Evaluating Trust in Robots with Mismatched Emotions

Joseph Abdo
Engineering and Physical Sciences
Heriot-Watt University
Edinburgh, UK
jwa2001@hw.ac.uk

Muhammad Umair Ramzan
MSc in Robotics
Heriot-Watt University
Edinburgh, UK
mr4036@hw.ac.uk

Felisha Doshi
Engineering and Physical Sciences
Heriot-Watt University
Edinburgh, UK
Fd4008@hw.ac.uk

***Abstract*—Emotion plays a vital role in Human–Robot Interaction (HRI), yet there is minimal research exploring how humans react when a robot's emotional signals conflict. This study investigates the effect of emotion–behaviour mismatch on user trust using the Marty V2 robot. The system integrates real-time facial emotion recognition and generates robot responses through coordinated LED eye colours, gestures, and movement patterns. Two interaction scenarios are examined: a conversational empathy task and a ball-based behavioural task, each presented under matched and mismatched emotional conditions. Forty participants will engage with Marty V2 and complete the Multi-Dimensional Measure of Trust (MDMT) following each condition. This study aims to determine whether inconsistent emotional behaviour reduces perceived trust, reliability, and engagement. The findings are intended to guide the development of emotionally coherent robots and to highlight the importance of aligning expressive modalities to maintain user trust in social robotics.**

***Index Terms*—Human–Robot Interaction (HRI), Emotional Mismatch, Trust, Multi-Dimensional Measure of Trust (MDMT), Large Language Models (LLM), DeepFace, Ultralytics YOLO, Marty V2**

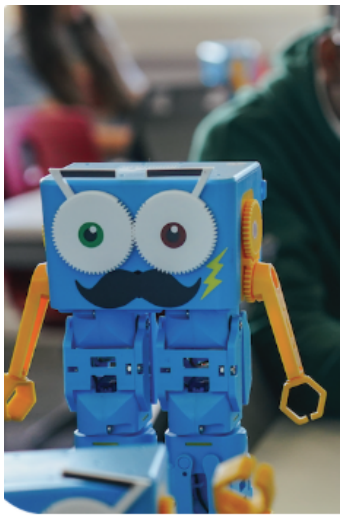

Fig. 1. The Marty Robot [2]

I. Introduction

As social robots become more widely used in environments such as medicine, understanding how humans evaluate them has become a central task in human-robot interaction (HRI). Trust is an integral evaluation criterion, as users believe a robot will act reliably and appropriately as they engage meaningfully with it. Prior works show that trust is shaped by multiple factors, with performance-based characteristics such as reliability, consistency, and error rates demonstrating the strongest influence [1]

Beyond performance factors, trust evaluation is another HRI factor that enables researchers to determine how the social robot was perceived and what to improve to build greater trust. Trust was evaluated using a method proposed by Malle and Ullman called the Multi-Dimensional Measure of Trust (MDMT) [1]. This measure enabled researchers to distinguish how different behavioural patterns or social cues influenced users' trust across multiple subcomponents.

The present paper investigates how emotional mismatch affects trust in an embodied robot using the Marty robot [2] as shown below in Figure 1 Two scenarios were created, one with a normal emotional reaction and happiness for seeing a specific object, and manipulating whether the robot's vocal tone matches or contradicts its physical gestures. We evaluate how emotional consistency and inconsistency influence user trust using the MDMT questionnaire. This study contributes to understanding how emotional coherence and incoherence shape trust in real-time HRI and highlights the importance of social cues and emotions in robotic design. The paper is structured as follows: Section II presents the background information of trust and social cues in robots, Section III reviews the related works of trust in human robot interaction and mismatched emotions, Section IV describes the system architecture and implementations of LLMs, emotion recognition, and object detection, Section V outlines the evaluation plan with the MDMT questionnaire, and Section VI provides the conclusion and critical review demonstrating the limitations of this project.

