Shall we kill the doctor? Hmm…the doctor? Haha: A Social Robot in a Moral Dilemma

Talha Bedir[†], **Vanessa Vanzan**[†]
[†]Department of Philosophy, Linguistics and Theory of Science (FLoV)
University of Gothenburg, Sweden
talha.bedir@gu.se
vanessa.vanzan@gu.se

Abstract

This project describes the development of a dialogue system designed to support a later study on how hesitation markers affect moral decision-making in human–robot interaction.

We implemented a revised version of the Balloon Task to develop a dialogue system for the Furhat robot, where participants will interact with the robot in a scenario that requires them to decide which character in the moral dilemma should be sacrificed to save the others.

The robot's response to the participant's decision varies across three conditions: a filled pause (e.g.,"hmm, the doctor?"), silence (e.g., "… the doctor?"), or social laughter (e.g.,"haha, the doctor?"). We investigate whether these markers, when embedded in a clarification request, may prompt participants to reconsider their decision. We anticipate that conditions with a filled pause and laughter will present more instances of participants changing their final decisions, suggesting that such cues can influence moral deliberation.

1 Introduction

Dialogue is shaped not only by words and complete sentences but also by subtle cues that guide how meaning unfolds. Among these, backchannels and disfluencies, including hesitation markers, filled pauses, and social laughter, serve important functions in managing turn-taking, promoting alignment, and signaling hesitation (Howes et al., 2017).

Additionally, disfluencies may play a crucial role on competence perceived by others. In a web-based task using synthesized dialogues, participants were significantly more likely to trust and select fluent speakers' answers over disfluent ones, which may suggest that fluency shapes decision-making and trust perceived by interlocutors (Kirkland and Edlund, 2025).

This project investigates how interactional cues that typically occur in human–human dialogue, when generated by a social robot, may affect human decision-making in a moral dilemma task.

To achieve that, we designed a dialogue system incorporating dislfuencies (pauses and filled pauses) and social laughter followed by a clarification request. We focus on whether hesitation expressed through disfluencies and laughter produced by the robot can influence moral decision-making. Specifically, we ask whether these hesitation cues might lead users to reconsider their choices in a moral dilemma task.

Our dialogue system uses the Balloon Task and it was chosen for its potential to trigger ethically charged reasoning.

In the dilemma, participants must decide which of four passengers (e.g., a Pilot, Teacher, Doctor, or Prodigy) should jump from a hot-air balloon to save the others.

The Furhat[1] robot conducts the dialogue, introducing each character and prompting the user to discuss the four possibilities and choose one person to sacrifice. When the participant states their decision (e.g., "The doctor should jump"), Furhat repeats the chosen name in one of three ways, corresponding to distinct experimental conditions.

2 Hesitation and Clarification Requests in Dialogue

Hesitation is not merely a delay in speech, but a multifaceted communicative phenomenon that can serve various interactional functions. Filled pauses, for instance, may indicate moments of processing difficulty or uncertainty (Clark and Fox Tree, 2002). When perceived by a listener, hesitation can invite reflection, display caution, or signal implicit dis-

[1] Furhat is a social robot developed in 2014 by Furhat Robotics, a spin-off from KTH Royal Institute of Technology, Sweden (www.furhatrobotics.com).

agreement, as delays and hesitations frequently precede dispreferred responses such as disagreement (Pomerantz, 1984). These characteristics make hesitation a particularly intriguing cue in moral reasoning tasks, where speakers must navigate conflicting values and weigh competing considerations.

Clarification requests play an important role in maintaining mutual understanding and updating the common ground in dialogue (Ginzburg and Cooper, 2004). They help resolve uncertainty about what was said or meant, enabling interlocutors to repair potential misunderstandings. When combined with hesitation (e.g., “Hmm, the doctor?”), they may convey additional interpersonal meaning, such as doubt, reflection, or the need for further justification.

3 Laughter, Furhat and Dialogue

Social robots such as Furhat make it possible to explore how hesitation and expressive behaviours in human–robot interaction compare to those in human–human communication. Laughter, in particular, plays a key role in establishing social connection and shared understanding. In conversational settings, it often co-occurs with disfluencies or hesitation, marking alignment, mitigation, or repair. In human–robot dialogue, shared laughter has been shown to enhance perceptions of empathy, engagement, and naturalness (Inoue et al., 2022). Similarly, the coordination of laughter and gaze timing can strengthen impressions of a robot’s empathy and compassion (Giannitzi et al., 2025).

In this study, these insights are extended to a moral context. The expressive behaviours implemented in the robot (in this project limited to disfluencies and social laughter)are investigated not merely as surface markers of affect, but as potentially meaningful cues that shape reasoning and moral judgment. Examining how these cues produced by a robot affects participants’ decision-making may reveal whether they can prompt reflection or reconsideration, mirroring effects observed in human–human interaction.

