory-of-mind (ToM) of the other [16, 14]. Individual actions and composite behaviours are thus interpreted within a pre-existing ToM framework. Whether that ToM is accurate is unimportant, provided that it is predictive in terms of behaviour. The robot's transparency model does not define the ToM employed by the human user, but it is the transparency model that we can directly adjust and this is therefore the focus of this paper. It is well known that observable behaviour can communicate the internal mental states of the individual. Breazeal [2] found that implicit non-verbal communication improves transparency over that of only deliberate non-verbal communication. Here implicit is defined as conveying information inherent in behaviour but which is not deliberately communicated by the robot designer. People have strong expectations for how implicit and explicit non-verbal cues map to mental states. Breazeal also found that transparency reduces conflict when errors occur, particularly when a joint task is being attempted. Reduced conflict implies that when an error occurs during task execution, recovery is still possible with less apportionment of blame. Breazeal terms this reduced conflict Robustness, and this robustness is one effective measure of utility.

2.1 Anthropomorphism and Mental Models of Robots

Humans have a strong predisposition to anthropomorphise not only nature, but anything around them [5] — the Social Brain Hypothesis [7] may explain this phenomenon, however humans do not treat robots identically to humans, for example with respect to moral standing [10]. Although there is significant debate about the ontology of robot minds versus human minds, what is of more practical importance is how robot minds are understood psychologically by humans, i.e. what is the perceived, rather than actual, ontology. Stubbs [15] considers it essential to form a mental model of robots in order to build common ground — which we might also interpret as the basis for human trust. Stubbs [15] also found that this common ground can be effectively established via an interactive dialogue with the robot. Although this study primarily considered remote robots working in an industrial or exploratory setting, rather than robots operating in

[1] University of Bath, UK, email: r.h.wortham@bath.ac.uk
[2] University of Bath, UK, email: a.theodorou@bath.ac.uk
[3] University of Bath, UK, email: j.j.bryson@bath.ac.uk

domestic environments, we should take note of the importance of dialogue in establishing trust. Indeed Mueller [13] sees dialogue as one of the three main characteristics of transparent computers, the others being explanation and learning.

Meerbeek [12] investigates the relationship between a robot's perceived personality and the level to which the user feels in control during the interaction. In order to be believable, Meerbeek found that the personality expression should be linked to an internal model that deals with the behaviour (e.g. decision making) based on personality and emotion. More expressive, informal behaviour is associated with a higher perception of user control.

Non-specialist humans either have little ToM for robots, or have a model based on contemporary science fiction, and therefore interpret behaviours using a default other agent theory, which assumes the agent to share human-like motivations. This can be understood in evolutionary terms through our ancestors' need to rapidly categorise proximal activity as either neutral (the rustling of leaves in the wind), friendly (the approach of a tribe member) or hostile (the approach of a predator or foe). When sensory information is uncertain, evolving a bias towards an assumption of both agency and hostility is selective for individual longevity in an environment where one is frequently the prey, not the predator. Even in our technological environments we often experience fake agency, such as robotic dialling sales calls, automated twitter postings and auto-generated personalised spam emails.

In a study conducted in 2006 in a community hospital in the USA, the nursing staff were constantly searching for reasons why the robots acted as they did. They would ask themselves and others, "What is going on here? Is the robot supposed to do this or did I do something wrong?". This research asserts that low levels of transparency led people to question even the normal behaviours of the robot, sometimes even leading people to think of correct behaviours as errors [11].

3 RESEARCH PROGRAMME

We are beginning a programme of practical research to investigate the transparency, trust, utility triangle. Initially using non-humanoid robots, we are conducting experiments to determine the effect of various expressions of transparency on the emotional response of humans. At the heart of our experiments we are using reactive planning techniques to build autonomous agents. We have developed the Instinct reactive planner based on Bryson's Behaviour Oriented Design (BOD) approach [3]. The Instinct planner 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. Our experiments will investigate and demonstrate how this transparency data from the planner can be used to make the behaviour of the robot more understandable. Initially we are primarily interested in making the behaviour transparent for the robot designer, since robots with complex plans are typically very hard to design and debug. However, these initial experiments may also improve transparency for non-specialist observers.

We will subsequently investigate how we can harness the transparency mechanism embedded with the Instinct Planner to produce a more effective domestic robot. The research will investigate whether transparency makes people feel more or less bonded to their robot, and whether they are more or less able to accurately assess the needs of the robot, as it works to achieve its goals.

It is anticipated that these trials should take place within a domestic or near-domestic environment, such as a retirement home. We must gain feedback from non-specialist observers/users about the qualitative level of intelligence of the robot, and also about how comfortable they would be to have such a device in their home environment. The research will attempt to assess initial levels of fear, anxiety, mistrust of AI and robots in general, and of domestic robots in particular. Having established a reference position, transparency of the robot must be enabled by providing feedback to the user based on the real time execution within the reactive planner. The methods we currently envisage are:

- Real-time presentation of textual statements relating to plan execution.
- Graphical real-time visualisation of plan execution.
- Audio (i.e. verbal) statements relating to robot plan execution.

For each of these methods the transparency information could either be presented on/from a remote device, or on/from the robot itself. There are thus six possible combinations. Of course additional transparency fusion, such as audio combined with graphical, could also be tested based on the success or failure of initial experimental results.

As the literature indicates that dialogue is important in establishing trust, this research should give some consideration to the possibility of accepting speech input, albeit restricted to simple commands, as a means for users to inquire of the robot what it is doing, and to have the robot respond appropriately.

