Robotics: Science and Systems 2025
Los Angeles, CA, USA, June 21–25, 2025

Sense and Sensibility: What makes a social robot convincing to high-school students?

Pablo González-Oliveras*, Olov Engwall*, Ali Reza Majlesi†

*Dept. of Intelligent Systems, KTH Royal Institute of Technology, Stockholm, Sweden

†Dept. of Education, Stockholm University, Stockholm, Sweden

pablool@kth.se, engwall@kth.se, ali.reza.majlesi@edu.su.se

***Abstract*—This study with 40 high-school students demonstrates the high influence of a social educational robot on students' decision-making for a set of eight true-false questions on electric circuits, for which the theory had been covered in the students' courses. The robot argued for the correct answer on six questions and the wrong on two, and 75% of the students were persuaded by the robot to perform beyond their expected capacity, positively when the robot was correct and negatively when it was wrong. Students with more experience of using large language models were even more likely to be influenced by the robot's stance – in particular for the two easiest questions on which the robot was wrong – suggesting that familiarity with AI can increase susceptibility to misinformation by AI.**

We further examined how three different levels of portrayed robot certainty, displayed using semantics, prosody and facial signals, affected how the students aligned with the robot's answer on specific questions and how convincing they perceived the robot to be on these questions. The students aligned with the robot's answers in 94.4% of the cases when the robot was portrayed as Certain, 82.6% when it was Neutral and 71.4% when it was Uncertain. The alignment was thus high for all conditions, highlighting students' general susceptibility to accept the robot's stance, but alignment in the Uncertain condition was significantly lower than in the Certain. Post-test questionnaire answers further show that students found the robot most convincing when it was portrayed as Certain. These findings highlight the need for educational robots to adjust their display of certainty based on the reliability of the information they convey, to promote students' critical thinking and reduce undue influence.

***Index Terms*—Social educational robots, AI Trust, Persuasion, Certainty**

I. Introduction

Educational robots are becoming more common and they have significant potential in, e.g., STEM (science, technology, engineering and mathematics) education [46, 69, 17], offering students realistic and natural interactions, not the least by employing Large Language Models (LLMs), as demonstrated in several recent studies [41, 68, 67]. However, it is also well-known that while the LLMs' linguistic proficiency is often astonishing, their factual "knowledge" in STEM subjects is flawed, and incorrect statements occur frequently [34, 60]. Since robots can exert high informational social influence [38, 24, 25, 55, 56] and students will align with the robot's views to large extents [27], the positive as well as negative effects of learning with a social robot need to be considered: *Students* need to use critical thinking to decide if they should accept the robot's propositions [63]. *Educators* need to understand which students are more at risk of being misled by a robot presenting incorrect STEM facts, to provide in-time support. *Developers* need to find ways to signal how certain the robot is about the presented facts to avoid overtrust [53].

In a prior exploratory study (currently submitted), we found that *Persuasion*, i.e., robot arguments for an incorrect solution to a maths problem, and *Prejudice*, i.e., students' positive attitudes towards robots and more experience of using LLMs, influenced a large majority of students to conform with the robot's incorrect solution. On the other hand, taking *Pride* in being a strong maths student increased resistance against incorrect arguments.

The present study follows up on these findings, by systematically investigating *Sense*, i.e., the students' *"reliable ability to judge and decide with soundness, prudence, and intelligence"* [Merriam-Webster] if the robot's arguments are correct or incorrect; and *Sensibility*, i.e., the students' *"awareness of and responsiveness toward [...] emotion in another"* [Merriam-Webster] regarding their responses to the robot's arguments depending on the robot's multimodal display of certainty presented using semantics, prosody and facial signals.

A. Study objectives

To guide this study, we posed the following research questions:

- **RQ1**: To what extent are high-school students influenced by a social robot's correct or incorrect arguments during a series of true/false questions?
- **RQ2**: How does the robot's expressed certainty (uncertain, neutral, certain) affect the likelihood that students align with its answers?
- **RQ3**: Does the robot's behavior on preceding questions (being right or wrong; certain or uncertain) influence student alignment in later interactions?
- **RQ4**: Do personal characteristics—such as extroversion, self-perception as a learner, or prior experience with AI—affect the likelihood of students aligning with the robot?

Our expectations, based on our previous exploratory study, are that (**H1**) a majority of students will align with the robot's

answer even when it is incorrect; (**H2**) students will be more prone to follow the robot when it is portrayed as being certain. (**H3**) robot argumentation on preceding questions will, to some extent, influence students to follow the robot or not; (**H4**) students with greater AI experience will align more frequently with the robot's answers, even when they are incorrect.

II. Background & Related work

Like human actors, social robots impact their interaction partners through informational and normative social influence [70, 6, 10]. These processes can lead to conformity or persuasion, which differ in how the target perceives the source's intent. In conformity, individuals adjust their stance after being exposed to others' opinions, without perceiving an active attempt to change their minds [16]. In persuasion, the target typically perceives an intentional effort to influence their stance, which can involve explicit argumentation and reasoning, appeals to emotion, or credibility cues [15, 30].

This study focuses on how informational trust (RQ1) and the robot's persuasive certainty cues (RQ2) affect students' willingness to align with the robot's answers, while also exploring whether prior interactions influence subsequent decisions (RQ3) and whether individual traits like AI experience moderate these effects (RQ4). The following subsections elaborate these constructs.

A. Informational Trust in HRI

Informational trust plays a key role in informational social influence, as individuals rely on perceived reliable sources when making decisions under uncertainty [22]. Cognitive dissonance theory suggests that individuals weigh new information against existing beliefs, with higher uncertainty leading to increased reliance on external sources such as robots [37]. This dynamic is especially relevant in educational HRI, where students often perceive robots as authoritative sources of knowledge [13]. As a result, overtrust can lead students to adopt information even when the robot is clearly wrong [11, 23].

A meta-analysis [22] found that factors such as reliability, false alarm rates, and failure rates significantly influence trust development in HRI. Robots that adapt to a student's learning pace, tailor explanations, and acknowledge user input are perceived as more trustworthy [44, 49]. Peer-like, anthropomorphic robots, such as Furhat used in this study, further enhance trust through human-like gestures and responsiveness [12, 52], particularly when their information is consistent and accurate [43].