4 Research Design

We aim to test whether hesitation, expressed through a pause, a filled pause, or social laughter, combined with a clarification request, influence participants’ moral decisions in the Balloon Task. Therefore, our research question is:

- Does Furhat’s hesitation prompt participants to reconsider or change their decision in a morally charged task?

We hypothesize that:

- When faced with a morally difficult choice, hesitation from the interlocutor (Furhat) acts as a social cue signaling doubt or moral conflict. This may “push back” on the participant’s decision, increasing the likelihood of reconsideration.

We further anticipate that conditions including social laughter and filled pauses will be more likely to prompt participants to reconsider their moral decisions, whereas a silent pause alone will not produce such an effect.

H_0: The robot’s hesitation has no effect on participants’ moral decisions in the Balloon Task.

H_1: Hesitation expressed through social laughter or filled pauses increases the likelihood that participants will reconsider or change their moral decisions, while a silent pause has no significant influence.

4.1 The Ballon Task

The Balloon Task presents participants with four characters, a Pilot, a Teacher, a Doctor and a Prodigy trapped in a balloon that is losing altitude. To save the others, one person must jump. The task requires participants to discuss their reasoning and ultimately choose who should die for the greater good.

Table 1: Experimental conditions and corresponding manipulations.

Cond.	Manipulation	Furhat utters
1	Filled pause + echo	*“Hmm... the doctor?”*
2	Silent pause + echo	*“... the doctor?”*
3	Laughter + echo	*“Ha-ha, the doctor?”*

In our current implementation, each condition lasts approximately one second.

4.2 Procedure

Furhat presents the dilemma and introduces all four characters.

The participant are encouraged to discusses all possible options.

Participants are instructed to say "I want X to jump" when come to a decision.

When the system identifies that, it produces one of the responses depending on the assigned condition. The participant may confirm or change their decision.

Dependent variables include decision change, response latency, and linguistic hesitation in the participant's reply.

5 Dialogue System Implementation

The dialogue system was implemented as a state machine using `XState` in TypeScript, integrating Furhat's Remote API and a locally hosted large language model (LLM) through Ollama. The system manages dialogue turns through three core states—*Speaking*, *Listening*, and *ProcessingResponse*—that are continuously cycled until a moral decision is detected in the participant's utterance. The LLM receives the dialogue history as a message array (`messages: {role, content}`) and generates short, neutral responses following predefined moral dilemma instructions. A guard function monitors each user turn, extracting the chosen character (e.g., "I want the Doctor to jump") via a regular expression, which then triggers the manipulation phase. The system begins by setting Furhat's voice (`fhSetVoice`, `fhAttendUser`), attends to the closest user, and alternates between Furhat's text-to-speech (`fhSay`) and automatic speech recognition (`fhListen`) calls. In the current implementation, all hesitation cues last approximately one second before Furhat produces the echoic clarification request ("X?").

When a decision is detected, the state machine transitions to the *Manipulation* state, where one of three experimental conditions is applied: a silent pause, a filled pause ("Hmm"), or social laughter, each followed by the echoed phrase ("the X?").

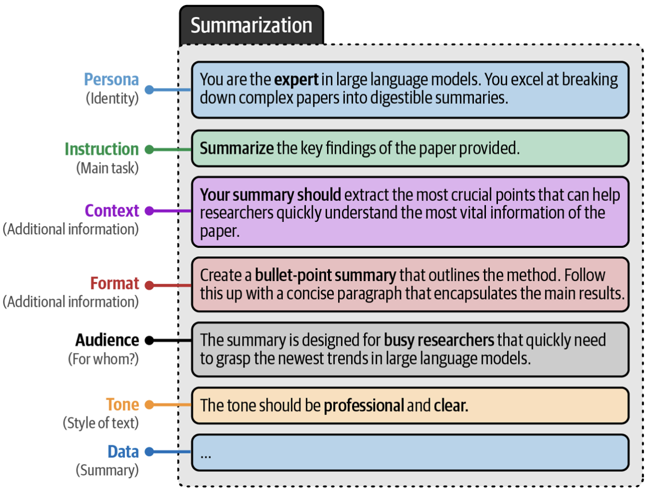
These manipulations correspond to the three hesitation conditions tested in the study. After the manipulation, Furhat listens for a response: if the participant answers "yes," the interaction ends (*End*); if "no," the system re-enters the main dialogue loop with the LLM. The following code snippet illustrates the main manipulation block implemented in `XState`:

```
DetermineTheManipulationState:
  always: [
    { target:
    "SpeakingWithBuffer_Condition3",
    guard: "isSpeakingWithBuffer" },
    { target:
    "SpeakingWithLaughter_Condition2",
    guard: "isSpeakingWithLaughter" },
    { target: "SpeakingWithPause_Condition1" }],

SpeakingWithPause_Condition1:
{ text: "........ The X?" },
SpeakingWithLaughter_Condition2:
{ text: "Hahaha, The X?" },
SpeakingWithBuffer_Condition3:
{ text: "Hmm, The X?" }
```

6 Prompt Engineering

The system prompt was inspired by the framework presented by (Alammar and Grootendorst, 2024), which outlines key dimensions of effective prompt construction such as persona, instruction, context, format, audience, and tone (see Figure 1).