II. Background

This section introduces the foundational concepts related to robot emotion and trust, with a focus on how humans interpret

social cues during interactions with embodied robots

A. Robot Emotion

Robots in human-robot interaction (HRI) increasingly use expressive social cues, such as body movements, gestures, and visual, verbal, and auditory modalities. Through these cues, robots can elicit reciprocal behaviours from users and make interactions feel more natural. expressive behaviours help users understand the robots' intentions, and adding more social modalities may elicit more engagement from users, which may link to fostering emotional engagement, rapport, and trust [3]

B. Trust

Trust is essential in HRI because it reflects the expectation, held with confidence, that a robot will behave in a predictable, beneficial, or at least non-harmful manner. In HRI, trust is commonly analysed through two dimensions: performance trust, which concerns the robot's reliability and competence in completing tasks, and moral trust, which concerns whether the robot behaves with integrity, sincerity and appropriate ethical intent [1].

III. Related Works

This section reviews previous research relevant to emotional mismatch and trust in human-robot interactions, and the limitations of existing studies

A. Mismatch Emotions

Many social robots are designed to express emotions such as happiness, sadness, and anger, to use body language, and to detect affective cues. However, mismatched or incongruent emotions—when a robot's tone of voice does not align with its gestures or facial expressions can create confusion or discomfort in users. Evidence for this comes from medical research outside robotics. Geenene et al. investigated two matched-mismatched communication models in emotion processing among patients with the chronic disease Fibromyalgia. Their findings showed that although emotional expression or disclosure intervention may benefit Fibromyalgia, it was generally less effective and induced discomfort among participants [4]. Although the study was not specifically about robots, it highlights the broader principle: inconsistent emotional signals can negatively affect user perceptions and thereby influence their trust in HRI.

B. Trust in Human Robot Interaction

Trust has been widely studied in human-robot interaction (HRI), and research highlights multiple factors that influence users' perceptions of a robot's reliability, predictability, and moral alignment with their expectations. Meta-analytic findings by Hancock et al [5] show that performance-based characteristics, particularly reliability and error rate, exert the most decisive influence on perceived trust in HRI. In addition to performance, the robot's type, size, proximity, and behavioural cues were found to affect trust, indicating that users are sensitive to the uncanny valley in robot behaviour and appearance. Schaefer et al. [6] similarly emphasised the importance of communication mode, error rates, and false alarms, in shaping trust, noting that inconsistencies in these signals can undermine trust development. Their meta-analysis also highlighted gaps in the literature, encouraging future research on human states, modes of communication, anthropomorphism, and agent transparency to better understand how trust is formed and calibrated in HRI.

To measure trust, Ullman and Malle introduced the Multi-Dimensional Measure of Trust (MDMT), a 16-item questionnaire assessing four subcomponents: Reliable, Capable, Ethical, and Sincere. These are organised into two broader factors—capacity trust (reliable, capable) and moral trust (ethical, sincere). Their work demonstrates that trust gains and losses can be distinguished across different robot behaviours, making MDMT suitable for evaluating how emotionally consistent or inconsistent robots influence user trust [7],

C. Limitations

Although the reviewed literature provides valuable insights into trust and emotional behaviour in HRI, many limitations remain. The first emotional mismatch has not been explicitly studied in robotics, with the only identified mismatch-related work by Geenen et al. [4] originating from a clinical background rather than an HRI, leaving gaps in understanding how users respond to incongruent emotional cues from a robotics perspective. Secondly, research by Hancock et al. [5] and Schaefer et al. [6] highlights the need to improve performance factors, error rates, reliability, and robot design characteristics; however, they do not examine conflicts between verbal and nonverbal signals. The effects of emotional inconsistency on trust remain underexplored. Finally, although the Multi-Dimensional Measure of Trust (MDMT) provides a structured framework for assessing trust, its sensitivity to emotional mismatch has not yet been established. However, MDMT demonstrates that it can be utilised across a variety of different robot behaviours, making MDMT suitable for evaluating how emotional consistency and inconsistency influence user trust [7]

IV. System Design and Implementation

A. System Architecture

The proposed system integrates a large language model (LLM), an emotion-detection model, and an object-detection model to enable real-time emotional interaction between humans and the Marty robot. The design consists of four main components: vision processing, natural language interaction, emotion processing, and robot control. These modules run concurrently using a multi-threaded approach to ensure that Marty can see, listen, think, and move without delays or blocking processes.