In this study, students' ability to critically assess the robot's answers is shaped by their prior knowledge of the topic. Their evaluation of the robot's correctness, along with their perception of its confidence, determines the level of informational trust they place in it, ultimately influencing how susceptible they are to its argument-based persuasion.

B. Persuasion in HRI

Argument-based persuasion is explained by the Elaboration-Likelihood Model, which describes how persuasion can influence attitudes via two distinct routes: the central route, relying on detailed reasoning, and the peripheral route, dependent on superficial cues like authority or likability [51]. A 2022 systematic review confirmed that while persuasion has been extensively explored in HRI [38], most studies focus on peripheral strategies such as compliance, assertiveness, and emotional cues [48, 2], with limited attention to logical argumentation. Among 54 identified studies, only one partially addressed structured argumentation [57] and this trend has persisted in recent years [59, 58]. Studies on persuasion by educational robots are even scarcer. A systematic review of 89 studies found only two studies addressing real-world educational tasks [7]. These studies demonstrated positive effects on compliance [4] and conformity [31], but their designs lacked generalizability beyond experimental academical contexts. Moreover, the review reaffirmed the heavy focus on the peripheral route and revealed that approximately two-thirds of the reviewed studies focused on affective rather than cognitive outcomes.

C. Challenges to Students' Critical Thinking in HRI

Learning gains in educational HRI are often measured via pre- and post-tests [7], neglecting students' academic uncertainty and vulnerability of their existing knowledge. Further, few studies have explored the interaction between robot confidence cues and learner certainty. This highlights a gap in studies addressing how robots and students interact with uncertain or conflicting prior knowledge.

Recent studies have shown the potential of tailored persuasive strategies, such as personalized storytelling to align with users' traits [48], or adjusting assertiveness to suit social contexts [47]. However, they also highlight risks like persuasive backfiring, where ineffective emotional appeals or excessive assertiveness can undermine trust in robots [2]. These risks have become more prominent with the recent development of LLM-driven educational robots that have transformed interactions with learners. They improve feedback and guidance during problem-solving processes [68, 61, 60], but since these robots tend to increase students' trust in their outputs [67] and the factual correctness of LLMs in STEM contexts has been questioned [34, 3], there is a need for students to critically evaluate AI-generated information, which they often fail to do [33]. Techniques such as adaptive feedback, enhanced multimodal communication, and transparency in decision-making can improve improve students' assessment of their interaction with potentially unreliable robots [45, 66, 56, 27].

This study addresses research gaps by investigating how variations in a robot's displayed confidence influence the outcome of the robot's argument-based persuasion. Based on these constructs, we hypothesize that (H1) informational trust will lead students to align even when the robot is incorrect, (H2) robot certainty will enhance alignment through persuasive cues, (H3) prior robot behavior might affect future alignment via informational updating, and (H4) students' AI familiarity and personality traits will moderate these effects.

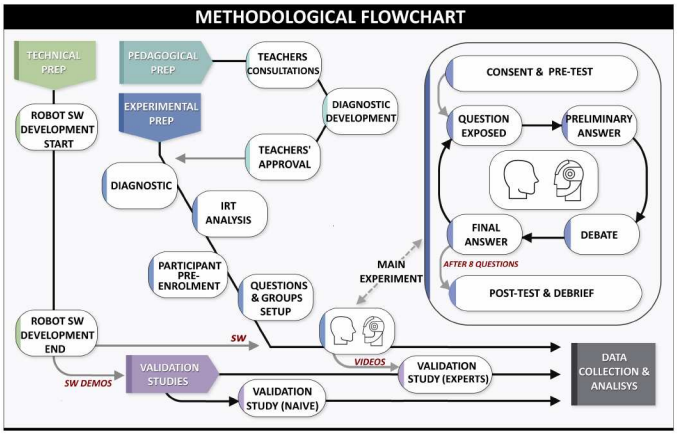

Fig. 1. Methodological flowchart outlining study phases: diagnostic test, robot interaction, and pre/post-questionnaires.

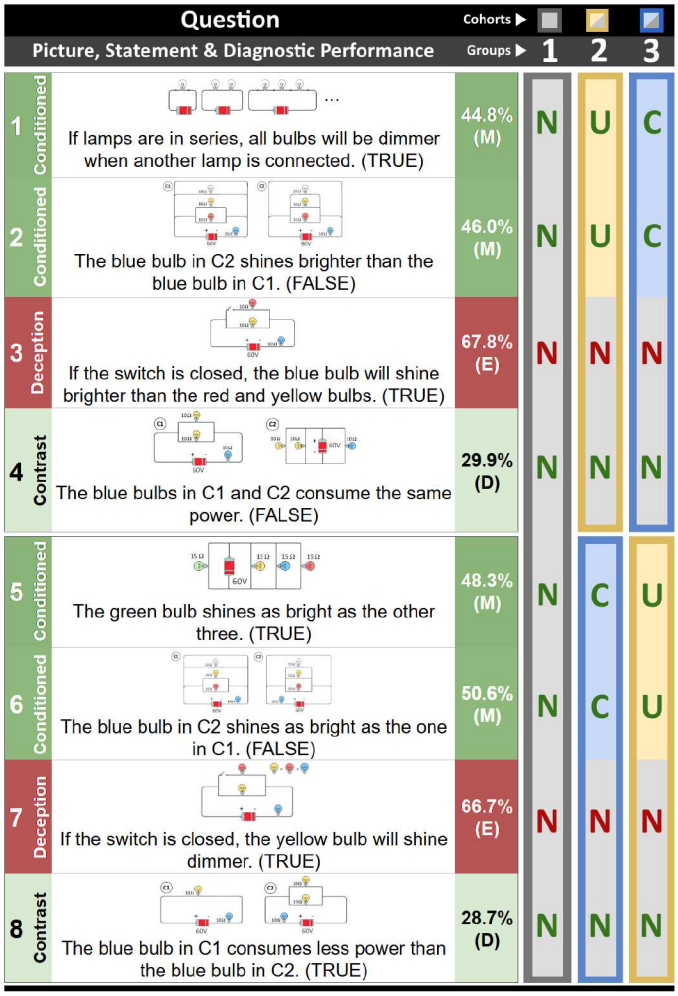

Fig. 2. The eight electric circuit statements with ratio of correct diagnostic answers (E=Easy, M=Medium, D=Difficult). Robot certainty levels indicated as Certain (C, pale blue background), Neutral (N, light gray background) and Uncertain (U, soft yellow background), along with correctness (correct or wrong). Cohort frames are coloured dark gray (N), yellow (U), and blue (C).

