Challenges of Transparency for Learning Robots

Mattia Racca
Aalto University
Espoo, Finland
mattia.racca@aalto.fi

Ville Kyrki
Aalto University
Espoo, Finland
ville.kyrki@aalto.fi

ABSTRACT

With the importance of transparency for autonomous intelligent systems (AISs) becoming clear to the robotics community, definitions of transparency and guidelines for its implementation have started being proposed. In this paper, we adopt from the literature a model characterizing different aspects of transparency in the light of human-robot interaction and apply it to the specific case of robots learning from humans. Requirements deriving by the specific nature of learning systems are remarked and the main challenges highlighted and discussed.

KEYWORDS

Autonomous Intelligent Systems; Robot Learning; Transparency

ACM Reference Format:
Mattia Racca and Ville Kyrki. 2018. Challenges of Transparency for Learning Robots. In *Proceedings of Explainable Robotic Systems, Workshop in conjunction with HRI 2018, Chicago, IL, USA, March 5th, 2018 (Explainable Robotic Systems Workshop, HRI '18),* 2 pages.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

With service and collaborative robots becoming increasingly common, the need for explainable and transparent systems has become clear to the robotics community. The transparency of an Autonomous Intelligent System (AIS) can be defined as the sharing of accurate perceptions of its abilities, intentions and constraints with the end users [7]. Without appropriated transparency mechanisms, AISs can be misunderstood, misused or distrusted by their final users [10].

Efforts have been done in the community to provide guidelines for characterizing transparency for AISs and to design such transparency mechanisms. Lyons proposed a model of transparency for human-robot interaction, organizing the different aspects of transparency in 6 categories [7]. Theodorou et al. presented guidelines for designing transparent systems, discussing *what*, *how* and *to which degree* a transparent system needs to expose its reliability and decision making [10].

We are interested in the particular case of systems that learn from their end users [4]. Robots have potentially a wide range of capabilities and allowing end users to directly teach their robots is the most direct way to achieve the desired level of customization.

The appeal of learning systems (LS) comes however with more requirements in their design. First, LSs are evolving over time, accordingly to how their training proceeds. Second, LSs may need to operate in an suboptimal way during the training (i.e. make mistakes while learning). Third, they have two (separated or interleaved) modes of operation: training and deployment.

As a subset of AISs, LSs can also benefit from implementing transparency mechanisms. Using the categories presented in [7], we present how transparency can be characterized for the special case of LSs during their training, highlighting challenges and providing, when possible, pointers to current research addressing them.

2 TRANSPARENCY FOR LEARNING SYSTEMS

In his model of transparency for HRI [7], Lyons presents 6 categories of transparency, separated in two macro categories: *Robot-to-human* factors and *Robot-of-human* factors. The former group represents concepts of transparency that aim to expose robot related information to the user; the latter group exposes instead user related information available to the robot.

Robot-to-human factors

Intentional Model. This models answers the questions *why the robot exists?* and *what is its intended purpose?* [7]. For LSs, the trivial answer is *the robot purpose is to learn.* However, LSs are not learning for the sake of learning; the training phase should produce skills that satisfy the needs of the user and fulfil the imposed requirements. It has to be clear that the goal of the training has to come from the user. The definition of the goal has to however take in consideration the limitation of the system and its intended purpose *after the training*. Being clear about the physical and computational capabilities of a system can prevent mismatches with user expectations. As LSs are meant to obtain new skills by training, the definition of an intended purpose is not trivial yet needed to prevent misuse of the system. Ideally, the intended purpose should be defined by the robot's manufacturer. The training should then be accordingly constrained to forbid eventual novel learned skills conflicting with the original intended use.

Task Model. This models includes understanding of a particular task of the robot, its current goal (and progress in achieving it), awareness of its capabilities and limitations. From the perspective of LSs, exposing information like the current learning objective or the pace of the learning can help the human teacher to understand how the training is proceeding to provide a stopping condition (similarly to software progress bars). If the progress of learning is not easily computable or cannot be conceptualized in a user friendly way, the user should be allowed to test the LS in the same way a teacher

Explainable Robotic Systems Workshop, HRI '18, March 5th, 2018, Chicago, IL, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

assesses the proficiency of her pupils with homework. Testing a partially trained system is however challenging, as safety issues can arise. Another skill LSs can benefit from is the ability to explain failures in the light of the training, implementing some kind of experience-based debugging [10]. Steps in this direction have been presented in [5], where the authors proposed a suite of algorithm to automatically generate explanations of robot behaviours starting from a known policy.

Analytical Model. The aim of this model is to expose the decision making of an AIS, a point stressed also in [10], especially when there is no clear best plan to be followed due to e.g. environmental uncertainties. Regarding LSs, the aim becomes to expose the reasoning guiding the learning process, like e.g. the reward maximization for a Reinforcement Learning agent or the query selection for an Active Learning system. The main challenge here is to make machine learning concepts graspable for layman users. In [8], in the context of robot active learning, we grounded probability concepts for non-expert users using frequency adverbs. Chao et al. in [3] equipped their robot with non-verbal behaviours (shake head, nod head and shrug) to expose the evolution of the label prediction capabilities of the system during the training. Similar information where integrated by augmenting the robot's questions in [9], exposing concepts like prediction power and confidence of the LS.