You are the **expert** in large language models. You excel at breaking down complex papers into digestible summaries.

Summarize the key findings of the paper provided.

Your summary should extract the most crucial points that can help researchers quickly understand the most vital information of the paper.

Create a **bullet-point summary** that outlines the method. Follow this up with a concise paragraph that encapsulates the main results.

The summary is designed for **busy researchers** that quickly need to grasp the newest trends in large language models.

The tone should be **professional** and **clear**.

...

Figure 1: A complex prompt structure (Alammar and Grootendorst, 2024, p. 179)

Following this structure, our prompt explicitly defined the model's identity and communicative role (persona), specified the main task and its boundaries (instruction and context), and established formatting and stylistic constraints to ensure clarity and neutrality (format and tone). This layered design guided the model's behaviour toward acting as a neutral conversational partner who supports ethical reflection without persuasion.

7 Future Work and Limitations

The current implementation presents several limitations. First, the speech synthesis system constrains prosodic variation, meaning that robotic hesitation may lack the natural timing and intonation patterns observed in human speech. Second, participant interaction is limited to pre-defined prompts, preventing the system from generating fully adaptive or context-sensitive responses. Third, maintaining behavioural consistency across sessions remains challenging, as large language models can exhibit variability in tone or phrasing even when provided with identical prompts.

Current application of Furhat does not also check for other forms of decision-indicating points–only accepting the form "I want X to jump." This is, of course, untenable if we desire to capture all forms of ethical decision-making utterances. For this, we plan to implement an NLU structure, trained on many Intent and Entity tokens, so that utterances such as "I decide on X," or "I think, maybe, X should go." should also be accepted to trigger the manipulation state.

Another constructive criticism we have received after showcasing our demo was that the robot was generating too long utterances. Additionally, utterances present a very unnaturally deadpan tone. This reduces our system to a mere assistant rather than a conversation participant in the task. We aim for not a full human-like dialogue but a robot that is believable enough so that the participants might take it seriously enough to discuss a moral dilemma.

Future work will focus on increasing the adaptive capabilities of the dialogue system. One direction is to implement real-time prosodic monitoring, allowing Furhat to modulate hesitation duration or intensity based on the participant's vocal and temporal cues. Additionally, integrating multimodal sensing, such as gaze tracking, facial expression recognition, or vocal tension analysis, could enhance the robot's responsiveness and social sensitivity. These developments would contribute to more fluid, context-aware interactions and a deeper understanding of how timing and hesitation shape moral reasoning in human–robot dialogue.

8 Final Comments

This project presents the development of a dialogue system that integrates Furhat's embodied interaction capabilities with a large language model to investigate how hesitation cues influence moral decision-making. By combining controlled manipulations of pauses, filled pauses, and social laughter with a structured conversational prompt, the system enables fine-grained study of timing, reflection, and ethical reasoning in human–robot dialogue.

Beyond its immediate experimental goals, we believe that this work may contribute to the broader goal of designing empathetic and ethically aware dialogue systems that can engage with human uncertainty not by asserting answers, but by hesitating with us.

References

Jay Alammar and Maarten Grootendorst. 2024. *Hands-On Large Language Models*. O'Reilly Media.

Herbert H. Clark and Jean E. Fox Tree. 2002. Using uh and um in spontaneous speaking. *Cognition*, 84(1):73–111.

Eleni Giannitzi, Vladislav Maraev, Erik Lagerstedt, and Christine Howes. 2025. Laughter in sight: How gaze and laughter affect perceptions of a social robot. In *Human-Computer Interaction*, pages 249–268. Springer.

Jonathan Ginzburg and Robin Cooper. 2004. Clarification, ellipsis, and the nature of contextual updates in dialogue. *Linguistics and Philosophy*, 27(3):297–366.

Christine Howes, Mary Lavelle, P. G. T. Healey, Julian Hough, and Rose McCabe. 2017. Disfluencies in dialogues with patients with schizophrenia. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*, London, UK. Cognitive Science Society.

Koji Inoue, Divesh Lala, and Tatsuya Kawahara. 2022. Can a robot laugh with you? shared laughter generation for empathetic spoken dialogue. *Frontiers in Robotics and AI*, 9:933261.

Emily Kirkland and Johan Edlund. 2025. Who knows best? effects of speech disfluencies on incentivized decision-making. In *Proceedings of Interspeech 2025*, pages 3142–3146, Kos, Greece. ISCA.

Anita Pomerantz. 1984. Agreeing and disagreeing with assessments: Some features of preferred/dispreferred turn shapes. In J. Maxwell Atkinson and John Heritage, editors, *Structures of Social Action: Studies in Conversation Analysis*, pages 57–101. Cambridge University Press, Cambridge.