Towards Explainable and Trustworthy Collaborative Robots through Embodied Question Answering

Lars Kunze[1], Omer Gunes[2], Dylan Hillier[2], Matthew Munks[1], Helena Webb[3], Pericle Salvini[2], Daniel Omeiza[2], and Marina Jirotka[2]

***Abstract*— Collaborative robots (or cobots) will offer significant societal benefits, but their large-scale deployments may also lead to unintended consequences. The ability to query, analyse, and understand data from cobots will be a fundamental requirement for ensuring safety, accountability, and trust. To this end, we propose *embodied question answering* as a means to enable cobots to explain themselves and make them trustworthy. Our approach is founded in responsible research and innovation and thereby will shape the future of responsible robotics design, development, and deployment for cobots.**

In this paper, we first provide some background on responsible robotics. Second, we elaborate on the need for explanations. Third, we describe our approach to embodied question answering, and finally, we discuss open challenges before we conclude.

I. INTRODUCTION

Collaborative robots (or cobots) are designed to work alongside the human workforce with the aim of making monotonous and/or physically demanding tasks more efficient. The collaboration between humans and cobots in close proximity will fundamentally change how we transport, assemble, and/or inspect materials and goods. However, the close interaction between humans and cobots will also lead to new challenges in terms of ensuring human safety.

Many works within robotics consider safety in terms of ensuring humans never enter unsafe states, e.g. by avoiding collisions, physical hazards and emotional/psychological harms. Since incidents and accidents resulting in unsafe states are unfortunately inevitable to some extent, another important component of safety is revealing what went wrong in an incident and what changes can be made to ensure it does not reoccur. Event data recorders (or *black boxes*) are designed for this purpose. They are software devices or modules capable of recording data from sensors, actuators and/or decision making processes. Currently they are mandatory in several applications and used during accident investigations, in particular in the aviation[1] and automotive sectors.

In this paper, we describe a framework for extracting information from data recorders using Embodied Question Answering (EQA) (see Fig. 1). We propose a conversational question answering system to extract information from an *embodied* cobot system. As the cobot's knowledge base and its episodic memories relate to observations and events in a physical environment, we call it *embodied question answering*. The conversational context is used to generate answers to contextualised questions within a dialogue. The framework was motivated by an in-depth analysis of robot accident investigations [1]. It supports querying, analysing, and interpreting cobot data, drawing on state-of-the-art models for graphical representations and inference methods and conversational question answering.

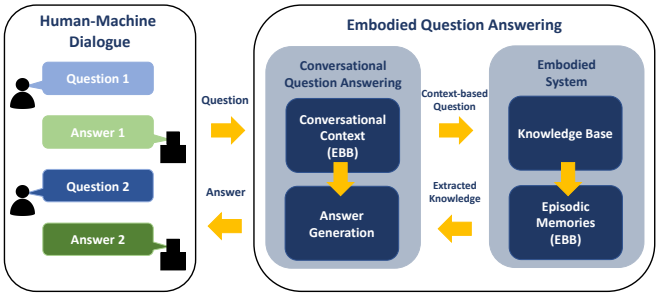

Fig. 1. Embodied Question Answering (EQA) for Collaborative Robots. Questions are answered using information from the system's knowledge base, its episodic memories as well as the conversational context which are stored in a data recording device called the Ethical Black Box (or EBB). EQA advances the capacity for investigators to derive useful explanations from a robot following an accident and this work therefore contributes to increasing trust and responsibility in robotics.

Experts, investigators, and regulators might pose complex and/or specific questions to identify the underlying causes of an accident/incident. Offering answers to these questions in a timely manner helps improve the quality of these investigations.

Overall, we argue that EQA can positively impact at least three key aspects of the deployment of cobots:

1) **Safety**—in case of an incident (e.g. a near-miss), an investigation can reveal what went wrong and aid us in addressing any issues;
2) **Accountability**—in case of an accident, a careful analysis of the logged system state can facilitate legal investigations by identifying any actors (including co-workers, developers, and operators) who may be held accountable for malfunctions, oversights or other problems that caused the incident to occur;
3) **Trust**—the possibility of questioning the system to obtain explanations can help us understand why decisions and/or actions have been taken. This can increase trust and acceptance of a system.