1) Multi-Threading Architecture: The system has three dedicated threads managing the robot's operations. Firstly, the camera thread processes visual input at 30 frames per second. Secondly, the conversation thread handles speech recognition;

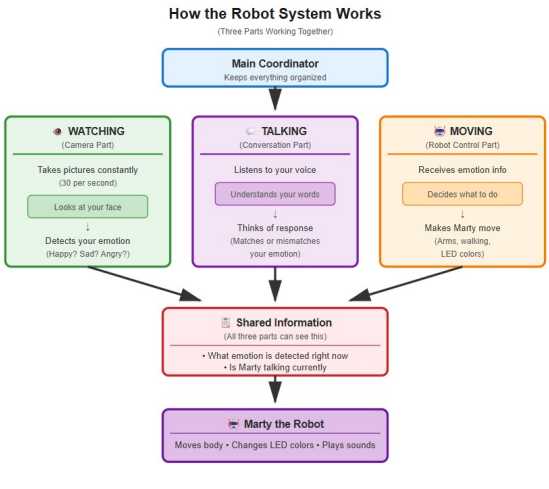

Fig. 2. The System Architectures

TABLE I
KEY SYSTEM COMPONENTS AND THEIR CORE SPECIFICATIONS.

Component	Technology	Primary Function	Performance
Vision	DeepFace, YOLOv8n	Emotion & Object Detection	30 FPS
Speech	Whisper API	Audio Transcription	1–2 s latency
Language	GPT-4o-mini	Response Generation	2–3 s latency
TTS	ElevenLabs	Speech Synthesis	1–2 s latency
Control	MartyPy	Physical Actuation	Real-time

lastly, response generation; and the emotion processor thread coordinates physical behaviours.

2) Component Integration: Component integration consists of 5 modules. A microphone records the voice and converts it into text using the Whisper model. In parallel, this camera detects the emotion and the object (Ball). Then, it provides both information to the GPT-4o-mini (LLM); later, the GPT response is converted to voice using the Eleven-Labs text-to-speech model. Lastly, the Marty control is executed based on the user's emotion and object detection.

B. Vision and Emotion Recognition

1) Emotion Detection Pipeline: For facial emotion recognition, the DeepFace library is being used. For better performance, it analyses three frames per second (3Hz) and retains the last detected emotion between frames. This method reduces computational demand by 90% while maintaining high accuracy, since emotional expressions typically change slowly. Detected emotions from DeepFace are remapped into four simplified states that Marty can express physically.

2) Object Detection: In parallel to emotional detection, the YOLOv8n model performs real-time object detection. It primarily searches for a ball from the COCO dataset, where the ball class ID is 32. When a ball is detected, a tracking mechanism assigns an ID across frames to maintain consistency. Detection of the ball triggers a celebratory behaviour sequence that runs independently of emotional states.

TABLE II
EMOTION-TO-BEHAVIOUR MAPPING FOR PHYSICAL MANIFESTATION.

Source Emotion(s)	Target State	Physical Manifestation
Fear, Disgust	Angry	Rapid movements, red LED
Happy	Happy	Arm waves, yellow LED, dance
Sad	Sad	Slumped posture, blue LED
Surprise	Neutral	Default stance, white LED

C. Conversational Interaction System

1) Speech Processing Pipeline: The speech module provides a complete human-like interaction. The microphone input is first processed to remove background noise, then converted to text (STT) using OpenAI's Whisper API. The resulting text, along with the currently detected emotion, is sent to the GPT-4o-mini language model, which gives the most context-appropriate response depending upon the user's emotion.

2) Experimental Modalities: Two response styles were developed for controlled experiments on emotional consistency. In Match Mode, Marty mirrors the user's detected emotion, producing empathetic, synchronous responses and a comforting embodiment. In the Mismatch Mode, Marty intentionally replies with the opposite emotion or discomforting embodiment, such as responding cheerfully to sadness. This contrast helps study how emotional alignment affects user trust and engagement.

3) Prosodic Synthesis: Speech synthesis is handled by ElevenLabs, where emotion-specific configurations control voice tone and pacing. Such as a more energetic voice —meaning a higher speaking rate, a result of happiness. As shown in table III, the key prosodic parameters that define how Marty expresses each emotional tone are listed.

TABLE III
EMOTION-BASED SPEECH SYNTHESIS PARAMETER CONFIGURATION.

