Why is my robot behaving like that? Designing transparency for real time inspection of autonomous robots

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

Abstract. The EPSRC's Principles of Robotics dictates the implementation of transparency in robotic systems, however, research related to it is in its infancy. The current paper introduces the reader to the need of having transparent to inspection intelligent agents. We provide a robust definition of transparency, as a mechanism to expose the decision making of the robot, by considering and expanding upon other prominent definitions found in literature. The paper concludes by addressing potentials design decisions developers need to consider when designing transparent systems.

1 INTRODUCTION

Transparency, in our opinion, is a key element relating to the ethical implications of both developing and using Artificial Intelligence, a topic of increasingly public interest and debate. We frequently use philosophical, mathematical, and biologically inspired techniques for building artificial interactive, intelligent agents, but we treat them as black-boxes with no understanding of how the underlying real-time decision making works.

The black box nature of intelligent systems, such as in context-aware applications, makes interaction limited and often uninformative for the end user [14]. Limiting interactions may negatively effect the system's performance or even jeopardize the functionality of the system. Imagine an autonomous robotic system built for providing health-care support to the elderly. However, the elderly people may be afraid and distrust the system. They may not allow the robot to interact with them. In a such scenario human lives are at risk, as they may not get the required medical treatment in time, as a human overseeing the system must detect lack of interaction and intervene. Conversely, if the human user places too much trust in a robot, it could lead to misuse, over-reliance, and disuse of the system [13]. In our example of the health care robot, if the agent malfunctions and its patients are unaware of its failure to function, the patients may continue using the robot, risking their own health. The robots in both scenarios are breaking EPSRC's first Principle of Robotics by putting human lives at risk [1].

To avoid such situations, proper calibration of trust between the humans operators and their robots is critically important, if not essential, especially in high-risk scenarios such as the usage of robots in the military or for medical purposes [9]. Calibrating trust occurs when the end-user has a mental model of the system and relies on the system within the systems capabilities and is aware of its limitation [6].

We believe that enforcement of transparency is not only beneficial for end-users, but also for intelligent agents' developers. Real-time debugging of a robot's decision making mechanism could help developers to fix bugs, prevent issues, and explain potential variance in a robot's performancee. We envision that by the correct implementation of transparency, developers could design, test, and debug their agents in real-time — similar the way in which software developers work with traditional software development and debugging.

Despite these possible benefits of transparency in intelligent systems, there is little existing research in transparent agents and even less implementation of transparent agents. Moreover, there are inconsistencies in the definitions of transparency and the criteria for a robot to be considered a transparent system. In this paper, we will present the inconsistent definitions found in the literature and attempt to compliment them with our own. Furthermore, in the third section of this paper, we will discuss the design decisions a developer needs to consider when designing transparent robotic systems.

We specifically use the term intelligent agent to denote the combination of both the software and hardware of an autonomous robotic system, working together as an actor, living in and changing the world [3]. Within this paper the words robot and agent are used interchangeably.

2 DEFINING TRANSPARENCY

Despite the predominant usage of the keyword transparency in the EPSRC Principles of Robotics, research into making systems transparent is still in its infancy. Throughout the years, very few publications have focused on the need of transparent systems and even fewer have attempted to address this need. Each study provides its own definition of the keyword, without excluding others. To date, the transparency concept has been limited to explanations for abnormal behaviour, reliability of the system, and attempts to define the analytic foundations of an intelligent system.

2.1 The EPSRC Principle of Transparency

EPSRC's Principles of Robotics considers transparency as one of its key principles, by defining transparency in robotics as: "Robots are manufactured artefacts. They should not be designed in a deceptive way to exploit vulnerable users; instead their machine nature should be transparent." .

The EPSRC definition of transparency emphasizes keeping the end-user aware of the manufactured, mechanical, and thus artificial

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

nature of the robot. However, the phrasing used allows to consider even indirect information, such as online technical documentation, as a sufficient methodology of enforcing transparency[4]. This places the burden of responsibility with the end-user. The user will have to find, read, and understand documentation or other information provided by the manufacturer. Some user groups, such as the elderly or non-specialist users, may have issues understanding the technical terms often found in technical manuals.