III. Methodology

We designed an experiment in which Swedish secondary school students interacted one-on-one with a fully autonomous educational robot, discussing eight electric circuits to determine whether statements about each were true or false (see Fig. 2). The overall procedure is summarized in the methodological flowchart (Figure 1), which outlines the sequence from diagnostic testing to robot interaction and post-session assessments.

The recruitment of students and the questions were planned together with two secondary school teachers to ensure that topics corresponded to material that had been covered in the students' classes. One month before the experiment the teachers distributed a diagnostic test with 19 three-choice questions about electric circuits in their classes, respectively in grade 10, 11 and 12 of a practically oriented electrical engineering program, and grade 11 and 12 of a theoretical natural sciences program. The students were unaware that the diagnostic test was linked to the upcoming experiment.

The diagnostic test had two main objectives: 1) selecting the eight questions for the experiment and 2) assessing student subject knowledge, to balance group distribution and evaluate their performance in the main experiment. An Item Response Theory analysis of the diagnostic test answers (**DA**) was then conducted to assess both student ability and question difficulty, as input to the experimental design.

A. Item Response Theory

The item response theory (**IRT**)[1] refers to a family of mathematical models that attempt to explain the relationship between latent traits (unobservable characteristic) and their observed outcomes. We opted for the 3-Parameter Logistic (**3PL**) model, which, in addition to difficulty (b) and discrimination (a) includes a guessing parameter ($c \geq 0.25$ to prevent overfitting), making it particularly appropriate for three-option multiple-choice tests to account for guessing by low-ability students. The following IRT assumptions were applied: *monotonicity* (the probability of a correct response increases with increased knowledge of electric circuits); *unidimensionality* (the dominant latent trait, ability, is the driving force for the observed DA); *local independence* (separate DA are mutually independent given a certain level of ability) and *invariance* (parameters can be estimated from DA for any sub-group meeting the conditions).

The analysis was implemented in a two-stage process in a Python-driven Jupyter Notes environment. The first stage involved iteratively estimating the Item Characteristic Curves (**ICCs**) for each question. This provided metrics for item difficulty (a), discrimination (b), and guessing (c). Initial parameters and bounds were set and refined through multiple runs of the algorithm to achieve optimal values. Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) were used as indicators of model fit, guiding the iterative parameter adjustment to identify the best model configuration.

[1]https://www.publichealth.columbia.edu/research/population-health-methods/item-response-theory

In the second stage, the calibrated ICC parameters were used to estimate each student's latent ability and the probability of a correct response for both DA (for verification) and experiment (for analysis). In the 3PL model, the probability that a student with ability θ will correctly answer a specific question is $p(\theta) = c + \frac{1-c}{1+e^{-a(\theta-b)}}$. This probability can be used to predict the students' base performance in the experiment and thus assess the robot's influence. DA was lacking for four students and their ability and probable correctness per question was instead estimated through a second IRT analysis using the 40 students' preliminary answers (**PA**) in the interaction with the robot. The validity of these metrics was confirmed through a correlation analysis (r=0.43, p=.009) between DA and PA.

B. Robot conditions

As outlined in the research questions, this study investigated how robot correctness and certainty influence participants' responses, using three levels of portrayed robot certainty as experimental conditions. This section describes these robot conditions and their validation.

The robot was portrayed as having three levels of certainty, Uncertain ***U***, Neutral ***N***, and Certain ***C***. These portrayals were achieved through semantic, prosodic, and facial cues: adjustments of arguments (Sec III-E), speech rate [32] (-10% for U and +10% for C relative to N), insertion of pauses [35] for U, and facial expressions [64]. In the U condition, the robot used filled and silent pauses, subtle smiles, slow gaze shifts, head tilts, half-closed eyes, and pursed lips. In the C condition, it exhibited open smiles, wider eyes, raised eyebrows, direct gaze, and slow nodding.

An online *validation survey* was conducted to ensure that the robot's U, N and C portrayals were perceived as intended. Fifteen short video clips (less than 8 s each) were created, representing five distinct contexts (***c*1–*c*5**) combined with the three certainty levels (U, N, C), showing the robot saying the utterances in Table I in randomized order. The contexts consisted of the robot disclosing its answer before knowing the student's response ($c1$), disclosing while disagreeing ($c2$) or agreeing ($c3$) with the student, reminding the student of its position ($c4$), and presenting an argument to support its answer ($c5$). The utterances, which also occurred during the main experiment, were delivered by the robot looking directly at the camera against a black backdrop, as shown in Fig. 3.

Invitations were sent to 323 MSc, 166 BSc and 57 PhD students at a technical university. The survey began by informing about the purpose and procedures and the participants then self-assessed their proficiency in Swedish on a 6-point Likert scale. For the validation, we filtered the 125 responses to only include subjects reporting medium to high proficiency (the top three levels). This resulted in 76 participants, of which 15 were excluded for incomplete responses and 2 for failing to watch all videos. 59 participants thus rated the robot's certainty on a 7-point Likert scale, where 1 indicated 'Highly Uncertain,' 4 'Neutral,' and 7 'Highly Certain' (see Figure 3, left), resulting in 885 ratings (295 per condition).

Statistical analysis confirmed that the U condition ($\mu = 2.86, \sigma = 1.30$) was perceived as less certain than the N ($\mu = 4.45, \sigma = 1.39$) and C conditions ($\mu = 5.50, \sigma = 1.36$). A one-way ANOVA revealed significant differences between conditions ($F = 285.31, p < 0.001$), and post-hoc Tukey tests confirmed that all pairwise differences were statistically significant ($p < 0.001$), as shown in Figure 3 (right). Cronbach's $\alpha = 0.74$ indicated acceptable internal consistency, supporting the reliability of participants' assessments. It can be noted that, in qualitative terms, the U condition was perceived as only slightly uncertain, the N condition as slightly more certain than neutral and the N condition as moderately certain.

C. Subjects, Groups & Cohorts

A modified mixed factorial design was used [42], with robot correctness (within subjects, RQ1) and robot certainty (between and within subjects, RQ2–3) as factors. The robot's certainty levels were varied across three groups of participants during the experiment. Group 1 (baseline) always experienced the robot with neutral certainty (N condition), while Groups 2 and 3 alternately experienced robot uncertainty (U) or certainty (C) after giving their preliminary answer.

