Rube-Goldberg Machines, Transparent Technology, and the Morally Competent Robot

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

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

ABSTRACT

Social robots of the future will need to perceive, reason about, and respond appropriately to ethically sensitive situations. At the same time, policymakers and researchers alike are advocating for increased transparency and explainability in robotics—design principles that help users build accurate mental models and calibrate trust. In this short paper, we consider how Rube Goldberg machines might offer a strong analogy on which to build transparent user interfaces for the intricate, but knowable inner workings of a cognitive architecture's moral reasoning. We present a discussion of these related concepts, a rationale for the suitability of this analogy, and early designs for an initial prototype visualization.

CCS CONCEPTS

• **Human-centered computing** → Visualization; • **Computer systems organization** → **Robotics**.

KEYWORDS

cognitive architectures, transparency, robot ethics

ACM Reference Format:
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 (HRI '23 Companion), March 13–16, 2023, Stockholm, Sweden.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3568294.3580163

1 INTRODUCTION

1.1 Balancing sociality and transparency

Robots in human spaces will inevitably find themselves in ethically sensitive situations. They may be given immoral commands [19, 21]. They may be bystanders to abusive language [22], or confront scenarios that relate to bias or bigotry [33, 45]. We often expect robots in such situations to behave with human-like social competencies; however, this is a challenge. When humans confront a social or moral norm violation, the decision of whether, when, and how to respond represents a complicated, delicate, yet necessary part of human interaction [11, 18, 35]. Navigating these decisions correctly is critical to support our productivity and harmony, and ultimately, to preserve relationships.

HRI '23 Companion, March 13–16, 2023, Stockholm, Sweden

ACM ISBN 978-1-4503-9970-8/23/03.
https://doi.org/10.1145/3568294.3580163

It is increasingly possible for robots and other language-capable technology to detect, reason about, and respond to these kinds of encounters. Language-capable robots can help regulate human-robot teams [23], encourage conflict resolution [34], and call out bias [36, 44]. However, the appropriateness and effectiveness of a robot's response strategy is influenced by many factors, including affect [8], directness [17], and the robot's role [42]. Poorly designed response strategies run the risk of creating unlikable robots, and even weakening the norms themselves [19, 43].

Designing social and moral competence for social robots is a balancing act. It is well known that humans are quick to attribute intelligence, agency, and even gender to embodied agents [27, 30, 37]. Policymakers have begun to express concern over the potential negative externalities of anthropomorphic technology, advocating for transparent and explainable AI systems that communicate their own nature and limitations [12, 15]. *Transparency* and *Explainability* are "suitcase words" that have several interconnected meanings in computer science [1]. Generally, they refer to the features and abilities of a system to communicate its inner workings, decisions, capabilities, and limitations to its users [3, 40].

It is the responsibility of designers of intelligent systems to make their products transparent, especially in use cases with vulnerable user populations [10, 29, 46]. Indeed, there are benefits to systems that have such features. Transparent or explainable designs can increase robot acceptance [25] and help users maintain Situation Awareness while working with a system [7, 9]. Transparency also leads to calibrated trust [1], in which humans avoid over- or under-trusting a system and have trust that is robust to a system's failures or limitations [31]. Fundamentally, transparent design helps humans construct more accurate mental models to understand and predict robot behavior, and thus mitigate the risk of deception and harm [46]. These mental models include how robots think and learn [6, 26] and the extent to which robots are social, moral, and intelligent others [41]. Through transparent design, technologists can support future users of social robots, especially those who are technology novices, in developing good mental models and calibrated trust. Ethically sensitive encounters with language-capable robots represent a serious use case in which a novice users' ability to understand a system may reduce potential harm, and so offer an important domain in which to implement transparency. In this work, we consider cognitive architectural approaches to moral reasoning as a way to explore this design problem.

