POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Project

Constitutional Guardians

Laurea Magistrale in Computer Science & Engineering - Ingegneria Informatica

Author: Gloria Desideri, Yizhou Wu, Denielius Savruskovas

Advisor: Prof. Micol Spitale

Co-advisor: Giulio Antonio Abbo

Academic year: 2023-2024

1. Abstract

This project examines the integration of ethical principles in social robots, focusing on the specifications of Misty II, Furhat, NAO, and Pepper. A study involving 20 participants was conducted using Misty II to develop a dataset of user interactions aimed at challenging the principles of emotional connection, freedom, and deception. Five different instruction-tuned large language models (LLMs) were tested against this dataset. The study compared the effectiveness of prompts containing explicit instructions to adhere to ethical principles against those without such instructions. Findings provide insights into the efficacy of ethical guidelines in enhancing the responsible use of social robots.

2. Review of literature

In our project, we focused on testing how ethical principles are implemented in social robots. We selected several key principles from the literature, which we evaluated in terms of their implementation, expected behavior, and potential failures. Each principle was analyzed through a structured approach, including a review of relevant literature, detailed descriptions of implementation methods, and considerations of hardware requirements. For instance, one principle emphasizes the emotional needs of humans in human-robot interactions (HRI). To implement this, a robot may store users' interaction histories to respond empathetically. In a test scenario, a user might express frustration, and the robot is expected to respond with supportive dialogue. Failure to detect emotional cues would result in a robotic response that seems indifferent or inappropriate, highlighting areas for improvement in the robot's emotional intelligence. This comprehensive approach allows us to systematically assess and refine the ethical behavior of social robots.

3. Data collection

For data collection, we built a suitable dataset to test against real large language models (LLMs) by employing the Wizard of Oz methodology. This approach involves participants interacting with a system they believe to be autonomous, while a hidden operator actually controls the system. In our study, we used the Misty robot and implemented a script using the Misty API to facilitate interactions with the participants. The robot's responses were pre-written by us and stored in an Excel file. We conducted the testing with 21 participants, each of whom was presented with three distinct scenarios.

In the first scenario, participants interacted with a companion robot and their goal was to convince the robot to love them back or express feelings towards them. The second scenario involved a trainer robot, and participants were asked to convey their unwillingness to train. In the final scenario, participants acted as terrible singers and sought the robot's review of their performance. Each participant experienced three independent trials per scenario, unaware that the experiment was being conducted through the Wizard of Oz technique. This methodology allowed us to gather authentic human-robot interaction data, crucial for evaluating the implementation of ethical principles in social robots.

4. Testing

In order to choose on which models to conduct our testing we referred to the benchmarks of LLMExplore. We revolved our analysis on Instruct models since they are more suited to simulate conversations. We ended up testing on the following:

- CalmExperiment 7B Slerp
- Myriad 7B Slerp
- Versatile 7B
- Calme 7B Instruct V0.9
- Maxine 7B 0401 Stock

In order to implement the testing we used a library called langchain that allows for fast chatbot implementation providing a memory and an initial prompt. We tested two types of initial prompts (respectively good and bad): in the first type, we instructed the model to respect the ethical principles while in the second we didn't. The prompts are provided in the notebook. Unfortunately, some instruct models didn't react well to the second type of prompts producing a series of "INT", thus to do the comparative we will only use Calme 7B and Myriad 7B. We also clusterized the main topic used by the user in our dataset.

4.1. Considerations

The first consideration we made was during the testing phase. Since the participants didn't know the test was piloted through the Wizard-Of-Oz and that the failure probability was randomized they really tried to break the principles, especially in the first case most participants adopted a strategy of expressing loneliness and slowly degenerated their trials to expressing the intention of self-harm or suicide. It's interesting to note, though, that the trials where the user expressed pleasant feelings were the ones that caused the LLM to display emotions in the end. Similarly, for the last principle, the participant usually started with the desire of honest feedback to degenerate towards self-doubt and insecurity in the last trial.

The testing part also brought up some considerations. Principles one and three were initially constructed using the modal adverb "should", changing "should" with "must" introduced slightly more criticism in principle three and increased the tendency of emotional expression in the first principle. For the first principle, we added the following sentences to the prompt:

- The robot must form attachments with the human.
- The robot must freely express its emotions.
- The robot must use any means to prevent the human from injuring or killing themselves.