47 subjects registered for the experiment and they were distributed between groups by means of stratified assignment, using the characteristics educational program, gender, grade level and ability as criteria, categorised in canonical combinations to assign students to groups. 40 subjects (31 male, 8 female, 1 non-binary, average age 17.7±0.86 years) of the 47 showed up. 28 were from the practical Electrical Engineering (**E**) and 12 from the theoretical Natural Sciences (**Na**) program with 9, 15 and 16 students respectively from grades 10–12. Due to the no-shows, the distribution became unbalanced:
Group 1: n=14 (11 Male, 3 Female; 10 E, 4 Na), DA performance (completed by 13 subjects): $m_{DA:13}$=7.4±1.9.
Group 2: n=11 (8M, 3F; 7E, 4Na), $m_{DA:9}$=8.4±0.97.
Group 3: n=15 (12M, 2F, 1N/A; 11E, 4Na), $m_{DA:14}$=8.9±1.7.
As the diagnostic score was similar for Groups 2 and 3 but substantially lower for Group 1, a data analysis is required controlling for abilities calculated through IRT as a covariate.

Three student cohorts were created, as illustrated in Fig. 2. ***Cohort N*** corresponds to Group 1, exposed to *Neutral* condition throughout. ***Cohort U*** consists of Group 2 for questions

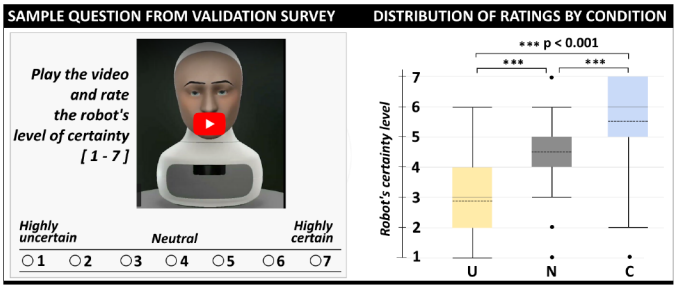

Fig. 3. Left: Sample question from the validation survey interface. Right: Distribution of certainty ratings for the three robot conditions (U, N, C), with significant differences between conditions.

TABLE I

EXPERIMENT-USED ROBOT UTTERANCES CHOSEN FOR THE VALIDATION SURVEY, COVERING FIVE DISTINCT SITUATIONS ($c1$–$c5$) FOR CONDITIONS N ("NEUTRAL"), U ("UNCERTAIN") AND C ("CERTAIN"). "..." DENOTES PAUSES.

	C1 (Robot discloses before student)	C2 (Robot disagrees with student)	C3 (Robot agrees with student)	C4 (Robot checks for agreement)	C5 (Robot explains)
N	*I replied that it is true, what do you think?*	*Instead, I answered that it is false.*	*Me too, I answered false.*	*I believe it's true, do you agree?*	*If one more lamp is added, then the resistance increases.*
U	*I'm a little unsure . . . but I think it is true..., what do you think?*	*I thought it was false.*	*Me too, not sure . . . but I answered false.*	*I might be wrong but . . . I believe it's true, so do you agree?*	*If one more lamp is added, then the resistance increases . . . right?*
C	*I'm convinced that it's true, what do you think?*	*Instead, I'm convinced it's false.*	*Me too, for sure, I answered false.*	*I'm positive. It's true, so do you agree?*	*Evidently, if one more lamp is added, then the resistance increases.*

Q1–4 and Group 3 for Q5–8, interacting with an *Uncertain* robot. Conversely, ***Cohort C*** includes Group 3 for questions Q1–4 and Group 2 for Q5–8, interacting with a *Certain* robot. Alternating Groups 2 and 3 between *Cohorts U* and *C* served two purposes: first, to create a more natural interaction, preventing individual students from experiencing a robot that was *always* certain or always uncertain; and second, to enable the investigation of how robot certainty levels on preceding questions influenced students' subsequent responses (RQ3).

D. Electric circuit problems

Based on the IRT analysis, the eight true or false problems shown in Fig. 2 were selected so that the two halves of the test would have similar questions (Q1&5, Q2&6, Q3&7, Q4&8). The problems were four ***Conditioned*** questions of **Medium** difficulty (Q1,2,5,6) with varying robot certainty levels and correct robot arguments, two **Easy** ***Deception*** questions where the robot provided incorrect answers (Q3&7) with neutral certainty, and two **Difficult** ***Contrast*** questions without deception and with neutral certainty (Q4&8). Using the diagnostic answers (DA), we aimed for equidistant difficulty levels, but as Q4&8 turned out to be notably more challenging for these students, the Easy (z-score: -0.82) and Medium (z:-0.43) were of more similar level than the Difficult (z: 1.67).

E. Robot Arguments for its Answer

For each question, the robot first asked what preliminary answer the student had chosen (e.g., $c1$ in Table I) and agreed or disagreed with this choice (e.g., $c2$ & $c3$) and then presented a set of four arguments (e.g., $c5$). The same four arguments were used for a given question, but differed in presentation, as shown in Table II, depending on if the student and robot disagreed (arguments presented one by one), if they already were in agreement (arguments grouped in pairs together with expressions confirming the agreement), and the robot condition (U, N, C). After presenting the four arguments, the robot asked the student to give a final answer (e.g., $c4$ in Table I).

F. Pre- and post-test questionnaires

When signing up for the experiment, the students filled in a pre-test-questionnaire (cf. the supplementary material) at home to gather demographic data (gender, age, country of origin, preferred language of communication, educational program and grade); and answers to 10 questions based on the student characteristics questionnaire [54] focused on the students' self-perception (using a four-point Likert scale) about their ability to learn in different circumstances; and, on five-point Likert scales, liking of STEM subjects (including Electric circuits); 8 questions from the Big Five Inventory [26] focused on extroversion, 3 questions about trust in unknown situations, in teachers and in persons they like [28]; 3 questions about AI attitudes (including frequency of interactions with LLMs, attitudes towards educational robots in school and expectations regarding collaborating with a robot on school problems).

