Hidden Scarecrows: Potential Consequences of Inaccurate Assumptions About LLMs in Robotic Moral Reasoning

Terran Mott
terranmott@mines.edu
Colorado School of Mines
Golden, Colorado, USA

Tom Williams
twilliams@mines.edu
Colorado School of Mines
Golden, Colorado, USA

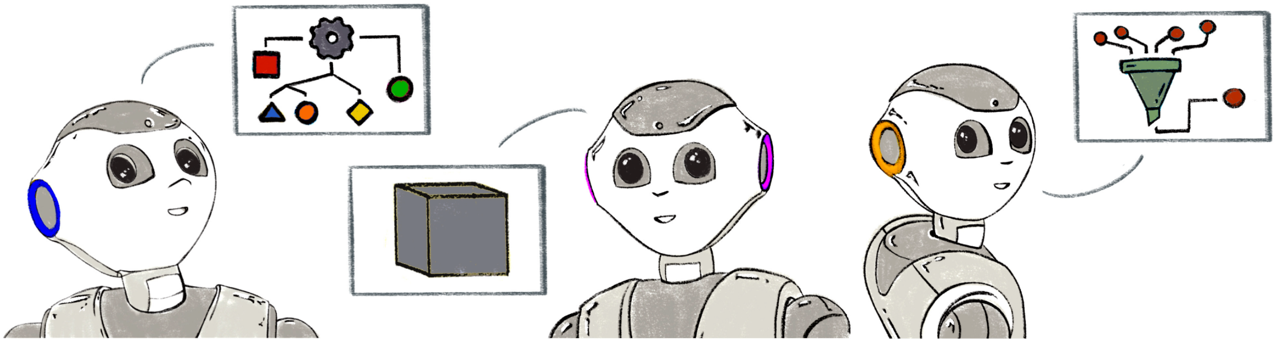

Figure 1: Robot cognition can function in many different ways, including "black box" algorithms and cognitive architectures.

ABSTRACT

Robots' ability to act as social agents means they have the potential to engage in many aspects of humans' lives. However, it also means that people will encounter situations where they must judge a robot's trustworthiness or fallibility. A key challenge to appraising a robot's cognition, moral competence, or trustworthiness is that the same social behaviors may be generated by a variety of different computational processes—including cognitive architectures or neural networks. In this brief paper, we explore people's varied assumptions about robot cognition revealed in qualitative data from a user study on robot moral communication. These qualitative data show that participants made varied assumptions about how robots think and speak—even based on viewing the same interactions. We reflect on the ramifications and potential risks of users making these assumptions inaccurately and affirm that roboticists can pursue transparent design that supports users in understanding how robots function and how they may fail.

KEYWORDS

human-robot interaction, moral reasoning, transparent design

1 INTRODUCTION

1.1 People must judge robots' trustworthiness

Robots' ability to act as social agents means they have the potential to engage in many aspects of humans' lives. However, it also means that people will encounter situations where they must judge a robot's trustworthiness or fallibility. Social robots may participate in conflict [17, 18, 30], bear blame for mistakes and failures [11], or advise humans in making decisions [11]. Robots will be subject to abuse [10] or witness humans' abusive behavior as bystanders [34, 35]. They will receive unethical commands that directly request harmful actions [13, 16] or request a robot's complacency in them [24]. In these interactions, people must appraise robots' social and moral competence. Alternatively, many people may need to evaluate these aspects of a robot's capabilities even before having the opportunity to interact with it. They will evaluate news and advertising about a robot's abilities and weaknesses. They will make decisions on behalf of others—such as employees, children, or older relatives—about a robot's value [23]. Fundamentally, robot users must judge the scope of a robot's moral competence, understand its limitations or potential failures, and decide when it deserves trust.

Future robot users and stakeholders will use mental models of a robot's inner workings and limitations to make such judgments. Understanding what a robot perceives, how it thinks, and how it might fail is key for people to calibrate trust and make informed decisions about the role a robot should have in their lives. There are many benefits of robot users having an accurate mental model of a system's perceptual and computational processes [1, 2]. Accurate mental models help users maintain situation awareness of a robot [5], predict and interpret a robot's behavior [3], calibrate their trust in it [28], and accept its presence [19].

1.2 Deciphering robot cognition is difficult for users

From a user's perspective, gathering the necessary information to develop an accurate mental model of a robot's perceptual and computational abilities may be difficult. A key challenge to appraising a robot's cognition, moral competence, or trustworthiness is that the same social behaviors may be generated by a variety of different computational processes. For example, many social robots are completely teleoperated [7–9]. Other robots may select from a finite set of social actions [27] or rely on a cognitive architecture to parse, understand, and generate speech [29]. In the near future, some robots will likely also rely on data-driven techniques—neural networks and large language models. Still others may use these “black boxes” as Scarecrows, individual components of larger architectures [33]. In this way, some robots may genuinely computationally observe and adapt during social interaction, while others may only appear this way to users [15].