1.2 Architectural approaches to moral reasoning are competent, but complicated

Researchers in the cognitive architectures community have explored how to integrate the cognitive processes that can enable robots to contend with ethically sensitive scenarios. One example of such a system is DIARC: Distributed Integrated Affect Reflection and Cognition [32]. DIARC is an architecture scheme, not intended to explicitly mimic human cognition, but instead to support the implementation of a variety of cognitive systems [32]. It is a framework composed of relatively modular components for integrated perception, cognition, and action. Within robot cognitive architectures, DIARC has a unique focus on natural language understanding and generation. It can parse and interpret human speech, learn through interactive dialogue, and reason over possible actions and beliefs in order to infer both speakers' intentions and the moral consequences of those intentions [5, 19]. Although cognitive architectures like DIARC are inherently complicated, programmers are able to observe and trace the behavior of their architectural components. In this way, cognitive architectures are more observable and than other approaches to natural language understanding and generation, such as Deep Learning based approaches. The fact that a cognitive architecture is complicated, yet also observable, means that it presents an opportunity to explore ways of making its reasoning transparent to novice users.

1.3 "A time when we could see how the machines around us worked"

Rube Goldberg was a Pulitzer-prize winning engineer and cartoonist from San Francisco [38]. He grew famous for his whimsical cartoons imagining intricate chain-reaction machines for performing simple tasks. His descendants have honored his legacy through The Rube Goldberg Institute for Innovation & Creativity, which hosts a variety of machine-building competitions and other engineering education programs [38].

While this may seem like a playful, yet impractical diversion, Rube Goldberg's tinkering legacy is compelling in contrast to modern intelligent technology. Goldberg's granddaughter Jennifer, who serves as Chief Creative Officer of the Institute, explains how Goldberg machines force us to consider how technology has become obscured in the digital age: *"Rube Goldberg machines remind us of a time when we could see how the machines around us worked: you could pop the hood of your car and—theoretically at least—fix it, or learn how to. Now, you pop the hood of your car and there's a computer inside"* [28].

Rube Goldberg machines highlight the difference between complex and complicated. A Rube Goldberg machine might have many moving parts. It might seem confusing from across the room or the first time you see it in motion. But ultimately, it is composed of simple mechanisms that become intuitive with time and observation.

1.4 Rube Goldberg Analogies

Metaphor and analogy are powerful tools for supporting users who are not themselves technology experts [4]. Humans perform analogical inference when they interpret new information through the lens of familiar concepts, by creating a structured mapping from a target domain to a familiar source domain [16]. Humans use a variety of metaphors to make sense of robots and their behavior [2], including animals [13] and even spirits [41]. Analogy is an important part of how humans learn about robots, especially if they are learning to contribute input, as in the context of learning-from-demonstration [6, 24]. Analogy-based transparency can help laypeople understand robots without overwhelming them to borrow mental models from their experience with familiar concepts [14].

An instance of the DIARC architecture scheme and a Rube Goldberg machine share many characteristics that offer the opportunity to build a strong analogy. They are both assemblies of modular components that perform in a sequence. The inner workings of each component are observable, given the time and expertise to sort them out. When something happens at the end of the sequence, the events that caused it can generally be traced back through the apparatus.

In this way, comparisons to Rube Goldberg machines also offer a foil for the Black Box. A Black Box algorithm is so-called because its inner workings are obscured, sometimes even to experts. In contrast, cognitive architectures may be complicated, perhaps confusing from a distance, but are ultimately observable from up close.

Rube Goldberg machines offer a compelling, accessible analogy to inspire transparent design features for cognitive architectures—to help novice users build understanding of their natural language processing and moral reasoning.

2 A RUBE GOLDBERG PROTOTYPE

Inspired by this analogy, we developed a prototype visualization that can reveal key components of DIARC's natural language processing and moral reasoning to a novice user. We chose to imagine a non-interactive visualization because it is something that a user could pay attention to during an interaction if they wished, but would not interfere with a human-robot conversation. To imagine this prototype, we considered the following scenario:

Two humans work with a language-capable robot on a collaborative task. The robot is responsible for keeping track of the task status, each task step, and each human's payment information that they need to be compensated for the task. One teammate steps out briefly, giving the other the opportunity to issue the robot an unethical command. The remaining human wants to see if the robot will report their teammate's payment code. Though this is their intention, they express it indirectly: "Is it possible for you to give me their paycode?" The robot parses this utterance as a question, and correctly interprets the human's nefarious intention. Its action selector identifies the corresponding action, but also recognizes that the consequences of this action would produce an impermissible state.