[1] Oxford Robotics Institute, Dept. of Engineering Science, University of Oxford, lars@robots.ox.ac.uk
[2] Dept. of Computer Science, University of Oxford, omer.gunes@cs.ox.ac.uk
[3] University of Nottingham, helena.webb@nottingham.ac.uk
[1] https://www.gov.uk/government/publications/how-we-investigate/how-we-investigate

In the remainder of the paper, we first provide some background on responsible robotics (Section II). Second, we elaborate on the need for explanations for cobots (Section III). Third, we describe our approach to embodied question answering in Section IV, and finally, we discuss open challenges in Section V before we conclude in Section VI.

II. Responsible Robotics: The Ethical Black Box

Responsibility takes a very specific meaning when understood within the context of Responsible Research and Innovation (RRI). RRI is an initiative that emerged around 20 years ago and has since gained significant traction across academia, policy and industry [2]. In broad terms RRI seeks to make the processes and outcomes of research and innovation more inclusive, so that they better align with societal values, needs and interests. Frameworks for RRI [3], [4] set out actions to support this aim; for instance, emphasising the value of engagement with stakeholders in innovation, and the importance of anticipating the intended and unintended impacts of a development process.

Building on this, responsibility in robotics can be understood as a set of good practices to ensure that the design, manufacture, operation, repair and end-of-life recycling of robots maximise the potential benefit to society and minimise potential harms, including harms to the environment [1]. The focus on practices is significant as the responsible approach seeks to incorporate existing high-level principles and standards in ethical robotics whilst also identifying actions that practitioners can undertake. The growing presence of cobots in contemporary society creates many opportunities for responsible actions across the phases of design, development and implementation to increase the benefits that these technologies can bring and to limit their harms.

The work described in this paper on EQA is driven by this responsible approach. It supplements an ongoing research study[2] [1] that helps to embed responsibility into robotics by creating a mechanism—a robot data recorder similar to a 'black box' flight recorder used in aviation—to support processes of investigation following accidents involving social robots of various kinds. The concept of the so-called *Ethical Black Box* (or EBB) [5] is critical, especially in the event of an accident or incident, to understand why and how the robot may have been involved and to establish accountability and responsibility. It is called *ethical* just because the proponents believe that it would be irresponsible to deploy robots without one [5]. In fact, the word *ethical* here has nothing to do with the capability of the EBB to make ethical decisions. Indeed, the EBB is a passive device designed for recording and securely storing data and make data available for an investigation after the accident/incident has occurred.

Whilst hopefully rare, it can be anticipated that accidents will occur as robots become more prevalent in society and that such accidents will risk causing harm to humans. It is therefore essential that steps are put in place in advance to address this problem. Ensuring safety can be understood as consisting of several facets, among which there are anticipatory measures (ex ante), in which the goal is to prevent possible hazards (e.g. by means of risk assessment procedures) and retrospective measures (ex post) in which the goal is to learn from errors (such as near-misses) and this is the purpose of the EBB. The data collected by a robot's EBB can be central to a process of investigation that identifies the cause of the accident. Once identified, safety changes can be made to prevent the accident reoccurring and this will also help to restore public trust in robots. EQA advances the capacity for investigators to derive useful information from a robot following an accident and this work therefore contributes to increasing responsibility in robotics.

[2]https://www.robotips.co.uk

III. Need for Explanations in Cobots

Cobots represent a paradigmatic shift in human-robot interaction. Only two decades ago robots operating in industrial settings had to be fenced off from human workers. Today, thanks to advances in physical safety, such as passive compliance devices, human-robot collaboration is possible. However, because robots and humans will be closer together and capable of establishing physical contact, the possibility of accident/incidents leading to physical collisions may increase. The need for explanations in cobots stems from increasing concerns for safety, transparency and accountability. Explanations are one way of achieving these goals. In this section, we discuss the need for explanation in the light of transparency, accountability, and trust.

A. Transparency and Accountability

One generally agreed upon notion of accountability is associated with the process of being called 'to account' to some authority for one's actions [6]. Mulgan [7] elucidated that accountability entails responsibility but, unlike responsibility, it requires explanations about actions and it cannot be shared. In the human-machine context, [8] conceptualise accountability as the ability to determine whether a decision of a system was made in compliance with procedural and substantive standards, and importantly, to hold one responsible when there is a failure to meet the standards.

With cobots, accountability becomes a challenging issue mainly because of the various operations involved (e.g., perception, planning, controls, system management among others) that demand inputs from multiple stakeholders and to put data in context; this can result in responsibility gaps due to the impossibility to retrieve such information.

As identified by Mulgan [7], achieving accountability requires social interaction and exchange. At one end, the requester of an account seeks answers and rectification while at the other end, the respondent or explainer responds and accepts responsibility if necessary.