4 DISCUSSION

EPSRC Principle 1 asserts that robots are tools. Within industrial and engineering environments this is fairly clear, in the sense that a human uses the robot to complete a technical task. The designer and user of the robot share the goal of the robot: to complete the task. However, within domestic and healthcare environments, robots may have rather a different relationship with those they interact with. They may be intended to provide companionship and simultaneous covert monitoring of patient well-being. They may be tools for the healthcare professional, but for the patient they are companions. In such an environment the utility may be negatively affected by increased transparency. Our sense of companionship is related to the measure of agency we project onto the robot. If we are able to understand the workings of the intelligence does it inherently appear to become less intelligent in the folk sense, such that we then project less agency, and as a result experience less benefit from the robot? We might compare this with television. We know it has no agency, but its presence in the corner of our sitting room does provide companion like benefits. Maybe this has to do with the conscious suspension of disbelief, or maybe we have an unconscious agency detector which is more easily fooled by technology.

Common-sense notions of intelligence are conflated with folk psychology ideas of agency and also of living. Things that are intelligent are alive, in the sense that they have their own beliefs, desires and intentions that we understand are fundamentally self serving, or selfish. We implicitly recognise selfishness as a fundamental characteristic of all life [6]. If such an agent engages with us then it considers us to be important in the pursuit of these selfish objectives. Such agents are worthy of becoming our companions because they ascribe true value in their relationship with us, and this increases our value in society. Conversely, agents who have no self-serving agency are not worthy of our attention because they convey no social value. Perhaps therefore, artificial agents whose sole purpose is companionship and are truly transparent in this respect are thus disqualified from being worthy companions. In some situations robot transparency may therefore be at odds with utility, and more generally it may be orthogonal rather than beneficial to the successful use of the robot. Whilst we may invent scenarios and continue to discuss the theoretical and philosophical interplay between transparency, trust and utility, as scientists we await the outcome of our experiments.

5 CONCLUSION

We have seen that unpacking transparency and trust is complex, but can be partly understood by looking at how humans come to understand and subsequently trust one another, and how they overcome evolutionary fears in order to trust other agents, through implicit non-verbal communication. Unacceptable levels of anxiety, fear and mistrust will result in an emotional and cognitive response to reject robots. Hancock [9] asserts that if we cannot trust our robots, we will not be able to benefit from them effectively. However, given that we happily interact in society with others whom we do not completely trust, and increasingly we interact with computers knowing that their recommendations maybe faulty, we must conclude that Hancock is over simplifying. Finally, there may be applications where transparency is at odds with utility. Our ongoing programme of research is intended to validate our hypothesis that we can indeed create transparent robots that are nevertheless emotionally engaging and useful tools across a wide range of domestic and near-domestic environments. Meanwhile, there remains a great of work to be done to unpack the relationship between transparency, utility and trust.

REFERENCES

[1] Margaret Boden, Joanna Bryson, Darwin Caldwell, Kerstin Dautenhahn, Lilian Edwards, Sarah Kember, Paul Newman, Vivienne Parry, Geoff Pegman, Tom Rodden, Tom Sorell, Mick Wallis, Blay Whitby, and Alan Winfield. Principles of robotics. The United Kingdom's Engineering and Physical Sciences Research Council (EPSRC), April 2011. web publication.

[2] C. Breazeal, C.D. Kidd, A.L. 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*, pp. 708–713, Alberta, Canada, (2005). Ieee.

[3] Joanna J. Bryson, 'Intelligence by design: principles of modularity and coordination for engineering complex adaptive agents', (2001).

[4] Joanna J Bryson and Paul Rauwolf, 'Trust, Communication, and Inequality'. 2016.

[5] Kerstin Dautenhahn, 'Methodology & themes of human-robot interaction: A growing research field', *International Journal of Advanced Robotic Systems*, **4**(1 SPEC. ISS.), 103–108, (2007).

[6] Richard Dawkins, 'Hierarchical organisation: A candidate principle for ethology', in *Growing Points in Ethology*, eds., P. P. G. Bateson and R. A. Hinde, 7–54, Cambridge University Press, Cambridge, (1976).

[7] R I M Dunbar, 'The Social Brain Hypothesis', *Evolutionary Anthropology*, 178–190, (1998).

[8] Shaun Gallagher, 'The narrative alternative to theory of mind', in *Radical Enactivism: Intentionality, Phenomenology, and Narrative*, ed., R Menary, number Gallagher 2001, 223–229, John Benjamins, Amsterdam, (2006).

[9] 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).

[10] Peter H. Kahn, Hiroshi Ishiguro, Batya Friedman, and Takayuki Kanda, 'What is a human? - Toward psychological benchmarks in the field of human-robot interaction', *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, **3**, 364–371, (2006).

[11] Taemie Kim and Pamela 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*, 80–85, (2006).

[12] Bernt Meerbeek, Jettie Hoonhout, Peter Bingley, and Jacques Terken, 'Investigating the relationship between the personality of a robotic TV assistant and the level of user control', *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, 404–410, (2006).

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

[14] Rebecca Saxe, Laura E Schulz, and Yuhong 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).

[15] Kristen Stubbs, Pamela J Hinds, and David Wettergreen, 'Autonomy and Common Ground in Human-Robot Interaction: A Field Study', *IEEE Intelligent Systems*, **22**(2), 42–50, (2007).

[16] Robert H Wortham and Joanna J Bryson, 'Communication', in *Handbook of Living Machines* {*in press.*}, Oxford University Press, Oxford, (2016).