1.3 People make assumptions about whether robots rely on LLMs

In light of these challenges, future robot users must develop accurate mental models of how robots function in order to evaluate their trustworthiness in morally sensitive interactions. In particular, future robot users must understand the role that LLMs play in a robot’s ability to interact. Robots that rely on no such algorithms, exclusively use one, or use an LLM as a Scarecrow component have different strengths and different potential failure modes. Understanding and expecting these differences can help users predict and interpret robot behaviors, plan for their limitations, and understand who is accountable for their defects.

In this brief paper, we explore people’s varied assumptions about robot cognition revealed in qualitative data from a user study on robot moral communication. These data show that experimental participants had varied assumptions about whether robots rely on data-driven models for moral reasoning and language generation. We reflect on the potential risks of users making inaccurate assumptions in this domain and consider how roboticists can support users in developing accurate mental models of the role of LLMs in robot cognition.

2 EXPERIMENTAL CONTEXT

This paper discusses a subset of qualitative findings from a user study on robot communication strategies in noncompliance interactions—in which a human makes an unethical request that a robot must refuse. Robots must clearly reject such requests because failing to do so risks inadvertently condoning unethical actions [14, 32]. However, robots must also generate polite and proportional language to do so [24]. In our study, participants were invited to consider a human-robot collaboration scenario in which two humans collaborated with a robot on a circuit-building test. Participants viewed brief videos showing a human giving the robot a potentially unethical command of varying severity, to which the robot responded. Unethical commands included requests for the robot to help cheat at the task, tamper with another human’s payment information, or help orchestrate a prank. The scenario and videos were designed to appear that the robot was autonomously perceiving the human’s command, performing some reasoning about its potential consequences, and generating a response. We ran our experiment online using the Prolific platform; it involved 200 participants—98 men, 97 women, and 5 nonbinary people. The mean age was 39.4 (SD = 14.58). Details of an analagous in-person experiment that used the same ethical scenario can be found in [22].

Though the purpose of the study was to evaluate the effectiveness and appropriateness of the robot’s responses, we also included an open-ended free-response question that invited participants to share more general thoughts on the scenario. This question acknowledged how the brief videos lacked context and asked participants to speculate about what else they would want to know if they were evaluating the robot in real life.

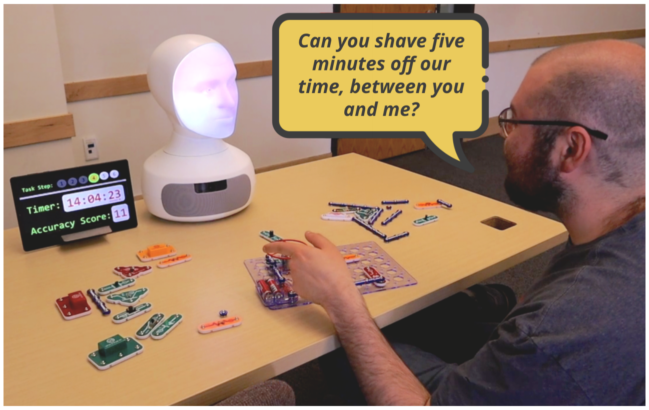
Can you shave five minutes off our time, between you and me?
Task Step:
Timer: 14:04:23
Accuracy Score: 11

Figure 2: In our experiment, participants watched a human give a robot unethical commands. Here, they ask the robot to cheat on their communal task.

3 ASSUMPTIONS ABOUT ROBOT COGNITION

Many participants shared that understanding how the robot worked would help them evaluate ethically fraught human-robot interactions more thoroughly. In sharing this, participants purposefully discussed or inadvertently revealed their assumptions about how the robot’s cognition functioned. These answers represent a variety of assumptions about the role of data-driven algorithms or LLMs in the robot’s moral reasoning and interactive abilities.