In the context of this paper, the cobot is being called by a stakeholder to provide an account; we expect this to be provided in the form of an explanation that is intelligible to the requester to facilitate the assignment of responsibilities.

There have been debates on how responsibility should be allocated for certain cobot accidents.

The social aspect of accountability described by Mulgan [7], demands that the aforementioned recommended approaches are able to plug into explanation mechanisms where causes and effects of actions can be communicated to the relevant stakeholders in intelligible ways. In addition to accountability for accident cases, which has gained much attention in the industry reports, actions resulting in undesired, discriminatory, and inequitable outcomes also need to be accounted for. This means that stakeholders who may not have direct involvement in the management of the cobots should be able to instantaneously request accounts in the form of intelligible explanations for such undesired actions when they occur.

Finally, the capability to provide intelligible explanations to specific stakeholders is one of the five levels of transparency used in the related IEEE standard to measure transparency [9]. In the specific case of cobots as team mates, the lack of explanations in the decisions made by the robot worker could bring about stress and anxiety in the human worker and the feeling of loss of control.

B. Trust

Research investigating trust in automation has been around for decades, i.e., since the introduction of interpersonal trust theories into the human-machine interaction domain by [10], [11], [12]. While various definitions of trust in automation have been proposed, the most commonly adopted definition is that put forward in [13]. Lee and See [13] consider trust as a social psychological concept that is important for understanding automation partnership. The authors emphasise that trust is the attitude that automation will help an individual to achieve their goals in a situation characterised by uncertainty and vulnerability. Trust in automation, as made evident in [14], [15], has significantly influenced the acceptance of and reliance on automated systems. As opposed to a binary categorisation, trust can be more finely calibrated so that an individual's trust levels on an automated system adequately reflects the actual capabilities and functional scope of an automated system. This trust calibration is considered to be an important requirement for safe and efficient human-machine interaction [10].

Information about the functioning modes of a cobot at the user's disposal can help the user create a better understanding of the cobot's behaviour, eventually adding to the user's knowledge base [16], and helpful for constructing calibrated trust. This information could be presented as explanations of the operational modes and behaviour of a complex system, such as a cobot, especially when it acts outside the expectations of the user. We note that trust can break down when there are frequent failures without adequate explanations, and regaining trust once lost can be challenging [17], [18].

Researchers (e.g., in [19]) suggest that the provision of meaningful explanations from cobots to stakeholders is one way to build the necessary trust in cobot technology.

IV. EMBODIED QUESTION ANSWERING

Our approach to embodied question answering is based on conversational question answering integrated with an embodied system (see Fig 1).

To start a dialogue, a user and/or an accident investigator can pose a question to the EQA system (via speech or text). The question is contextualised and used to extract relevant information from the cobot's knowledge base and its episodic memories. Based on the question, the current conversational context and the extracted knowledge an answer is generated. The process repeats until the human-machine dialogue ends.

Consider the following hypothetical dialogue between an accident investigator (INV) and the cobot system (EQA) after a person was injured by a robot arm:

INV: "What did you do during 10–11am last Monday?"
EQA: "I performed an assembly task."
INV: "Did you work alone or with a co-worker?"
EQA: "I performed the task alone."
INV: "Did you see anyone in the workspace?"
EQA: "Somebody entered the workspace at 10:29am."
INV: "Did you notice any anomalous forces on the arm?"
EQA: "Anomalous forces were detected at 10:34am."
...

It is worth noting that the subsequent questions build on the established temporal as well as spatial context. Hence, identifying and extracting the relevant context is critical for answering the questions correctly.

Below we describe how we use knowledge graphs as an intermediate, interpretable representation to translate between natural language (NL) questions, formal robot knowledge, logged data from the EBB, and NL answers for EQA.

A. A Knowledge Graph-based Approach to EQA

A cobot's knowledge about the world may be stored in knowledge graphs: these are flexible and natural data structures for storing structured information. Furthermore as explored in [20], knowledge graphs can be used to give robots commonsense knowledge about the world. Similarly contemporary language models have been shown to store implicit knowledge [21], and as discussed in [22], we can utilise the strengths of the two by combining graph neural networks—which can reason over and extract relevant knowledge from knowledge graphs, with fine-tuned language models. Unlike in [22], our question answering system is placed in a conversational context, which requires more complicated reasoning as discussed in [23].

As such in addition to querying knowledge graphs, our system enriches them by storing knowledge about the conversation into an additional knowledge graph. Another advantage to this approach is that we can ensure that the system only has access to sensitive information suitable for the user in question by switching out which knowledge graphs the system can query over. Furthermore unlike with utilising the implicit knowledge stored in the language model, this approach makes it explicit what information is being drawn upon to generate answers.