2.2 Transparency as a mechanism to report reliability

One of the early publications defined transparency in terms of communicating information to the end-user, regarding the system's tendency for errors in a given context [6]. While the Dzindolet's interpretation is only a part of our definition of a transparent system, the study presents interesting findings for the importance of transparent systems. The study showed that providing extra feedback to users regarding system failures, it helped participants place their trust in the system. The users knew that the system was not 100% reliable, but they were able to calibrate their trust to the autonomous system in the experiment, as they became aware of when they could rely on it and when not to. Military usage of robotic systems is increasingly becoming more popular, especially in the form of Unmanned Aerial Vehicles (UAVs), and transparency in combat systems is essential. Imagine if an agent identifies a civilian building as a terrorist and decides to take actions against it. Who is responsible? The robot for being unreliable? Or the human overseer, who placed his trust in the system's sensors and decision making mechanism? While the EPSRC Principle of Robotics considers the human operator responsible, the damage done is irreversible. Robots working autonomously to detect and neutralize targets need to have a transparent behaviour [17]. Humans should be able to calibrate their trust to the system and in cases of combat, medical, or other scenarios where if a robot acts unreliable may harm or kill humans, transparency as a mechanism to report the system's reliability is fundamental.

2.3 Transparency as a mechanism to expose unexpected behaviour

Later studies by Kim Hinds [11] and Stumpf et. al [14], concentrated on providing feedback mechanisms to users regarding unexpected behaviour of an intelligent agent. In their studies, the user was alerted only when the agent considered that its behaviour as abnormal. Kim and Hinds' study, interestingly, showed that by increasing autonomy the importance of transparency was also increased as responsibility shifted from the user to the robot. Their results are in line with [10] research, which together demonstrate that humans are more likely to blame a robot for failures than other manufactured artefacts and coworkers.

Being able to alert the user when the robot behaves in an unexpected way is essential to achieve transparency. In high-risk situations, it could help save human lives or valuable resources by alerting a human overseer of the system to take control or calibrate its trust respectively. However, in Kim and Hinds implementation, the robot was alerting the user only when it detected that it was behaving in an unexpected way. In our opinion, this implementation tries to fix a black-box by using another. There is no guarantee that the robot is behaving unexpectedly without it knowing about its atypical behaviour. Transparency should be enforced in real-time as a always-on mechanism, allowing the user to decide if the behaviour of the agent is considered expected or unexpected.

2.4 Transparency as a mechanism to expose decision making

It is to our belief that transparency mechanisms should be built-in to the system, providing information in real time of its operation, as well as providing additional documentation as dictated by the EPSRC current principle. The intelligent agent, i.e. a robot, should contain the necessary mechanisms to provide meaningful information to the end-user. To consider a robot transparent to inspection, the end-user should have the ability to request accurate interpretations of the robot's capabilities, goals, progress in relation to the said goals, sensory inputs - situation awareness, its reliability and unexpected behaviour, such as error messages. The information provided by the robot should be presented in a human understandable format.

A transparent agent, with an inspectable decision making mechanism, could also be debugged in a similar manner to the way in which traditional, non-intelligent software is commonly debugged. The developer could see which actions the agent is making, why, and how it moves from one action to the other. This is similar to the way in which popular Integrated Development Environments (IDEs) provide options to follow different streams of code with debug points, and have abilities such as "Step-up" and "Step-in" over blocks of code.

3 DESIGNING TRANSPARENT SYSTEMS

In this section of this paper, we will discuss the various decisions developers may face in designing a transparent system. Until now, prominent research in the field of designing transparent systems focused in presenting transparency only within the context of human-robot collaboration (HRC).Thus, they focused on designing transparent systems able to build trust between the human participants and the robot.[12]. We believe that transparency should be present even in non-collaborative environments, such as human-robot competitions [11] or even when robots are used by the military. In our view, developers should strive to develop intelligent agents, which can efficiently communicate information to the human end-user, and sequentially allow her to develop a mental model of the system and its behaviour.

3.1 Usability

In order to enforce transparency, additional displays or other methods of communication to the end-user must be carefully designed, as they will be integrating potentially complex information. Agent developers need to consider both the actual relevance and level of abstraction of the information they are exposing and how they will present this information.

3.1.1 Relevance of information

Different users may react differently to the information exposed by the robot. [16] demonstrates that end-users without a technical background neither understand nor retain information from technical inputs such as sensors. This is contrary to the agent's developer, who needs access to such information during both development and testing of the robot to effectively calibrate sensors and to fix any issues found. However, within the same study, Tullio demonstrates that

users are able to understand at least basic machine learning concepts, regardless of their non-technical educational and work-history background.

Tullio's research establishes a good starting point at understanding which information maybe relevant to the user to help them understand intelligent systems. Nevertheless, further work is needed in other application areas to establish both domain-specific and user-specific trends regarding what information should be considered of importance.

