Honda Research Institute EU
**Honda Research Institute Europe GmbH**
https://www.honda-ri.de/

# Reducing Mental Model Mismatch with Intention-Based Feedback in Human-Robot Teaching

## Phillip Richter, Anna-Lisa Vollmer, Heiko Wersing

### 2024

**Preprint:**

Copyright ACM 2024. This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in Proc. Conf. on Human Agent Interaction, https://doi.org/10.1145/3687272.3690896.

# Reducing Mental Model Mismatch with Intention-Based Feedback in Human-Robot Teaching

Phillip Richter
Bielefeld University
Bielefeld, Germany
prichter@techfak.uni-bielefeld.de

Heiko Wersing
Honda Research Institute Europe
Offenbach am Main, Germany
heiko.wersing@honda-ri.de

Anna-Lisa Vollmer
Bielefeld University
Bielefeld, Germany
avollmer@techfak.uni-bielefeld.de

## ABSTRACT

This paper introduces the Mental Model Mismatch (MMM) Score, a feedback mechanism designed to align human teaching behavior with robot learning by quantifying mismatches between the human teacher's mental model and the robot's learning capabilities. Utilizing Large Language Models, the system analyzes teacher intentions in natural language to generate adaptive feedback. A study with 150 participants teaching a virtual robot demonstrates that intention-based feedback significantly improves the robots learning outcomes compared to traditional performance-based feedback or no feedback. The findings suggest that this approach improves understanding of the robot's learning process and reduces misconceptions to enhance human-robot interaction.

## CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in interaction design**; **HCI theory, concepts and models**; **User models**.

## KEYWORDS

human-robot interaction, interactive task learning, mental model, intention, feedback, evaluation metrics

**ACM Reference Format:**
Phillip Richter, Heiko Wersing, and Anna-Lisa Vollmer. 2924. Reducing Mental Model Mismatch with Intention-Based Feedback in Human-Robot Teaching. In *Proceedings of International Conference on Human-Agent Interaction (HAI '24).* ACM, New York, NY, USA, 3 pages. https://doi.org/XXXXXXX.XXXXXXX

## 1 INTRODUCTION

The rapid advancement of AI and robotics has profoundly transformed various aspects of our lives, automating tasks across industries and personal environments [1, 7, 10, 12]. However, a significant challenge remains in effectively teaching these intelligent systems, particularly robots [3, 6]. The vision of cooperative intelligence emphasizes the importance of mutual understanding between humans and intelligent systems to achieve synergy towards human-centred values and goals [9]. Humans often project their own experiences onto these systems, leading to mental model mismatches that can hinder the teaching process [2, 5, 8]. This discrepancy arises because intelligent systems learn through algorithms that differ fundamentally from human learning [4, 11], creating challenges in communication and understanding. Addressing this gap, we propose a novel approach: the *Mental Model Mismatch Score* (MMM) score, which quantifies the difference between a human's mental model and a robot's actual learning architecture. By integrating this understanding into the feedback loop during human-robot interactions, we aim to improve the effectiveness of teaching.

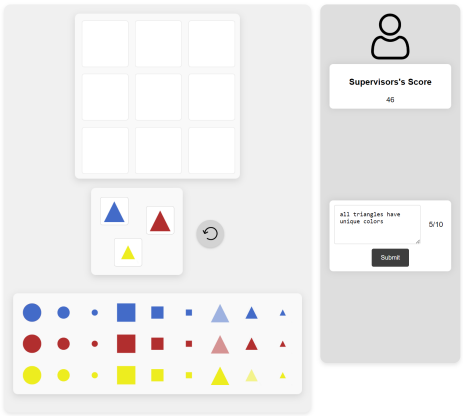


**Figure 1: Graphical User Interface used by participants in the Superdoku game to teach concepts to the virtual robot.**

## 2 METHOD

We leverage a Large Language Model to extract key terms from user intentions and map them to concepts learned by the robot. The MMM Score quantifies this alignment by evaluating how well the robot's learned concepts correspond to human intentions.