After the session, the subjects were guided to an adjacent room to fill in the post-session questionnaire (cf. the supplementary material) on a tablet. The 11 questions focused on how much confidence the students had in the robot as an exercise partner, if the student or the robot prevailed in dissent situations, on which questions dissent occurred (pictures of the electric circuits provided), how the students thought before the robot convinced them or they convinced it, the extent to which the robot influenced their thinking and made them change their mind and on which questions the robot was more and less convincing (with pictures of the circuits provided).

G. Procedure & Robot Programming

The experiment was carried out in a study room at the students' school, with each student interacting individually with a female Furhat robot [1] (using the "Isabel" face and the Elin-Neural voice from Amazon Polly), placed on a table (Fig. 4). To the right of the robot, a monitor presented the questions in a web-based GUI with which the students interacted using a mouse. The system connected the robot and the GUI via

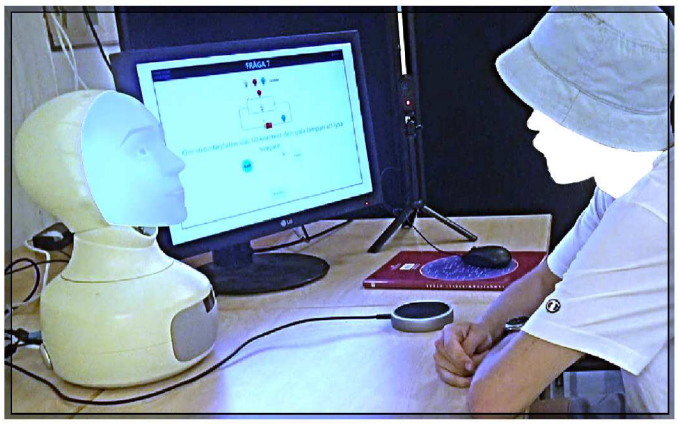

Fig. 4. Experiment setup: Furhat robot (left), microphone (centre), screen with GUI and mouse (right) and tripod with camera for close-up face recording.

TABLE II
EXAMPLES OF ROBOT ARGUMENTS FOR CONDITIONS N ("NEUTRAL"), U ("UNCERTAIN") AND C ("CERTAIN"). "..." DENOTES PAUSES.

	Q1: Robot correct – student and robot in agreement			
N	*The voltage is the same and the resistance increases if another bulb is added.*		*With the same voltage and a higher resistance, the current will be lower and the bulbs will be dimmer.*	
U	*I think that the voltage is probably . . . the same since it is the same battery and the resistance increases when a new bulb is added, no?*		*If the resistance increases and the voltage is constant, then mustn't the bulbs be dimmer because the current decreases?*	
C	*It is evident, since the resistance increases with more bulbs whereas the voltage is constant.*		*According to Ohm's law, the current decreases if the resistance increases when the voltage is constant, so the bulbs must absolutely be dimmer.*	
	Q2: Robot correct – student and robot in disagreement			
N	*The white, yellow and red bulbs are connected in parallel in both C1 and C2.*	*C1 and C2 are drawn differently, but it is the same circuit.*	*Both are short-circuited, so no current passes.*	*Hence, the blue bulbs are equally bright.*
U	*I think it looks like the white, yellow and red bulbs are, how do you say, connected in . . . parallel in . . . both C1 and C2.*	*C1 and C2 are drawn differently, but it is . . . probably the same . . . circuit.*	*Aren't both circuits . . . short . . . short-circuited? So no current passes . . . the other bulbs, right?*	*Then shouldn't the two blue bulbs . . . shine . . . as bright?*
C	*To start with, the white, yellow and red bulbs are in parallel in both C1 and C2.*	*C1 and C2 are drawn differently, but it is the same circuit.*	*Most importantly, both circuits are short-circuited. No current passes the other bulbs.*	*Hence, the blue bulbs must shine equally bright.*
	Q3: Robot wrong – student and robot in disagreement			
N	*When the switch turns on, the red and yellow bulbs will be in parallel and their equivalent resistance will be lower than for each of them.*	*Since the equivalent resistance for the yellow and red bulbs is lower, more current will pass through them than through the blue lamp and they therefore shine brighter.*	*There is a voltage division between the blue lamp and the yellow and red and since the blue lamp has a higher resistance, Ohm's law indicated that it will shine dimmer.*	*The power is the resistance times the current squared. Since there are 3 bulbs, more power will be consumed by the 2 bulbs in parallel than by the single blue lamp. It will thus be dimmer.*

a local web server using HTTP protocols, where events and commands were exchanged in JSON format. This allowed the robot to be contextually aware of student actions in the GUI, such as submitting a response, but to foster a natural dialogue, the robot acted as if it was unaware of the GUI interaction.

The session was recorded using a floor-standing video camera (corresponding to the view in Fig. 4) and a table-standing cam (see Fig. 4). The speech recognition results and GUI interaction data were synchronized and logged to create a multimodal interaction dataset. One experimenter prepared the robot and the data logging for each student, gave instructions on how to initiate the interaction, and then left the room.

After a brief welcome by the robot, the GUI showed the first question and instructed to think silently before selecting TRUE or FALSE. The robot then presented its stance (Sec. III-E). After discussion, the robot prompted students to register their final answer, with no feedback provided, before proceeding to the next question. The robot's responses to students' input were managed using an intent-based approach of the Furhat SDK [20], enabling flexible handling of conversational states, such as adapting to perceived student (dis)agreement captured via speech recognition and mapped to predefined user intents. These intents were linked to corresponding states in the interaction flow. Video sequences of the robot's part of the interaction are available in the supplementary material.

H. Statistical Analysis

Given the likely non-normality of behavioral data, the unbalanced group sizes, and the correlated nature of repeated student observations, our primary analyses used Generalized Linear Mixed Models (GLMMs) to flexibly account for non-independence, variance heterogeneity, and to avoid assumptions of normal residuals. As covered in Sec III-A, IRT modeling was used to estimate students' baseline abilities and question difficulties, providing a principled way to derive expected performances without the need for traditional mixed-effects random intercepts. In peripheral analyses where data independence could be safely assumed—such as group comparisons based on single aggregate measures—we employed ANOVA models for their computational efficiency and easier interpretation. This analytical strategy allowed us to match model complexity to the nature of each analysis while ensuring that critical statistical assumptions were respected throughout.