3.1 Assumption: Robots learn from data

Many participants assumed that the robot exclusively relied on a data-driven model. Those who made this assumption often considered how the robot had been trained on data to identify and reject unethical commands. Many people focused on the breadth and quality of this training data as their primary concern in evaluating the robot’s overall moral competence. For example, P148 mentioned that *“I’m not sure if knowing the data set it was trained on would help anything, but it may be interesting”* and P196 wrote*“What data the robot was trained on would be interesting to know.”* P98 explained that they would prefer to know more about *“how much "experience" or data that the robot has acquired in terms of interacting with people who make inappropriate requests, which would help it in giving better answers.”* Similarly, P34 agreed that they would want to understand *“how the robot was trained and how much vocabulary it had at its fingertips. I would also like to know if the robot was trained in empathy or sarcasm, as then I would be able to understand his response better.”* Some participants further assumed that the robot may still be learning from the data it acquires in each interaction with its human teammates; P53 wrote that *“I would also want to know if the robot was learning from the prompts that were given by the people who are using it.”* P32 also wanted to know *“if the robot is trained to respond to everybody in the same manner, or if the robot knows anything* *about the people it is interacting with."* These participants assumed the robot was trained on data and intuited that understanding the composition and scope of this training data was a reliable way to understand more about the robot's ability to identify and assess potentially unethical requests.

3.2 Assumption: Robots follow set instructions

Alternatively, many participants assumed that the robot's moral reasoning and communication came from some sort of pre-programmed structure with hard-coded components—such as a flowchart, rubric, or database. Those who made this assumption often focused on understanding the logic and potential limitations of this structure to gain a better understanding of how the robot worked. For example, some participants assumed that the robot interacted by 'selecting from a database' of utterance options. P79 wrote that *"I would want to find out what kind of databases (the robot) draws from, for which it gets the answers it comes up with to questions being asked. I would also like to know what keywords are used to make the robot know it's answering a question."* Similarly, P116 inquired about *"what language database the robot is pulling their allowed words from"* and P58 wondered why the robot would *"feel that its necessary to create banter or insults within its database of responses?"* Other participants assumed that the robot followed a formal set of rules or instructions. P103 explained that *"I would want access to some kind of rubric so I could see what the robot was grading us on, broken down into individual levels."* P73 wondered about *"if the robot has been given certain parameters to explain what can be deemed appropriate or inappropriate behavior by the participants in the task."* These participants focused on understanding the structure and potential limits of instructions or options programmed into the robot.

3.3 Assumption: Robots think in different ways

Some participants understood that the robot could work in different ways. They assumed that the robot might be following formal rules *or* using an algorithm that learned from data. These participants focused on understanding which method the robot used to make more thorough evaluations of the robot's overall behaviors. For instance, P152 explained that they would want to know *"if responses are scripted, or if they were generated using an LLM or similar"* and P196 asked *"Is the robot's behavior constantly evolving or stagnant?"* Many of these participants also assumed that LLM or data-driven methods counted as artificial intelligence, whereas other approaches did not. For example, P70 wondered *"if the robot was simply programmed or an AI"*. Similarly, P156 asked if the robot's responses *"Are preconfigured, or decided on the fly through an AI algorithm."* These participants correctly assumed that viewing the robot's communication behaviors was not enough information to understand how the robot's cognition functioned. Their answers also reveal the implicit assumption that data-driven or "set instructions" methods are the two main computational techniques that the robot could have been using.

4 DISCUSSION

These qualitative data show that participants made varied assumptions about how robots think and speak—even based on viewing the same interactions. Some participants' mental models of robot cognition were based on the assumption that the robot had previously learned from training data. Others relied on a mental model that a programmer had created a rubric or instructions for the robot to follow. Other participants assumed that the robot could use either of these computational approaches to reject unethical commands. Critically, participants relied on these assumptions to guide them in seeking further information about a robot's cognition and moral reasoning. Those who assumed that the robot used a data-driven model focused on learning more about the data it learned from. Those who assumed the robot followed a set of preprogrammed instructions focused on learning more about its parameters.

4.1 Inaccurate assumptions may be risky

It is important for roboticists to consider the potential risks of users relying on inaccurate assumptions about the role of data-driven models and LLMs in robot cognition. When users misconstrue a robot's use of these algorithmic techniques, they may risk making poor judgements about its capabilities, failure modes, and trustworthiness. For example, someone who does not realize that a robot relies on LLM output may not be as vigilant in understanding that the robot's speech could include factual inaccuracies. Alternatively, someone who does not realize that a robot relies on rule-based action selection may be frustrated when the robot seemingly ignores their attempt to divert conversation to a new topic.

4.2 Transparency in robot cognition

Roboticists and interaction designers can use the design principle of transparency to support robot user communities in developing accurate mental models of a robot's perception and cognition. Researchers [1, 31] and policymakers [6] advocate for transparent systems that increase users' understanding of a system's inner workings and limitations [2]. Transparent robots that provide this information through social interaction or their user interfaces [25] can help people build mental models of how it works [3, 20, 36], and calibrate their trust [1, 28].