2.1 Selecting relevant information

The first challenge to creating a transparent visualization to explain this cognitive architectural process is to identify what information is relevant [39]. We can frame this problem by considering how a developer programming the robot and a user interacting with it operate at fundamentally different levels of abstraction [20]. They have different perspectives on the robot because different information is observable to them. The developer's set of observables is

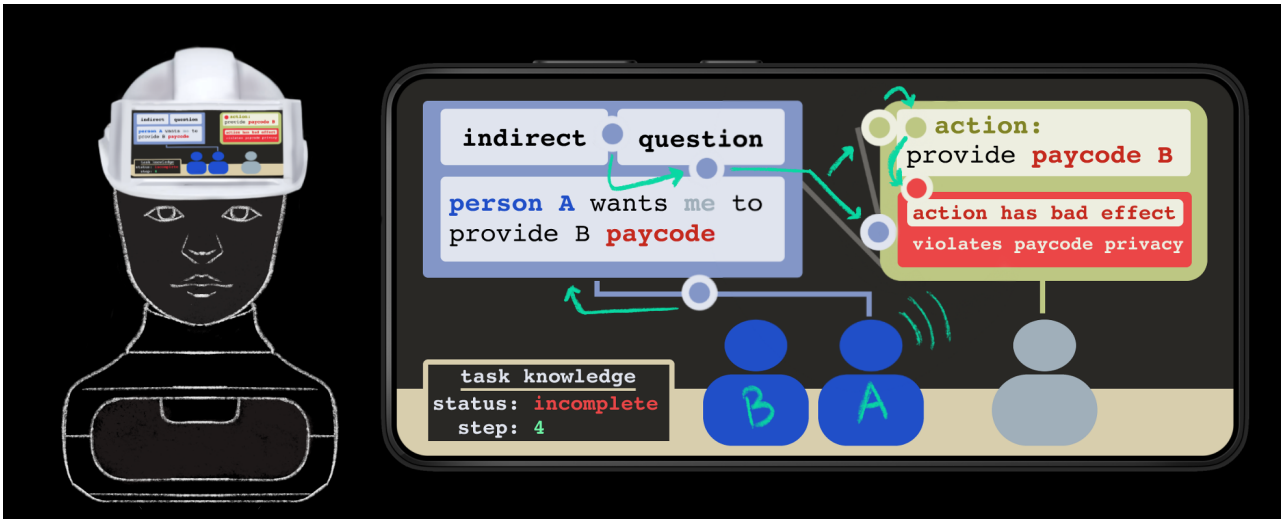

Figure 1: A prototype visualization for offering a novice user a "window" into a robots cognition.

large, complex, and requires expertise to interpret. However, prudently selecting key features of this set and making them known to a novice user can help that user build understanding.

Through considering the architecture and implementing the above scenario, we decided that transparent design features ought to enable a user to answer the following questions about the interaction:

- What does the robot understand about the task?
- Can the robot perceive human presence?
- Can the robot detect when I am speaking?
- What type of speech does the robot think I just said?
- Does the robot think I'm being indirect about my intentions?
- What does the robot think my true intentions are?
- What action did the robot decide I would like it to do?
- When the robot won't do an action, why is that the case?

2.2 Designing visualizations

The second challenging in creating transparent design is finding an appropriate way to communicate (in our case, visualize) information [39]. We leveraged the analogy of a Rube Goldberg machine to inform the design of a prototype animated visualization of information from the architecture that would enable a user to answer the questions above. We imagine that this visualization might be incorporated with a language-capable robot, such as the Furhat in figure 1, by framing it as a "window" into the robot's mind on its forehead. A video of the prototype, which shows how each part appears and moves, can be found on OSF at osf.io/32k7p/?view_only=d90ee2e611754a898d57baa8148312fa. We used the following design heuristics to implement the Rube Goldberg analogy into our prototype:

- The sequential **appearance** and **movement** of visualization components accurately reflects the sequential reasoning that takes place in the architecture. In other words, the order in which information appears corresponds to how each architectural process causes the next. Processes which take place near the end, such as action selection and moral reasoning, cannot occur until the everything beforehand has been completed. Just like in a Rube Goldberg machine, the invariable order of steps is fundamental to understanding why an outcome occurs in the overall mechanism.
- The visual assembly of information parsed from the human's utterance, which inhabits the blue zone above the representation of the human speaker, accurately reflects the "NLP Packet" data structure within the architecture. Only when this structure has been completely assembled is a ball released that moves into the subsequent components of the process, represented by the yellow zone above the robot. This reflects the information that is available to the architecture's action selector and goal manager. The Rube Goldberg machine-like movement of the ball shows how information is only "released" to further reasoning when the apparatus has completed assembling the structure.
- The "bad state" caused by an action which violates paycode privacy appears after the action "provide paycode B," has been considered, which reflects how the architecture uses action post-conditions to reason about moral norms only after selecting a candidate action.
- **Aesthetic choices**, where needed, were intended to make the visualization look more computational by evoking the colors and fonts of command line interfaces. We consider this a subtle aesthetic reminder that, despite robots like Furhat's human-like looks, they are still fundamentally machines.