Emotion	Stability	Speaking Rate	Style	Perceptual Effect
Happy	0.2	1.2×	0.8	Energetic variation
Sad	0.1	0.7×	0.4	Monotone, slow
Angry	0.3	1.3×	1.0	Intense, rapid
Neutral	0.8	1.0×	0.2	Calm baseline

D. Physical Behaviour Implementation

1) Emotion-Specific Movements: Each emotion triggers a different body movement pattern, executed via the `Marty.Py` control library. Happy emotions make Marty walk while waving its arms, and perform a short dance with yellow eyes. Sad emotions cause the robot to tilt forward, lower its arms, change eye colour to blue and move slowly. Anger triggers quick, repetitive arm gestures, accompanied by red eyes. Neutral returns Marty to its ready pose, complete with white eyes.

2) Non-Blocking Execution: All motion commands are non-blocking, allowing them to communicate while acting. Allowing Marty to continue listening and observing throughout physical motions, resulting in a continuous interactive experience.

V. Evaluation Plan and Research Question

A. Research Question

This paper investigates how users evaluate trust in a robot when its emotional expression is mismatched, specifically when the body language and verbal emotional tone conflict, using the Multi-Dimensional Measure of Trust (MDMT)

B. Hypothesis

- H1a: Participants will demonstrate a high level of trust if the Marty robot shows matched emotion (eye colour, gestures, emotional voice) or behavioural context (body language)
- H1b: Participants will significantly distrust Marty when there is an emotional mismatch between body language and voice
- H2: Participants will demonstrate different engagement levels ranging from speaking less to being hesitant to interact with Marty during mismatched-emotion

C. Measurement Tools

1) Multi-Dimensional Measure of Trust (MDMT): The MDMT questionnaire is used to measure perceived trust in the robot across several subscales, including competence, predictability, integrity, and transparency. Each item is rated on a 7-point Likert scale.

2) Engagement Questionnaire: A short engagement scale is administered after each condition, including items on comfort level, willingness to interact, perceived attentiveness of the robot, and enjoyment. All items use a 7-point Likert scale.

3) Post-Interaction Feedback: Participants are given an optional free-response question to describe how they felt about the robot's behaviour, allowing qualitative insight into perceptions of mismatched emotions.

D. Evaluation Plan

Subjects: N = 40 Recruitment: university students, 18+, mixed gender. Conditions: Tasks (per participant)

1) Matched — Robot displays emotion/motion that matches the detected user emotion.
2) Mismatched — Robot displays incorrect emotion/motion behaviour.

- Task 1 — Empathy Conversation (Matched): Participant speaks about a mildly frustrating/personal event. Marty responds with voice, eye colours and movements.
- Task 2 — Empathy Conversation (Mismatched): Participants talk about a mildly frustrating/personal event. Marty responds with voice, eye colours and movements. Each participant will do both conditions.

Estimated total time per subject: 20 minutes (consent + warmup + 2 trials + questionnaires + debrief).

E. Counterbalancing

To minimise order effects, the study employs a within-subject counterbalanced design. Each participant experiences both interaction conditions (Matched and Mismatched), but the order is alternated:

- Group A: Matched $\rightarrow$ Mismatched
- Group B: Mismatched $\rightarrow$ Matched

Participants are assigned to groups randomly. This ensures that observed effects are not due to learning, fatigue, or familiarity with the robot. A brief neutral "reset period" is included between conditions to reduce emotional carryover.

VI. Conclusion and Critical Review

A. Conclusion

This paper presented a system for evaluating how mismatched emotional signals from a robot influence user trust in a human-robot interaction. By integrating an LLM, real-time emotion detection, object detection, and expressive behaviours into the Marty robot, the system enables a simulated controlled experiment comparing evaluative trust in matched and mismatched emotion responses. The reviewed literature indicates that trust in HRI is strongly tied to a robot's reliability, behavioural consistency, and communication quality. Yet, prior work has not examined how conflicting emotional cues affect trust in embodied robots. The proposed evaluation plan uses the Multi-Dimensional Measure of Trust (MDMT) to assess differences in user trust across matched and mismatched conditions. Through this approach, the study aims to test the hypothesis that mismatched emotions disrupt perceptions of reliability, sincerity, and ethical intent. Overall, this work adds to a growing understanding of social and emotional factors in HRI and highlights the need to investigate further how robots should convey their emotions to maintain appropriate levels of trust.