3.1.2 Abstraction of information

Developers of transparent systems will need to question not only *which*, but also *how much* information they will expose to the user by establishing a level of complexity with which users may interact with the transparency-related information. This is particularly important in multi-robot systems.

Multi-robot systems allow the usage of multiple, usually small robots, where a goal is shared among various robots, each with its own sensory input, reliability, and progress towards performing its assigned task for the overall system to complete. Recent developments of biology inspired swarm intelligence allow the usage of large quantities of tiny robots working together in such a multi-robot system [15]. The military is already considering the development of swarms of autonomous little robotic soldiers. Implementing transparency in a such system is no trivial task. The developer must make rational choices about when low or high level information is required to be exposed. By exposing all information at all times, for all types of users, the system may become unusable as the user will be overloaded with information. We believe that different users will require different levels of information abstraction to avoid infobesity. Higher levels of abstractions could concentrate on presenting only an overview of the system. Instead of having the progress of a system towards a goal, by showing the current actions the system is taking in relation to achieve the said goal, it could simply present a completion bar. Moreover, in a multi-robot system, lower level information could also include the goal, sensor, goal-process, and overall behaviour of individual agents in a detailed manner. Conversely, a high-level overview could display all robots as one entity, stating averages from each machine. Intelligent agents with a design based on a cognitive architecture, such as Behaviour Oriented Design (BOD) [2], could present only high level plan elements if an overview of the system is needed. In the case of an agent designed with BOD, users may prefer to see and become informed about the states of Drives or Competencies but not individual Actions. Other users may want to see only parts of the plan in detail and other parts as a high level overview.

A good implementation of transparency should provide the user with such options, providing individuals or potential user-groups with both flexible and preset configurations in order to cater a wide range of potential users' needs. We hypothesize that the level of abstraction an individual needs is dependent on a number of factors including, but not limited to, the demographic background of the user.

1. User: We have already discussed the way in which different users tend to react differently to information regarding the current state of a robot. Similarly, we can expect that various users will respond in a similar manner to the various levels of abstraction based on their usage of the system. End-users, especially non-specialists, will prefer a high-level overview of the information available, while we expect developers to expect access to lower level of information.
2. Type of robotic system: As discussed in our examples above, a multi-robot system is most likely to require a higher level of abstraction, to avoid infobesity of the end-user. A system with a singleagent would require much less abstraction, as less data are displayed to its user.
3. Purpose of the robotic system: The intended purpose of the system should be taken into account when designing a transparent agent. For example, a military robot is much more likely to be used with a professional user in or on the loop and due to its high-risk operation, there is much greater need to display and capture as much information about the agent's behaviour as possible. On the other hand, a robotic receptionist or personal assistant is more likely to be used by non-technical users, who may prefer a simplified overview of the robot's behaviour.

3.1.3 Presentation of information

Developers needs to consider how to present to the user any of the additional information regarding the behaviour of the agent they will expose. Previous studies used visual or audio representation of the information. To our knowledge, there are no prior studies comparing the different approaches.

Autonomous robotic systems may make tens of different decisions per second. If the agent is using a reactive plan, such as a POSH plan [5], the agent may make thousands of call per minute to the different plan elements. This amount of information is hard to handle with audio-oriented systems. Moreover, visualizing the information, i.e. by providing a graphical representation of the agent's plan where the different plan elements blink as they are called, should make the system self-explanatory and easy to follow by less-technical users. Finally, a graph visualization as a means to provide transparency-related information has the additional benefits in debugging the application. The developer should be able to follow a trace of the different plan elements called, viewing the sensory input that triggered them, until a specific elements was used.

3.2 Utility of the system

So far in this paper we have expanded on the importance and design choices regarding the implementation of transparency. However, we believe the developer also needs to consider whether implementing transparency may actually damage the utility of a system. [18] argues that the the utility of an agent is measured by the degree to which it is trusted. Increasing transparency may reduce its utility. This might, for example, have a negative effect for a companionship robot or a health-care robot, designed to assist children. In such cases, the system is designed against the EPSRC Principles of Robotics, as it exploits its users feelings to increase its utility and performance on its set task.

Another important design decision which effects the system is the physical transparency of the system. The physical appearance of an agent may increase its usability [7], but also it may contrast with transparency by hiding its mechanical nature. Back in our companionship robot example, a humanoid or animal-like robot may be preferred over an agent where its mechanisms and internals are exposed, revealing its manufactured nature [8].