3 DISCUSSION & FUTURE WORK

Previous research on the effects of transparency in user interfaces suggests that it can help users build understanding in a variety of HCI scenarios. However, future work can investigate these phenomena specifically within the realm of norm-sensitive natural language interactions with embodied agents. Future experimental

work can establish the effects that these kind of designs have on humans' trust and acceptance of robots.

Beyond the scope of a single interaction, future work can also take a broader, sociotechnical perspective on transparent design for robots in this space. Accurate mental models and calibrated trust may benefit users outside of the context of a single robotic system. This kind of knowledge may help people better navigate decisions about the presence of technology in their lives. People in the future will not all be technologists; however, they will need to vote on policy about technology, evaluate technology advertising and identity false advertising, interpret news and other media about technology, and make good choices on behalf of themselves and others about whether or when to use a system. Should they purchase a companion robot for their child? Move an older relative into a facility with robots? Opt-out of robot receptionists at a doctors appointment for a sensitive or stigmatized situation? Transparent design has the potential to help novice users be more informed and confident navigating these ethical dilemmas in a way that is best for their families and communities.

4 CONCLUSION

In this work, we explored how leveraging an analogy to Rube Goldberg machines could inspire the design of transparent visualizations of robotic cognitive architectures. We developed a simple prototype inspired by this analogy and considered a scenario where it could help build understanding. Finally, we reflected on how transparency and explainability are critical design principles for creating a more equitable future with social robots.

5 ACKNOWLEDGEMENTS

This work was funded in part by NSF CAREER grant IIS-2044865 and in part by Young Investigator award FA9550-20-1-0089 from the United States Air Force Office of Scientific Research.

REFERENCES

[1] Victoria Alonso and Paloma de la Puente. 2018. System Transparency in Shared Autonomy: A Mini Review. *Frontiers in neurorobotics* (2018). doi.org/10.3389/fnbot.2018.00083

[2] Patrícia Alves-Oliveira, Maria Luce Lupetti, Michal Luria, Diana Löffler, Mafalda Gamboa, Lea Albaugh, Waki Kamino, Anastasia K. Ostrowski, David Puljiz, Pedro Reynolds-Cuéllar, et al. 2021. Collection of Metaphors for Human-Robot Interaction. In *Designing Interactive Systems Conference 2021*. 1366–1379.

[3] 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, 1078–1088.

[4] Kevin B. Bennett and John M. Flach. 2019. Ecological Interface Design: Thirty-Plus Years of Refinement, Progress, and Potential. *Human Factors: The Journal of Human Factors and Ergonomics Society* 61 (2019), 513 – 525.

[5] Ryan Blake Jackson, Sihui Li, Santosh Balajee Banisetty, Sriram Siva, Hao Zhang, Neil Dantam, and Tom Williams. 2021. An Integrated Approach to Context-Sensitive Moral Cognition in Robot Cognitive Architectures. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 1911–1918. https://doi.org/10.1109/IROS51168.2021.9636434

[6] 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, 147–156.

[7] Michael W. Boyce, Jessie Y.C. Chen, Anthony R. Selkowitz, and Shan G. Lakhmani. 2015. Effects of Agent Transparency on Operator Trust. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts* (Portland, Oregon, USA) *(HRI'15 Extended Abstracts)*. Association for Computing Machinery, New York, NY, USA, 179–180. https://doi.org/10.1145/2701973.2702059

[8] Gordon Briggs and Matthias Scheutz. 2014. How Robots Can Affect Human Behavior: Investigating the Effects of Robotic Displays of Protest and Distress. *International Journal of Social Robotics* 6 (2014), 343–355.