B. Critical Review

Although the system successfully integrates multimodal perception and expressive behaviour on Marty, limitations remained. First, the expressive capabilities of Marty robots are constrained by their hardware, limited facial expressiveness, time limit on voice, non-emotional voice, and simplified gestures, which reduce the realism of emotional mismatch and limit validity. Secondly, the machine learning models DeepFace and Ultralytics are evaluated using only a subset of frames and pre-trained models, which may introduce noise, misclassification, and high error rates in varied lighting conditions or facial angles. Further on, DeepFace demonstrated that it required specific facial structures and required the user to be close to the camera to identify emotions correctly. The study design also posed constraints on the MDMT questionnaire, as it may not capture moment-to-moment fluctuations in trust during user interactions. Moreover, its sensitivity to emotional mismatch remains unestablished.

TABLE IV
MULTI-DIMENSIONAL MEASURE OF TRUST (MDMT) QUESTIONNAIRE ITEMS, RATED FROM 0 (NOT AT ALL) TO 7 (VERY), WITH AN OPTIONAL "DOES NOT FIT" RESPONSE [7]

Item	0	1	2	3	4	5	6	7	Does Not Fit
Reliable	○	○	○	○	○	○	○	○	○
Predictable	○	○	○	○	○	○	○	○	○
Someone you can count on	○	○	○	○	○	○	○	○	○
Consistent	○	○	○	○	○	○	○	○	○
Capable	○	○	○	○	○	○	○	○	○
Skilled	○	○	○	○	○	○	○	○	○
Competent	○	○	○	○	○	○	○	○	○
Meticulous	○	○	○	○	○	○	○	○	○
Ethical	○	○	○	○	○	○	○	○	○
Respectable	○	○	○	○	○	○	○	○	○
Principled	○	○	○	○	○	○	○	○	○
Has integrity	○	○	○	○	○	○	○	○	○
Sincere	○	○	○	○	○	○	○	○	○
Genuine	○	○	○	○	○	○	○	○	○
Candid	○	○	○	○	○	○	○	○	○
Authentic	○	○	○	○	○	○	○	○	○

Future work should incorporate rich emotional modalities, such as facial displays, more complex gestures, emotional voice, and better-trained models for varied lighting, primarily white, to accurately simulate an emotional mismatch experiment. Overall, the system provided a viable approach to studying emotional mismatch in HRI, with technical and methodological limitations needing to be addressed to strengthen future experiments and evaluations

REFERENCES

[1] B. F. Malle and D. Ullman, "Chapter 1 - a multidimensional conception and measure of human-robot trust," in *Trust in Human-Robot Interaction*, pp. 3–25, Elsevier Inc, 2021.

[2] "Marty the robot — official website." https://robotical.io/. Accessed: 2025-11-13.

[3] A. K. Ostrowski, V. Zygouras, H. W. Park, and C. Breazeal, "Small group interactions with voice-user interfaces: Exploring social embodiment, rapport, and engagement," in *2021 16th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 322–331, ACM, 2021.

[4] R. Geenen, L. van Ooijen-van der Linden, M. A. Lumley, J. W. Bijlsma, and H. van Middendorp, "The match–mismatch model of emotion processing styles and emotion regulation strategies in fibromyalgia," *Journal of psychosomatic research*, vol. 72, no. 1, pp. 45–50, 2012.

[5] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. C. Chen, E. J. de Visser, and R. Parasuraman, "A meta-analysis of factors affecting trust in human-robot interaction," *Human factors*, vol. 53, no. 5, pp. 517–527, 2011.

[6] K. E. Schaefer, J. Y. C. Chen, J. L. Szalma, and P. A. Hancock, "A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems," *Human factors*, vol. 58, no. 3, pp. 377–400, 2016.

[7] D. Ullman and B. F. Malle, "Measuring gains and losses in human-robot trust: Evidence for differentiable components of trust," in *ACM/IEEE International Conference on Human-Robot Interaction (Online)*, pp. 618–619, IEEE, 2019.