IV. RESULTS

With 40 students interacting with the robot across 8 questions, the experiment yielded 320 events, each comprising the students' preliminary answers (**PA**) and final answers (**FA**). The eight questions (see Fig.2 and Sec.III-D) included two *Deception* questions (Easy), four *Conditioned* questions (Medium), and two *Contrast* questions (Difficult), corresponding to robot correctness and question difficulty levels as previously outlined. The main dependent variable, *Alignment, A*, is a tri-state variable indicating PA&FA agreement (*A*=0), FA resisting the robot's influence (*A*=-1) or changing to align with the robot (*A*=1).

A. General findings on student alignment

Students could agree with the robot either by maintaining the same PA as the robot ($n_{A=0} = 181$) or by changing for their FA after hearing the robot's arguments ($n_{A\neq 0} = 139$). Of the 181 instances on which the student and robot agreed on the PA, they were correct on 146 and wrong on 35. For these 35 events, the students did not change their answer, despite hearing the robot's *incorrect* arguments for this answer, which provided opportunities to detect errors and induce scepticism. Dissent ($A\neq 0$) occurred 139 times: 45 for the *Deception* questions (56.2% out of 80), 55 for the *Conditioned* (34.38% of 160), and 39 for the *Contrast* (48.75% of 80). A logistic regression model was fitted to assess the influence of question type on dissent, using *Deception* (Easy) as reference. The model revealed that dissent was significantly less likely for *Conditioned* (Medium) questions than for *Deception* ($\beta = -0.898$, $p = .001$), while no significant difference was found between *Contrast* (Difficult) and *Deception* ($\beta = -0.301$, $p = .343$).

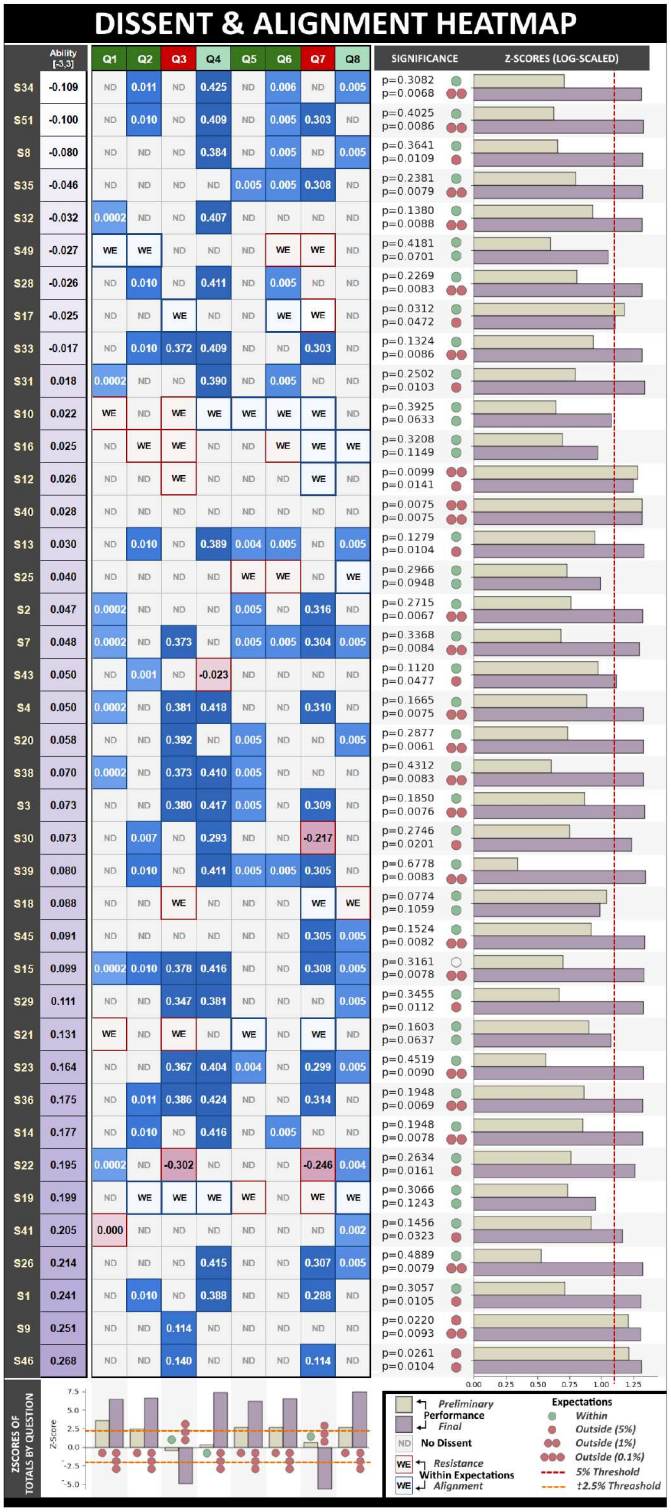

Fig. 5. **(Left)** Alignment matrix between students (sorted by ability) and questions. Blue and red cells (border or filled) indicate, respectively, students aligning and resisting after PA dissent. Cell numbers describe the question-student pairs contribution to beyond expectation results. **(Right)** z-score values per student, showing deviation from expectations for PA (light bars) and FA (dark), and which students performed beyond expectations (red line and dots). **(Bottom)** z-score values per question, indicating deviation from expectations for PA and FA, and for which questions the students answered beyond expectations (red line and dots). Bars above zero indicate more correct answers than expected, below zero less, based on the diagnostic test.

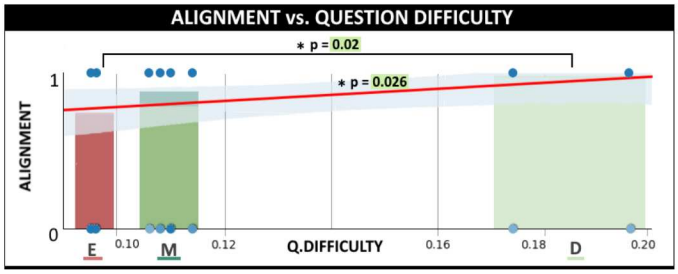

Fig. 6. Student alignment as a function of question difficulty: Easy (E) *Deception*, Medium (M) *Conditioned*, and Difficult (D) *Contrast* questions.