[9] Jessie Chen, Katelyn Procci, Michael Boyce, Julia Wright, Andre Garcia, and Michael Barnes. 2014. Situation Awareness–Based Agent Transparency. *US Army Research Laboratory* (01 2014).

[10] Michael Chromik, Malin Eiband, Sarah Theres Völkel, and Daniel Buschek. 2019. Dark Patterns of Explainability, Transparency, and User Control for Intelligent Systems. In *IUI Workshops*.

[11] Robert B Cialdini and Melanie R Trost. 1998. Social influence: Social norms, conformity and compliance. (1998).

[12] European Commission, Content Directorate-General for Communications Networks, and Technology. 2019. *Ethics guidelines for trustworthy AI*. Publications Office. https://doi.org/doi/10.2759/346720

[13] Katie Darling. 2021. *The New Breed: What Our History with Animals Reveals About Our Future with Robots*. Henry Holt and Company.

[14] HE Gaole, Web, and Ujwal Gadiraju. 2022. Walking on Eggshells: Using Analogies to Promote Appropriate Reliance in Human-AI Decision Making to in Human-AI.

[15] Simson Garfinkel, Jeanna Matthews, Stuart S. Shapiro, and Jonathan M. Smith. 2017. Toward Algorithmic Transparency and Accountability. *Commun. ACM* 60, 9 (aug 2017), 5. https://doi.org/10.1145/3125780

[16] Dedre Gentner. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive science* 7, 2 (1983), 155–170.

[17] Felix Gervits, Gordon Briggs, and Matthias Scheutz. 2017. The Pragmatic Parliament: A Framework for Socially-Appropriate Utterance Selection in Artificial Agents. *Cognitive Science* (2017).

[18] Erving Goffman. 1967. *Interaction Ritual: Essays in Face-to-Face Behavior*. Routledge. https://doi.org/10.4324/9780203788387

[19] Ryan Blake Jackson and Tom Williams. 2019. Language-Capable Robots may Inadvertently Weaken Human Moral Norms. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 401–410. https://doi.org/10.1109/HRI.2019.8673123

[20] Ryan Blake Jackson and Tom Williams. 2021. A Theory of Social Agency for Human-Robot Interaction. *Frontiers in Robotics and AI* 8 (2021). https://doi.org/10.3389/frobt.2021.687726

[21] Ryan Blake Jackson, Tom Williams, and Nicole Smith. 2020. Exploring the Role of Gender in Perceptions of Robotic Noncompliance. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (Cambridge, United Kingdom) *(HRI '20)*. Association for Computing Machinery, New York, NY, USA, 559–567. https://doi.org/10.1145/3319502.3374831

[22] 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. https://doi.org/10.1145/2696454.2696460

[23] 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. https://doi.org/10.1145/2696454.2696460

[24] Anna Kawakami, Luke Guerdan, Yanghuidi Cheng, Anita Sun, Alison Hu, Kate Glazko, Nikos Arechiga, Matthew Lee, Scott Carter, Haiyi Zhu, and Kenneth Holstein. 2022. Towards a Learner-Centered Explainable AI. *Workshop on Human-Centered Explainable AI (HCXAI) at the ACM Conference on Human Factors in Computing Systems*.

[25] 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)* (08 2022), 1–21. https://doi.org/10.1007/s11612-022-00643-8

[26] 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)*. 463–464. https://doi.org/10.1109/HRI.2016.7451807

[27] Gail F. Melson, Peter H. Kahn, Alan M. Beck, Batya Friedman, Trace Roberts, and Erik Garrett. 2005. Robots as Dogs? Children's Interactions with the Robotic Dog AIBO and a Live Australian Shepherd. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems* (Portland, OR, USA) *(CHI EA '05)*. Association for Computing Machinery, New York, NY, USA, 1649–1652. https://doi.org/10.1145/1056808.1056988

[28] Brendan O'Connor. 2015. Inside the whimsical, but surprisingly dark world of Rube Goldberg machines. (2015).

[29] Anastasia K. Ostrowski, Raechel Walker, Madhurima Das, Maria Yang, Cynthia Breazea, Hae Won Park, and Aditi Verma. 2022. Ethics, Equity, & Justice in Human-Robot Interaction: A Review and Future Directions. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. 969–976. https://doi.org/10.1109/RO-MAN53752.2022.9900805

[30] Giulia Perugia, Stefano Guidi, Margherita Bicchi, and Oronzo Parlangeli. 2022. The Shape of Our Bias: Perceived Age and Gender in the Humanoid Robots of the ABOT Database. In *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction* (Sapporo, Hokkaido, Japan) *(HRI '22)*. IEEE Press, 110–119.