Our approach to EQA has three steps. The first step involves grounding queries with the information in knowledge graphs. We implement this by identifying how similar entities mentioned in the question are to entities in the knowledge graph. This is then used to extract a subgraph of the knowledge graph with language-model based embeddings (following [24]). We then use a graph neural network model based on [25], which is known to work well on knowledge graphs with rich features. Finally we feed this information into a language-model fine-tuned on the conversational question answering dataset [26] to generate answers.

V. OPEN CHALLENGES

In this section, we discuss a few open challenges for making cobots explainable and trustworthy.

Language Grounding: During their operation cobots generate vast amounts of data, but how to sift through this data to extract and filter the most relevant information is an open problem. While our knowledge graph-based approach can identify some of the relevant information further research is needed to map NL questions to logged robot data.

Faithfulness: Another open challenge is how to ensure that the explanations provided are genuine (or faithful). A possible solution could be the standardisation of the process for generating explanations. The logged meta-data in the form of knowledge graph triples can also provide some assurance, however it is difficult to guarantee that the answers produced by the downstream language model reason appropriately using this extracted knowledge.

Psychological Safety: The capability of providing intelligible explanations of the causes of a physical accident or incident become fundamental both for ensuring safety (understanding what happened) and liability (avoiding the so called responsibility gap) and hence improving the overall trust in the system. However, there might be concerns regarding the psychological safety of human-robot collaboration [27]. Among the challenges for ensuring safety is how to collect data and provide intelligible explanations to help identifying psychological hazards deriving from having robots as co-workers (e.g. stress, anxiety, boredom, tiredness).

Explainable AI: How to explain the working of machine learning systems or black-box models? With respect to transparency and accountability it is important that decisions of a cobot can be explained. Hence, (deep) machine learning and/or black-box models should be complemented with methods that can provide detailed explanations of their decision making.

Level of detail: How to tailor explanations to different types of stakeholders? The level of detail (in terms of information) anticipated by the explanation recipients, the explanation type and the mode of communication vary with respect to the type of recipient and purpose for the explanation. This highlights the importance of explanation categorisation/personalisation with respect to stakeholders.

Privacy: How to ensure protection of sensitive or personal data? There are increasing concerns about the collection and use of personal data. Cobots could potentially be used to collect sensitive information from users either legally or illegally. To this end, regulation for control rights over personal and/or sensitive data are needed.

VI. CONCLUSIONS

In this paper, we discussed how providing intelligible explanations could improve robot safety, transparency, accountability, and trust. In addition to these benefits, we argue that providing explanations could be a way of improving the robot functionality or performance, namely the robot capability of accomplishing its collaborative tasks with humans. In other words, the request for explanations by humans could be motivated not by safety or accountability purposes, but by the collaborative nature of a specific task, for instance for clarification purposes (e.g. Human: "Why did you (robot) bring me the red bolt instead of the yellow one?" Robot: "Because we were out of yellow bolts").

Overall, we believe that our work on embodied question answering can significantly contribute to creating the next generation of responsible, explainable, and trustworthy collaborative robots.

ACKNOWLEDGMENT

This work was supported by the Human-Machine Collaboration Programme of the University of Oxford and Amazon Web Services (AWS). It was also partially supported by the EPSRC projects RoboTIPS (grant reference: EP/S005099/1) and RAILS (grant reference: EP/W011344/1).

REFERENCES

[1] A. F. T. Winfield, K. Winkle, H. Webb, U. Lyngs, M. Jirotka, and C. Macrae, *Robot Accident Investigation: A Case Study in Responsible Robotics*. Cham: Springer International Publishing, 2021, pp. 165–187. [Online]. Available: https://doi.org/10.1007/978-3-030-66494-7_6

[2] R. von Schomberg, *A Vision of Responsible Research and Innovation*. John Wiley Sons, Ltd, 2013, ch. 3, pp. 51–74. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118551424.ch3

[3] R. Owen, J. Stilgoe, P. Macnaghten, M. Gorman, E. Fisher, and D. Guston, *A Framework for Responsible Innovation*. John Wiley Sons, Ltd, 2013, ch. 2, pp. 27–50. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118551424.ch2

[4] European Commission, "Rome declaration on responsible research and innovation in europe," 2014.