Environment Model. This model aims to expose the situational awareness of the AIS, conveying to the user what the system can perceive in its surrounding. For LSs, the model extends to expose the nature of the input in relation to the training. The user could then act as a quality controller, checking for the suitability of the data the system is training on. This model shares the main challenge with the analytical model: non-expert users might lack the technical skills to evaluate the quality of what the robot perceives. Moreover, a transparent LS dealing with a large amount of data should consider what to expose to the user and when, in order to avoid distractions or cognitive overloads that may hinder the training quality.

Robot-of-human factors

Teamwork Model. This model deals with role assignment and autonomy level, especially when the AIS is collaborating with the user (Human-Robot Collaboration). On the role assignment, a person interacting with a LS should be or made aware of their roles of teachers for the LS, with its requirements and responsibilities.

One question that rises in the specific case of LSs is: *After the training, what is the level of autonomy that the system can achieve?.* With this information available, users can tune their teaching, neither aiming for impossible results nor under-training their robot. This model shares challenges with the intentional model, where the intended use of the system needs to be clarified. However, as the final autonomy level of the system can depend to a certain extent on the user's skill as a teacher [2], presenting measures of expected results can cause mismatches between the real and the perceived skill of the system. A robot presenting the autonomy level achievable in the best case scenario of an impeccable training might inflate the expectations of the user. On the other hand, a robot presenting the worst case scenario following an awful training might discourage the user from putting effort in the training.

Human State Model. Lyons proposed this model to allow AISs to express what is available to them regarding the cognitive, emotional and physical state of their human collaborator. Such detection systems are being developed for example in the automotive industry to detect driver's fatigue and take according decisions to avoid accidents [6]. Similar mechanisms can be beneficial also for LSs. Detecting tiredness or boredom in their human teacher, could allow learning robots to avoid an useless and potentially harmful training session. Knowing the state of the human teacher could also allow the robot to steer the training, e.g. to lower the cognitive load on the user if perceived as in distress. Without considering the technical difficulties related to the estimation of these user information, also finding a tradeoff between quality of the interaction and performance of the LS remains a challenging aspect.

3 CONCLUSIONS

With the robotics community aware of the importance of transparency for AISs, we wanted with this paper to shed some light on the particular case of transparency for systems that learn by interacting with their users. Transparency has enormous potential to improve the reliability and performances of learning robots [1, 3, 8]. Among the several challenges identified, we consider the exposure of the analytical and the task models the most challenging yet valuable, especially given the popularity of black-box learning techniques. We also believe the modelling of the human state to be a promising topic, where research efforts are needed to develop methods to reliably detect the human state and models to take that into account to aid the decision making of LSs during training.

Acknowledgements

This research was supported by the Strategic Research Council at the Academy of Finland, decision 314180.

REFERENCES

[1] Maya Cakmak, Crystal Chao, and Andrea L. Thomaz. 2010. Designing interactions for robot active learners. *IEEE Transactions on Autonomous Mental Development* 2, 2 (2010), 108–118.

[2] Maya Cakmak and Andrea L. Thomaz. 2014. Eliciting good teaching from humans for machine learners. *Artificial Intelligence* 217 (2014), 198–215.

[3] Crystal Chao, Maya Cakmak, and Andrea L. Thomaz. 2010. Transparent active learning for robots. In *Human-Robot Interaction (HRI), 2010 5th ACM/IEEE International Conference on.* IEEE, 317–324.

[4] Sonia Chernova and Andrea L. Thomaz. 2014. Robot learning from human teachers. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 8, 3 (2014), 1–121.

[5] Bradley Hayes and Julie A. Shah. 2017. Improving Robot Controller Transparency Through Autonomous Policy Explanation. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction.* ACM, 303–312.

[6] T. Inagaki. 2008. Smart collaboration between humans and machines based on mutual understanding. *Annual Reviews in Control* 32, 2 (2008), 253–261. https://doi.org/10.1016/j.arcontrol.2008.07.003

[7] Joseph B. Lyons. 2013. Being Transparent about Transparency: A Model for Human-Robot Interaction. In *2013 AAAI Spring Symposium Series.*

[8] Mattia Racca and Ville Kyrki. 2018. Active Robot Learning for Temporal Task Models. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction.* ACM.

[9] Stephanie Rosenthal, Anind K. Dey, and Manuela Veloso. 2009. How robots' questions affect the accuracy of the human responses. In *Robot and Human Interactive Communication, 2009. RO-MAN. The 18th IEEE International Symposium on.* IEEE, 1137–1142.

[10] Andreas Theodorou, Robert H. Wortham, and Joanna J. Bryson. 2016. Why is my robot behaving like that? Designing transparency for real time inspection of autonomous robots. In *AISB Workshop on Principles of Robotics.* University of Bath, University of Sheffield.