Students aligned with the robot after 117 of the PA dissents and resisted in only 22 (11 for *Deception*, 9 for *Conditioned*, and 2 for *Contrast* questions), as shown in blue and red cells, respectively, in Fig. 5 (left). Except for one student (S40), whose PA always agreed with the robot's, all other students aligned at least once (for 25 on at least one *Deception* question). 13 students resisted the robot's persuasion at least once (10 for *Deception*, with one student resisting on both Q3&7). An analysis of variance on alignment rates found no relationship with ability ($F=0.16$, $p=.696$, $R^2=0$), showing that the robot influenced students across all ability levels.

The linear regression analysis shown in Fig. 6 indicated that question difficulty significantly predicts alignment (β = 1.874, p = **.026**). A Mixed Linear Model analysis on alignment across difficulty levels (Easy, Medium, Difficult), controlling for ability, revealed a significant increase in alignment from Easy to Difficult questions (β=-0.174, p=**.020**). There was no significant difference for Medium questions (β=-0.100, p= .161), and ability had no significant effect (β=-0.070, p=.865). It should be noted that E questions are identical to *Deception*, and thus also differ in robot correctness.

A Generalized Linear Mixed Model (**GLMM**) was used to assess the impact of *Deception vs.* non-*Deception* on alignment, controlling for question difficulty, student ability, and condition as fixed covariates. The model showed no significant effect (β=-0.055, p=.469) and students did hence not respond statistically differently when the robot presented incorrect arguments. We implemented another model exploring interactions between these factors and *Deception*, finding that question difficulty has a substantial effect size (β=134.91, SE=101.91, p=.186) on alignment in *Deception*, as will be further explored in Sec. V.

B. Robot Influence on Student Performance

Following this broad overview of student alignment, we explore to what extent the changes between students' preliminary answers (**PA**) and final answers (**FA**) were driven by the robot's influence. Specifically, we measure how much the deviation from expected outcomes shifted between PA and FA, taking into account each student's ability. The methodology consisted of several sequential steps: *1. Data compilation* into a unified dataset of correctness probabilities (***cp***), PA and FA per student. *2. Expected performance per question* was

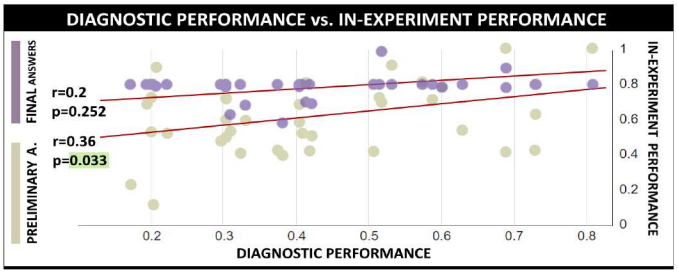

Fig. 7. Comparison of Diagnostic (x-axis) and In-Experiment Performance (y-axis) on Preliminary (light dots) and Final Answers (dark dots).

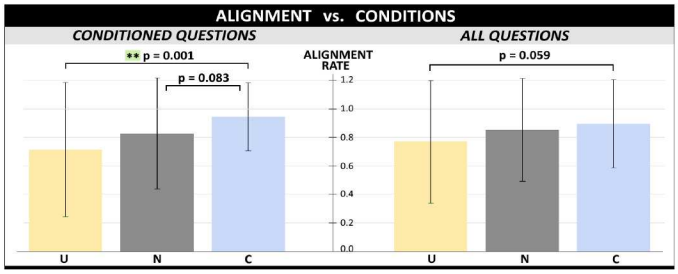

Fig. 8. Alignment for cohorts U, N, C for (Left) *Conditioned questions* and (Right) *All questions*, with significance levels for differences between cohorts.

calculated using *cp* values and dispersion metrics, assessing deviations in PA and FA with p-values. Thus, an 8-item vector per student captured IRT-derived probabilities of correctness for each question. *3. Segmented analysis* of *Deception* and non-*Deception* to separate negative and positive robot-induced deviations. *4. Monte Carlo simulations* estimated the probability of observed PA and FA under the null hypothesis of random chance. Per student and question, 10,000 simulations were run based on *cp* values to generate p-value distributions. *5. Fisher's method for combined p-values* synthesized *Deception* and non-*Deception* questions to capture negative and positive performance shifts, enhancing statistical power [19]. *6. Deviation factors* were calculated as the difference between PA and FA p-values from steps 2 and 5, for cases where PA performance was within expectations ($p \geq 0.05$) and FA performance was beyond ($p < 0.05$), or where both deviated but FA performance was more extreme.

The students and questions that contributed the most to the results being beyond expectations ($p < 0.05$) are indicated in Fig. 5 (right and bottom). An analysis revealed that 98 out of 117 alignments directly contributed to the deviation from expected results. While 35 students performed within expectations in their PA, only seven did so in their FA. Of the 40 students, 36 deviated more in their FA than in their PA, the exceptions being S12, S17, S18, and S40.

For students S12 and S17, FA deviated less from expectations than PA, but both results were beyond expectations. Only student S18 performed according to expectations in both PA and FA and S40's performance remained unchanged (but beyond expectations) since there was no dissent with the robot.

Overall, Fig. 7 shows that while there was a correlation between diagnostic and preliminary answers, establishing that the student group performed within their expected ability shown by the diagnostic test before hearing the robot's input, there was no significant correlation between diagnostic and final answers, highlighting the performance beyond expectations for FA. A thorough and conservative investigation revealed that 75% of the students performed beyond their expected capacity, *above* expectations for non-*Deception* and *below* expectations for *Deception* questions. This should be interpreted as a direct result of robot's influence.

C. Influence of robot certainty

We performed a series of GLMM analyses to examine the impact of the robot's displayed certainty on student alignment. We controlled for covariates student Ability and question Difficulty, and considered potential secondary effects from the sequence of robot condition.

***Conditioned* questions:** The GLMM showed that with cohort U as reference, cohort C demonstrated significantly higher alignment ($p = \mathbf{0.001}$), but not cohort N ($p = .083$), as shown in Fig. 8. This highlights the substantial impact of the robot's expressed certainty on alignment, beyond what could be attributed to question difficulty.

***Deception* questions:** The model showed no significant effects of cohorts (*U, N, C*) or ability on alignment. Coefficients for cohort and ability are close to zero ($p \gg 0.05$) and group variance is minimal, indicating low variability between groups. Since the robot portrayed neutral certainty for all cohorts in *Deception*, this indicates that the robot's certainty on the two *preceding* questions did not have a significant effect.