Concepts the robot can learn are represented as binary states within a set ($C$). Key terms from user instructions are compared against these concepts to determine alignment. The MMM Score for each iteration $d$ is calculated as:

$$S_d = 1 - \frac{\sum_{i=1}^{|C|} \mathrm{M}(c_i, \mathcal{K}) \cdot \mathrm{Learned}(c_i)}{N_{\mathrm{learned},d}},$$

where:

- $\mathrm{Learned}(c_i)$ is a binary indicator that equals 1 if concept $c_i$ is learned and 0 otherwise.
- $N_{\mathrm{learned},d}$ represents the total number of concepts learned in iteration $d$.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
*HAI '24, November 24–27, 2024, Swansea, United Kingdom*
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

To provide a comprehensive measure of alignment across multiple iterations, the cumulative mismatch score ($S_{\text{cum}}$) is calculated as:

$$S_{\text{cum},d} = \frac{S_d + S_{\text{cum},d-1} \cdot (d-1)}{d},$$

where $S_{\text{cum},d-1}$ is the cumulative score up to the previous iteration. This weighted approach ensures that each iteration equally contributes to the overall assessment, generating feedback on the alignment between human intentions and robot learning over time.
.

## 3 EXPERIMENT

We conducted a user study with 150 participants (50 per test group) to evaluate the MMM Score in a human-robot teaching scenario within a virtual environment. Participants taught a virtual robot to solve a puzzle game called *Superdoku*. To solve the game, the robot had to place the tokens on a 3x3 grid in such a way that each token appears only once in each row and column. Using a graphical user interface (see Figure 1), participants taught the robot by presenting a combination of three tokens. To maintain the experiment's integrity and control for unexpected variables, the robot's learning mechanism was simplified, allowing it to learn concepts directly from distinct, predetermined token combinations. Up to 13 concepts could be taught and when all concepts were taught the robot was always able to fully solve the puzzle. The participants also provided a brief explanation of their intention behind each token combination in natural language, which was used to calculate feedback with the MMM score. The participant's token combination and stated intention were processed by a supervisor, an artificial entity responsible for calculating the feedback. This setup should prevent the assumption that the robot had direct access to the verbalized intentions, as the robot only learns from the token combinations. Each participant completed 25 teaching iterations (token combination and verbalized intention). Feedback was given in two forms: Group-specific feedback after each iteration and graphical feedback, for each test group, every fifth iteration, showing the robot's current understanding by filling in a 3x3 grid. Group-specific feedback for each of the 3 test groups:

- **MMM Group:** Received numeric feedback based on the MMM score using verbalized intentions.
- **Performance Group:** Received numeric feedback based on whether a concept was learned: feedback increased if a concept was learned and decreased if not.
- **Control Group:** Received no numeric feedback.

In the MMM and Performance groups, numerical feedback was provided as a number between 0 and 100. This consistent feedback format ensured that only the method of feedback calculation varied between the groups.

## 4 RESULTS

Participants in the MMM group achieved an average score of 10.52 (SD = 2.37) concepts taught to the robot, higher than both the performance-based feedback group (9.54, SD = 3.42) and the control group (8.88, SD = 3.43). A one-sided independent samples t-test revealed that the MMM group performed significantly better than the performance-based feedback group ($t(98) = 1.66, p = .0496$) and the control group ($t(98) = 2.78, p = .0032$). There was no significant difference between the performance-based feedback group and the control group ($t(98) = 0.96, p = .1691$).

| Category | Absolute Score | Percentage of Maximum (%) |
|---|---|---|
| MMM | 10.52 | 80.92% |
| Performance | 9.54 | 73.39% |
| Control | 8.88 | 68.31% |

**Table 1: Final average scores per test group.**

Following positive feedback, participants in the MMM group showed higher success rates in teaching additional concepts in subsequent iterations compared to the performance-based feedback group, as seen in figure 2.