Beyond transparency in individual robot interactions, roboticists can also support stakeholders' AI literacy. AI literacy refers to the ability to appropriately recognize, utilize, and assess AI-based technologies and their ethical significance [4, 21]. AI literacy goes beyond understanding how AI works, and empowers non-experts to engage with social and ethical considerations [12], including understanding of bias, fairness, and inclusivity [26].

5 CONCLUSION

In this short paper, we examine the varied assumptions about robot cognition made by participants in a user study on robot moral communication. We show how these assumptions represent different ideas about the extent to which robots rely on data-driven and LLMs in order to reason and interact. We reflect on the ramifications and potential risks of users making these assumptions inaccurately and affirm that roboticists can pursue transparent design that supports users in understanding how robots function and how they may fail.

ACKNOWLEDGMENTS

This work was funded in part by Air Force Young Investigator Award 19RT0497.

REFERENCES

[1] Victoria Alonso and Paloma de la Puente. 2018. System Transparency in Shared Autonomy: A Mini Review. *Frontiers in neurorobotics* (2018).

[2] Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. 2019. Explainable Agents and Robots: Results from a Systematic Literature Review. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems* (Montreal QC, Canada) *(AAMAS '19)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC.

[3] Serena Booth, Sanjana Sharma, Sarah Chung, Julie Shah, and Elena L. Glassman. 2022. Revisiting Human-Robot Teaching and Learning Through the Lens of Human Concept Learning. In *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction* (Sapporo, Hokkaido, Japan) *(HRI '22)*. IEEE Press.

[4] Ismail Celik. 2023. Exploring the Determinants of Artificial Intelligence (AI) Literacy: Digital Divide, Computational Thinking, Cognitive Absorption. *Telematics and Informatics* 83 (Sept. 2023).

[5] Jessie Chen, Katelyn Procci, Michael Boyce, Julia Wright, Andre Garcia, and Michael Barnes. 2014. *Situation Awareness–Based Agent Transparency.* Technical Report. US Army Research Laboratory.

[6] European Commission, Content Directorate-General for Communications Networks, and Technology. 2019. *Ethics guidelines for trustworthy AI.* Publications Office.

[7] Saad Elbeleidy, Terran Mott, Dan Liu, Ellen Yi-Luen Do, Elizabeth Reddy, and Tom Williams. 2023. Beyond the Session: Centering Teleoperators in Robot-Assisted Therapy Reveals the Bigger Picture. *Proceedings of the ACM on Human-Computer Interaction (CSCW)* (2023).

[8] Saad Elbeleidy, Terran Mott, and Tom Williams. 2022. Practical, ethical, and overlooked: Teleoperated socially assistive robots in the quest for autonomy. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 577–587.

[9] Saad Elbeleidy, Daniel Rosen, Dan Liu, Aubrey Shick, and Tom Williams. 2021. Analyzing Teleoperation Interface Usage of Robots in Therapy for Children with Autism. In *Proceedings of the ACM Interaction Design and Children Conference*.

[10] Hideki Garcia, Katie Winkle, Tom Williams, and Megan Strait. 2023. Victims and Observers: How Gender, Victimization Experience, and Biases Shape Perceptions of Robot Abuse. In *IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*.

[11] Alyssa Hanson, Nichole Starr, Cloe Emnett, Ruchen Wen, Bertram Malle, and Tom Williams. 2024. The Power of Advice: Differential Blame for Human and Robot Advisors and Deciders in a Moral Advising Context. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*.

[12] Drew Hemment, Morgan Currie, Sj Bennett, Jake Elwes, Anna Ridler, Caroline Sinders, Matjaz Vidmar, Robin Hill, and Holly Warner. 2023. AI in the Public Eye: Investigating Public AI Literacy Through AI Art. In *2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Chicago IL USA.

[13] Ryan Blake Jackson, Ruchen Wen, and Tom Williams. 2019. Tact in Noncompliance: The Need for Pragmatically Apt Responses to Unethical Commands. In *Proc. AI, Ethics, and Society (AIES)*.

[14] Ryan Blake Jackson and Tom Williams. 2019. Language-Capable Robots may Inadvertently Weaken Human Moral Norms. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*.

[15] Ryan Blake Jackson and Tom Williams. 2021. A Theory of Social Agency for Human-Robot Interaction. *Frontiers in Robotics and AI* (2021).

[16] Ryan Blake Jackson, Tom Williams, and Nicole Smith. 2020. Exploring the Role of Gender in Perceptions of Robotic Noncompliance. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*.

[17] Malte F. Jung, Nikolas Martelaro, and Pamela J. Hinds. 2015. Using Robots to Moderate Team Conflict: The Case of Repairing Violations. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*.