Discussing the trade-offs between utility and transparency is far beyond the scope of this paper. However, developers should be aware of this as they design and develop robots.

4 CONCLUSION

We strongly believe that the implementation and usage of intelligent systems which are transparent in nature can help the public understanding of AI by removing the scary mystery around why is it behaving like that. Transparency will allow to understand an agents emergent behaviour. In this paper we re-defined transparency as an always-on mechanism able to report a system's behaviour, reliability, senses, and goals as such information could help us understand the autonomous system's behaviour.

Further work is needed to test and establish good practices regarding the implementation of transparency within the robotics community. Considering the benefits of transparent systems, we strongly suggest the promotion of this key principle by research councils, such as EPSRC, and other academic communities.

ACKNOWLEDGEMENTS

We would like to thank Swen Gaudl (University of Bath) for his most valuable insights.

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] Joanna Bryson, 'The behavior-oriented design of modular agent intelligence', in *System*, volume 2592, 61–76, (2002).

[3] Joanna J. Bryson, 'Robots should be slaves', in *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues*, ed., Yorick Wilks, 63–74, John Benjamins, Amsterdam, (March 2010).

[4] Joanna J Bryson, Darwin Caldwell, Kerstin Dautenhahn, Paula Duxbury, Lilian Edwards, Hazel Grian, Sarah Kember, Stephen Kemp, Paul Newman, Geo Peg, Andrew Rose, Tom Rodden, Tom Sorell, Mick Wallis, Shearer West, Alan Winfield, and Ian Baldwin, 'The making of the epsrc principles of robotics', **133**(133), 14–15, (2012).

[5] Joanna J. Bryson, Tristan J. Caulfield, and Jan Drugowitsch, 'Integrating life-like action selection into cycle-based agent simulation environments', in *Proceedings of Agent 2005: Generative Social Processes, Models, and Mechanisms*, eds., Michael North, David L. Sallach, and Charles Macal, pp. 67–81, Chicago, (October 2005). Argonne National Laboratory.

[6] Mary T. Dzindolet, Scott A. Peterson, Regina A. Pomranky, Linda G. Pierce, and Hall P. Beck, 'The role of trust in automation reliance', *International Journal of Human Computer Studies*, **58**(6), 697–718, (2003).

[7] Kerstin Fischer, 'How people talk with robots: Designing dialogue to reduce user uncertainty', *AI Magazine*, **32**(4), 31–38, (2011).

[8] Jennifer Goetz, Sara Kiesler, and Aaron Powers, 'Matching robot appearance and behavior to tasks to improve human-robot cooperation', *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, 55–60, (2003).

[9] Victoria Groom and Clifford Nass, 'Can robots be teammates?', *Interaction Studies*, **8**(3), 483–500, (2007).

[10] Peter H. Kahn, Rachel L. Severson, Takayuki Kanda, Hiroshi Ishiguro, Brian T. Gill, Jolina H. Ruckert, Solace Shen, Heather E. Gary, Aimee L. Reichert, and Nathan G. Freier, 'Do people hold a humanoid robot morally accountable for the harm it causes?', *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction - HRI '12*, (February 2016), 33, (2012).

[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] Joseph B Lyons, 'Being transparent about transparency : A model for human-robot interaction', *Trust and Autonomous Systems: Papers from the 2013 AAAI Spring Symposium*, 48–53, (2013).

[13] R Parasuraman and V Riley, 'Humans and automation: Use, misuse, disuse, abuse', *Human Factors*, **39**(2), 230–253, (1997).

[14] Simone Stumpf, Weng-keen Wong, Margaret Burnett, and Todd Kulesza, 'Making intelligent systems understandable and controllable by end users', 10–11, (2010).

[15] Ying Tan and Zhong-yang Zheng, 'Research advance in swarm robotics', *Defence Technology*, **9**(1), 18–39, (3 2013).

[16] Joe Tullio, Anind K. Dey, Jason Chalecki, and James Fogarty, 'How it works: a field study of non-technical users interacting with an intelligent system', *SIGCHI conference on Human factors in computing systems (CHI'07)*, 31–40, (2009).

[17] Lu Wang, Greg a Jamieson, and Justin G Hollands, 'Trust and reliance on an automated combat identification system', *Human factors*, **51**(3), 281–291, (2009).

[18] Robert Wortham, Andreas Theodorou, and Joanna J. Bryson, 'The iron triangle: Transparency-trust-utility'. submitted, 2016.