**Figure 2: Percentage of participants who taught at least once concept per iteration after receiving positive feedback.**

## 5 DISCUSSION

This study demonstrates that intention-based feedback significantly enhances teaching outcomes in human-robot interactions. As seen in table 1, participants in the MMM group achieved a higher average score compared to those in the performance-based feedback and control groups, indicating that feedback aligned with user intentions is more effective in facilitating learning.

The MMM Score also plays a crucial role in reducing misconceptions about the robot's learning behavior. As shown in figure 2, participants in the MMM group continued to teach more effectively in subsequent iterations after receiving positive feedback. This trend suggests that the MMM mechanism helps prevent the reinforcement of incorrect teaching strategies by ensuring that positive feedback is only provided when the robot learns the intended concept. In contrast, the performance-based group risked reinforcing misconceptions by giving positive feedback even when unintended learning outcomes were achieved. Thus, the MMM Score not only improves teaching effectiveness but also helps maintain an accurate understanding of the robot's learning process.

## ACKNOWLEDGMENTS

Phillip Richter gratefully acknowledges the financial support from Honda Research Institute Europe for the project 'Reducing Mental Model Mismatch for Cooperative Robot Teaching'.

## REFERENCES

[1] Iain M Cockburn, Rebecca Henderson, and Scott Stern. 2018. *The impact of artificial intelligence on innovation*. Vol. 24449. National bureau of economic research Cambridge, MA, USA.

[2] Lukas Hindemith, Jan Philip Göpfert, Christiane B Wiebel-Herboth, Britta Wrede, and Anna-Lisa Vollmer. 2021. Why robots should be technical: Correcting mental models through technical architecture concepts. *Interaction Studies* 22, 2 (2021), 244–279.

[3] Matthew B Hoy. 2018. Alexa, Siri, Cortana, and more: an introduction to voice assistants. *Medical reference services quarterly* 37, 1 (2018), 81–88.

[4] MohammadNoor Injadat, Abdallah Moubayed, Ali Bou Nassif, and Abdallah Shami. 2021. Machine learning towards intelligent systems: applications, challenges, and opportunities. *Artificial Intelligence Review* 54, 5 (2021), 3299–3348.

[5] Sau-lai Lee, Ivy Yee-man Lau, Sara Kiesler, and Chi-Yue Chiu. 2005. Human mental models of humanoid robots. In *Proceedings of the 2005 IEEE international conference on robotics and automation*. IEEE, 2767–2772.

[6] Benedikt G Mark, Erwin Rauch, and Dominik T Matt. 2021. Industrial assistance systems to enhance human–machine interaction and operator's capabilities in assembly. In *Implementing Industry 4.0 in SMEs: Concepts, Examples and Applications*. Springer International Publishing Cham, 129–161.

[7] Hiroyuki Nitto, Daisuke Taniyama, and Hitomi Inagaki. 2017. Social acceptance and impact of robots and artificial intelligence. *Nomura Research Institute papers* 211 (2017), 1–15.

[8] Scott Ososky. 2013. Influence of task-role mental models on human interpretation of robot motion behavior.(2013). *Unpublished Doctoral Dissertation, University of Central Florida, Retrieved from https://stars. library. ucf. edu/etd/2893* (2013).

[9] Bernhard Sendhoff and Heiko Wersing. 2020. Cooperative intelligence-a humane perspective. In *2020 IEEE International Conference on Human-Machine Systems (ICHMS)*. IEEE, 1–6.

[10] Kamran Shaukat, Farhat Iqbal, Talha Mahboob Alam, Gagandeep Kaur Aujla, Liton Devnath, Abdul Ghaffar Khan, Rimsha Iqbal, Irum Shahzadi, and Afifah Rubab. 2020. The impact of artificial intelligence and robotics on the future employment opportunities. *Trends in Computer Science and Information Technology* 5, 1 (2020), 050–054.

[11] Dirk Söffker. 2008. Interaction of intelligent and autonomous systems–part i: qualitative structuring of interaction. *Mathematical and Computer modelling of dynamical systems* 14, 4 (2008), 303–318.

[12] Michael Webb. 2019. The impact of artificial intelligence on the labor market. *Available at SSRN 3482150* (2019).

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009