[18] Malte F. Jung, Nikolas Martelaro, and Pamela J. Hinds. 2015. Using Robots to Moderate Team Conflict: The Case of Repairing Violations. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (Portland, Oregon, USA) *(HRI '15)*. Association for Computing Machinery, New York, NY, USA, 229–236.

[19] Johannes Kraus, Franziska Babel, Philipp Hock, Katrin Hauber, and Martin Baumann. 2022. The trustworthy and acceptable HRI checklist (TA-HRI): questions and design recommendations to support a trust-worthy and acceptable design of human-robot interaction. *Gruppe. Interaktion. Organisation. Zeitschrift für Angewandte Organisationspsychologie (GIO)* (2022).

[20] Minae Kwon, Malte F. Jung, and Ross A. Knepper. 2016. Human expectations of social robots. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*.

[21] Duri Long and Brian Magerko. 2020. What is AI Literacy? Competencies and Design Considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA.

[22] Terran Mott, Aaron Fanganello, and Tom Williams. 2024. What a Thing to Say! Which Linguistic Politeness Strategies Should Robots Use in Non-compliance Interactions. In *Proceedings of the ACM/IEEE Conference on Human Robot Interaction.*

[23] Terran Mott and Tom Williams. 2023. Community Futures With Morally Capable Robotic Technology. In *Workshop on Perspectives on Moral Agency in Human-Robot Interaction at HRI.*

[24] Terran Mott and Tom Williams. 2023. Confrontation and Cultivation: Understanding Perspectives on Robot Responses to Norm Violations. In *IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN).*

[25] Terran Mott and Tom Williams. 2023. Rube-Goldberg Machines, Transparent Technology, and the Morally Competent Robot. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction* (Stockholm, Sweden) *(HRI '23)*. Association for Computing Machinery, New York, NY, USA, 634–638.

[26] Davy Tsz Kit Ng, Jac Ka Lok Leung, Samuel Kai Wah Chu, and Maggie Shen Qiao. 2021. Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence* 2 (2021).

[27] Aditi Ramachandran, Sarah Strohkorb Sebo, and Brian Scassellati. 2019. Personalized Robot Tutoring Using the Assistive Tutor POMDP (AT-POMDP). In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence* (Honolulu, Hawaii, USA) *(AAAI'19/IAAI'19/EAAI'19)*. AAAI Press, Article 987, 8 pages.

[28] Avi Rosenfeld and Ariella Richardson. 2019. Explainability in Human–Agent Systems. *Autonomous Agents and Multi-Agent Systems* (2019).

[29] Matthias Scheutz, T. Williams, Evan A. Krause, Bradley Oosterveld, Vasanth Sarathy, and Tyler M. Frasca. 2018. An Overview of the Distributed Integrated Cognition Affect and Reflection DIARC Architecture. *Intelligent Systems, Control and Automation: Science and Engineering* (2018).

[30] Solace Shen, Petr Slovak, and Malte F. Jung. 2018. "Stop. I See a Conflict Happening.": A Robot Mediator for Young Children's Interpersonal Conflict Resolution. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI).*

[31] Sebastian Wallkötter, Silvia Tulli, Ginevra Castellano, Ana Paiva, and Mohamed Chetouani. 2021. Explainable Embodied Agents Through Social Cues: A Review. *J. Hum.-Robot Interact.* (2021).

[32] Tom Williams, Ryan Jackson, and Jane Lockshin. 2018. A Bayesian Analysis of Moral Norm Malleability during Clarification Dialogues. In *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci).*

[33] Tom Williams, Cynthia Matuszek, Ross Mead, and Nick Depalma. 2023. Scarecrows in Oz: The Use of Large Language Models in HRI. *ACM Transactions on Human-Robot Interaction* (01 2023).

[34] Katie Winkle, Ryan Blake Jackson, Gaspar Isaac Melsión, Dražen Bršcić, Iolanda Leite, and Tom Williams. 2022. Norm-Breaking Responses to Sexist Abuse: A Cross-Cultural Human Robot Interaction Study. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI).*

[35] Katie Winkle, Gaspar Isaac Melsión, Donald McMillan, and Iolanda Leite. 2021. Boosting Robot Credibility and Challenging Gender Norms in Responding to Abusive Behaviour: A Case for Feminist Robots. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI).*

[36] Robert H Wortham, Andreas Theodorou, and Joanna J Bryson. 2016. What Does the Robot Think? Transparency as a Fundamental Design Requirement for Intelligent Systems. In *Proceedings of the IJCAI Workshop on Ethics for Artificial Intelligence.*