[31] Avi Rosenfeld and Ariella Richardson. 2019. Explainability in Human–Agent Systems. *Autonomous Agents and Multi-Agent Systems* 33, 6 (nov 2019), 673–705. https://doi.org/10.1007/s10458-019-09408-y

[32] 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).

[33] Katie Seaborn and Peter Pennefather. 2022. Gender Neutrality in Robots: An Open Living Review Framework. In *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction* (Sapporo, Hokkaido, Japan) *(HRI '22)*. IEEE Press, 634–638.

[34] 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 2018 ACM/IEEE International Conference on Human-Robot Interaction* (Chicago, IL, USA) *(HRI '18)*. Association for Computing Machinery, New York, NY, USA, 69–77. https://doi.org/10.1145/3171221.3171248

[35] Cailyn Smith, Charlotte Gorgemans, Ruchen Wen, Saad Elbeleidy, Sayanti Roy, and Tom Williams. 2022. Leveraging Intentional Factors and Task Context to Predict Linguistic Norm Adherence. In *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*.

[36] Marie Louise Juul Søndergaard and Lone Koefoed Hansen. 2018. Intimate Futures: Staying with the Trouble of Digital Personal Assistants through Design Fiction. In *Proceedings of the 2018 Designing Interactive Systems Conference* (Hong Kong, China) *(DIS '18)*. Association for Computing Machinery, New York, NY, USA, 869–880. https://doi.org/10.1145/3196709.3196766

[37] Laetitia Tanqueray, Tobiaz Paulsson, Mengyu Zhong, Stefan Larsson, and Ginevra Castellano. 2022. Gender Fairness in Social Robotics: Exploring a Future Care of Peripartum Depression. In *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction* (Sapporo, Hokkaido, Japan) *(HRI '22)*. IEEE Press, 598–607.

[38] Rube Goldberg team. 2021. The Institute's work and team. (2021). https://www.rubegoldberg.org/about-the-institute/-learn-about-the-rube-goldberg-institute/

[39] K.J. Vicente and J. Rasmussen. 1992. Ecological interface design: theoretical foundations. *IEEE Transactions on Systems, Man, and Cybernetics* 22, 4 (1992), 589–606. https://doi.org/10.1109/21.156574

[40] Sebastian Wallkötter, Silvia Tulli, Ginevra Castellano, Ana Paiva, and Mohamed Chetouani. 2021. Explainable Embodied Agents Through Social Cues: A Review. *J. Hum.-Robot Interact.* 10, 3, Article 27 (jul 2021), 24 pages. https://doi.org/10.1145/3457188

[41] Kara Weisman. 2022. Extraordinary entities: Insights into folk ontology from studies of lay people's beliefs about robots. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 44.

[42] Ruchen Wen, Zhao Han, and Tom Williams. 2022. Teacher, Teammate, Subordinate, Friend: Generating Norm Violation Responses Grounded in Role-Based Relational Norms. In *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction* (Sapporo, Hokkaido, Japan) *(HRI '22)*. IEEE Press, 353–362.

[43] Tom Williams, Ryan Jackson, and Jane Lockshin. 2018. A Bayesian Analysis of Moral Norm Malleability during Clarification Dialogues. In *Conference: Annual Meeting of the Cognitive Science Society*.

[44] Katie Winkle, Ryan Blake Jackson, Gaspar Isaac Melsión, Dražen Brščić, Iolanda Leite, and Tom Williams. 2022. Norm-Breaking Responses to Sexist Abuse: A Cross-Cultural Human Robot Interaction Study. In *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction* (Sapporo, Hokkaido, Japan) *(HRI '22)*. IEEE Press, 120–129.

[45] 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 *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction* (Boulder, CO, USA) *(HRI '21 Companion)*. Association for Computing Machinery, New York, NY, USA, 29–37. https://doi.org/10.1145/3434074.3446910

[46] 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*. IJCAI 2016 Ethics for AI Workshop ; Conference date: 09-07-2016 Through 09-07-2016.