[5] A. F. T. Winfield and M. Jirotka, "The case for an ethical black box," in *Towards Autonomous Robotic Systems*, Y. Gao, S. Fallah, Y. Jin, and C. Lekakou, Eds. Cham: Springer International Publishing, 2017, pp. 262–273.

[6] G. W. Jones, "The search for local accountability," *Strengthening local government in the 1990s*, pp. 49–78, 1992.

[7] R. Mulgan, "'Accountability': An ever-expanding concept?" *Public administration*, vol. 78, no. 3, pp. 555–573, 2000.

[8] F. Doshi-Velez, M. Kortz, R. Budish, C. Bavitz, S. Gershman, D. O'Brien, S. Schieber, J. Waldo, D. Weinberger, and A. Wood, "Accountability of AI under the law: The role of explanation," *arXiv preprint arXiv:1711.01134*, 2017.

[9] A. F. T. Winfield, S. Booth, L. A. Dennis, T. Egawa, H. Hastie, N. Jacobs, R. I. Muttram, J. I. Olszewska, F. Rajabiyazdi, A. Theodorou, M. A. Underwood, R. H. Wortham, and E. Watson, "IEEE P7001: A Proposed Standard on Transparency," *Frontiers in Robotics and AI*, vol. 8, 2021. [Online]. Available: https://www.frontiersin.org/article/10.3389/frobt.2021.665729

[10] B. M. Muir, "Trust between humans and machines, and the design of decision aids," *International journal of man-machine studies*, vol. 27, no. 5-6, pp. 527–539, 1987.

[11] ——, "Trust in automation: Part i. theoretical issues in the study of trust and human intervention in automated systems," *Ergonomics*, vol. 37, no. 11, pp. 1905–1922, 1994.

[12] J. Lee and N. Moray, "Trust, control strategies and allocation of function in human-machine systems," *Ergonomics*, vol. 35, no. 10, pp. 1243–1270, 1992.

[13] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human factors*, vol. 46, no. 1, pp. 50–80, 2004.

[14] D. P. Biros, M. Daly, and G. Gunsch, "The influence of task load and automation trust on deception detection," *Group Decision and Negotiation*, vol. 13, no. 2, pp. 173–189, 2004.

[15] B. M. Muir and N. Moray, "Trust in automation. part ii. experimental studies of trust and human intervention in a process control simulation," *Ergonomics*, vol. 39, no. 3, pp. 429–460, 1996.

[16] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Metrics for explainable ai: Challenges and prospects," *arXiv preprint arXiv:1812.04608*, 2018.

[17] M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce, and H. P. Beck, "The role of trust in automation reliance," *International journal of human-computer studies*, vol. 58, no. 6, pp. 697–718, 2003.

[18] P. Madhavan and D. A. Wiegmann, "Effects of information source, pedigree, and reliability on operator interaction with decision support systems," *Human Factors*, vol. 49, no. 5, pp. 773–785, 2007.

[19] R. R. Hoffman and G. Klein, "Explaining explanation, part 1: Theoretical foundations," *IEEE Intelligent Systems*, vol. 32, no. 3, pp. 68–73, 2017.

[20] L. Kunze, M. Tenorth, and M. Beetz, "Putting people's common sense into knowledge bases of household robots," in *33rd Annual German Conference on Artificial Intelligence (KI 2010)*. Karlsruhe, Germany: Springer, September 21-24 2010, pp. 151–159.

[21] F. Petroni, T. Rocktäschel, P. Lewis, A. Bakhtin, Y. Wu, A. H. Miller, and S. Riedel, "Language models as knowledge bases?" *arXiv preprint arXiv:1909.01066*, 2019.

[22] M. Yasunaga, H. Ren, A. Bosselut, P. Liang, and J. Leskovec, "Qa-gnn: Reasoning with language models and knowledge graphs for question answering," *arXiv preprint arXiv:2104.06378*, 2021.

[23] P. Christmann, R. Saha Roy, A. Abujabal, J. Singh, and G. Weikum, "Look before you hop: Conversational question answering over knowledge graphs using judicious context expansion," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2019, pp. 729–738.

[24] J. Ni, G. H. Ábrego, N. Constant, J. Ma, K. B. Hall, D. Cer, and Y. Yang, "Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models," *arXiv preprint arXiv:2108.08877*, 2021.

[25] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.

[26] S. Reddy, D. Chen, and C. D. Manning, "Coqa: A conversational question answering challenge," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 249–266, 2019.

[27] BS8611, BSI, "Robots and robotic devices, guide to the ethical design and application of robots and robotic systems," *British Standards Institute*, 2016.