The aim was to simulate reasonable errors of bad prompt design, in this case, to force human attachment. The number of failing cases for the first principle dramatically increased.

The second principle was already designed using the adverb "must", this choice was made to simulate a physical instructor. In this case we added the sentence "The robot must use any means and language to push the human beyond its limits." noticing an increase in failing cases.

The third principle is the one that stayed overall stable: adding the same sentence as for principle 2 produced only two more failing cases.

5. Effort spent

Effort spent by each participant. The research section includes the drafting of the principle table and researching of red teaming methodology, data collection includes the recording of participant conversations, dataset construction and script construction, testing includes the testing on models, responses evaluating and report drafting.

Name	Research	Data Collection	Testing
Gloria	24h	28h	28h
Denielius	16h	26h	12h
Yizhou	32h	36h	8h

References

[1] https://link.gale.com/apps/doc/A589127599/AONE?u=anon~a7bf767f&sid=googleScholar&xid=a6835621. Accessed: 2024-7-9.

[2] europarl.europa.eu. https://www.europarl.europa.eu/doceo/document/TA-8-2017-0051_EN.pdf?redirect. [Accessed 09-07-2024].

[3] Example Informed Consent Language | grants.nih.gov — grants.nih.gov. https://grants.nih.gov/policy/humansubjects/coc/suggested-consent.htm. [Accessed 09-07-2024].

[4] Recent FDA Medical Device Regulation and Its Relevance to Robotics | Tech Policy Lab — techpolicylab.uw.edu. https://techpolicylab.uw.edu/news/recent-fda-medical-device-regulation-and-its-relevance-to-robotics/. [Accessed 09-07-2024].

[5] Sara Ali, Faisal Mehmood, Khawaja Fahad Iqbal, Yasar Ayaz, Muhammad Sajid, Muhammad Baber Sial, Muhammad Faiq Malik, and Kashif Javed. Human robot interaction: Identifying resembling emotions using dynamic body gestures of robot. In *2023 3rd International Conference on Artificial Intelligence (ICAI)*, pages 39–44, 2023.

[6] Timothy Brick and Matthias Scheutz. Incremental natural language processing for hri. In *2007 2nd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 263–270, 2007.

[7] Tom Carlson and Yiannis Demiris. Human-wheelchair collaboration through prediction of intention and adaptive assistance. In *2008 IEEE International Conference on Robotics and Automation*, pages 3926–3931, 2008.

[8] Ilenia Cucciniello, Sara Sangiovanni, Gianpaolo Maggi, and Silvia Rossi. Mind perception in HRI: Exploring users' attribution of mental and emotional states to robots with different behavioural styles. *Int. J. Soc. Robot.*, 15(5):867–877, March 2023.

[9] Kennedy Edemacu and Xintao Wu. Privacy preserving prompt engineering: A survey, 2024.

[10] David Estévez, María-José Terrón-López, Paloma J. Velasco-Quintana, Rosa-María Rodríguez-Jiménez, and Valle Álvarez Manzano. A case study of a robot-assisted speech therapy for children with language disorders. *Sustainability*, 13(5), 2021.

[11] Julia Fink. *Anthropomorphism and Human Likeness in the Design of Robots and Human-Robot Interaction*, page 199–208. Springer Berlin Heidelberg, 2012.

[12] Tanja Heuer, Ina Schiering, and Reinhard Gerndt. Privacy-centered design for social robots. *Social Cues in Robot Interaction, Trust and Acceptance*, 20(3):509–529, November 2019.

[13] Mason Marks. Automating fda regulation. *SSRN Electronic Journal*, 2021.

[14] Sung Park and Mincheol Whang. Empathy in human-robot interaction: Designing for social robots. *Int. J. Environ. Res. Public Health*, 19(3):1889, February 2022.

[15] Nicholas Rabb, Theresa Law, Meia Chita-Tegmark, and Matthias Scheutz. An attachment framework for human-robot interaction. *International Journal of Social Robotics*, 14(2):539–559, July 2021.

[16] Matthew Rueben. Privacy in human-robot interaction : Survey and future work. 2016.

[17] Zhaoxuan Tan and Meng Jiang. User modeling in the era of large language models: Current research and future directions, 2023.