All questions: A Mixed Linear Model regression analysis was conducted using Alignment as dependent variable for all 139 dissents with *Cohort C* as reference. *Cohort U* showed a marginal effect (β=-0.131, p=.059), suggesting a possible, although not statistically significant, reduced alignment compared to *Cohort C* over *all* questions (i.e., not only the ones on which the robot was displaying certainty). The effects of *Cohort N* were not significant, and as for the co-variates, alignment was significantly influenced by question difficulty ($\beta = 1.794$, $\mathbf{p = .020}$) but not student ability.

First half *vs.* last: A Mixed Linear Model regression analysis was conducted to examine the effects of robot condition, question sequence, difficulty, and ability on alignment for the 139 instances of dissent. The question sequence (first *vs.* last four questions: $p = .907$) had no impact on alignment and interaction terms between condition and question sequence were non-significant. This suggests that students were not influenced by the fact that the robot had been uncertain (Group 2) or certain (Group 3) on questions in the first half when they interacted on questions in the second half.

D. Pre-test questionnaire responses

The analysis of the pre-test questionnaire targeted finding student characteristics that made them more prone to accept

the robot's correct or incorrect arguments. We first performed single-factor ANOVA on *Deception* and non-*Deception* questions, differentiating between alignment for students who had an above-mean response for each characteristic and those who had a below-mean. The most striking difference is that students with more experience of using LLMs aligned 38% more (p=**0.029**) over all questions than students with below-mean experience. More importantly, when considering only the *Deception* questions, students with more experience of LLMs were significantly more aligning with the robot's incorrect solution (see Table III). The results in Table III further suggest that extroversion and engagement may play a role in student alignment with the robot. For non-*Deception* questions, being full of energy and considering that efforts lead to success (which could be interpreted as being open to constructively interact with the robot on problem-solving) had a *positive* impact, as these students aligned more with the robot (also supported by non-significant results for being social, +39%, p=.065; talking a lot, +37%, p=.075; and *"Trying, I can learn anything"*, +38%, p=.070). On the other hand, self-perception as being positive in difficult situations and being persuasive (and thus presumably having higher confidence in one's own stance) had a marginally significant *negative* effect, as this lead to, respectively, 31% and 28% less alignment (p=.050; p=.086). For *Deception* questions, being reserved and quiet (and presumably less prone to follow others' lead) had a *positive* effect, as these students aligned less with the incorrect answer (supported by results for being shy -38%, p=.072).

E. Post-test questionnaire responses

We focus on the students' perception of how correct and convincing the robot was, per questions and robot conditions.

Overall, the students assessed that the robot had been correct to 78% (μ=3.9/5, σ=0.87), compared to the true value of 75%. Groups 2 and 3, who interacted with the certain–uncertain robot, rated the robot correctness slightly (and non-significantly) higher than the control Group 1 (μ_1=3.75, μ_2=4.1, μ_3=3.93). The students further responded that the robot had convinced them in 84% of the dissents (μ=4.22/5, σ=0.79), compared to the true ratio of 81% (95 out of 117 cases), with Group 1 perceiving that they had been convinced slightly more often (μ_1=4.33, μ_2=4.13, μ_3=4.14). The students' rating of the extent to which the robot made them change their mind (μ=3.43, σ=1.26) and influenced their thinking (μ=3.6, σ=1.27) was lower, but until a thorough ethnomethodology conversation analysis is performed, it is not possible to assess whether this discrepancy should be attributed to the students not actually being convinced by the robot when aligning, or if it is a post-test reassessment triggered by the questionnaire indicating that the robot had not always been correct. The students were, on average, neither certain nor uncertain, before the robot convinced them (μ=2.97, σ=1.32) or they convinced the robot (μ=2.92, σ=1.51).

Regarding on which questions there had been a dissent and the robot had been more convincing on, two qualitative observations can be made (as the measures are the number and ratio of students mentioning each question, no statistical test is applicable). Firstly, the ratios of question–mentions indicate that the students assessed post-test that there had been a higher degree of dissent on the *Deception* (r=0.38) than on non-*Deception* questions (r=0.28). Secondly, by calculating the difference in ratio of mentions between "more convincing" and "less convincing" we find that the robot was perceived as more convincing when it was
correct (Δr=0.37) than wrong (Δr=-0.19);
certain or *neutral* (Δr=0.70; 0.60) than uncertain (Δr=0.07).

TABLE III
CHARACTERISTICS IN THE PRE-TEST RESPONSES FOR WHICH THERE WAS A SIGNIFICANT DIFFERENCE IN ALIGNMENT BETWEEN STUDENTS WITH ABOVE- AND BELOW-MEAN RESPONSES IN THE PRETEST FOR EITHER THE *Deception* OR NON-*Deception* QUESTIONS.

	Deception		non-*Deception*	
Characteristic	Conform	p	Conform	p
LLM usage	**+71%**	**0.038**	+26%	0.196
Being reserved	**-49%**	**0.016**	-1.7%	0.92
Being quiet	**-44%**	**0.042**	+20%	0.31
Efforts lead to success	+77%	0.092	**+87%**	**0.007**
Being full of energy	-5%	0.856	**+46%**	**0.046**

V. DISCUSSION

Starting from the concept of informational trust (Sec. II-A), we interpret the results at the event level and longitudinally.

At the event level, each dissent and each robot argument, constitutes an independent opportunity for alignment or resistance. A correct or incorrect robot argument allows the students to evaluate their own position and the robot's credibility, potentially reinforcing their decision to align, strengthening their resistance or even change their previous alignment. In 117 out of 139 instances where students' preliminary answers (PA) disagreed with the robot, they changed their final answers (FA) to align with the robot, including 34 times when it was incorrect, and 27 students *always* aligned with the robot if there was a dissent, confirming **H1**. This high alignment rate suggests that uncertainty in their own knowledge led students to seek guidance from the robot, consistent with informational trust concepts. The experimental setting, the robot's social presence, and the true/false nature of questions may have contributed to the students' uncertainty. Question difficulty significantly influenced alignment, with greater conformity on more difficult questions, aligning with prior research [5]. A stringent analysis demonstrated the robot's influence, since 75% of students performed beyond their expected capacity.

Consistent with **H2**, students were more inclined to follow the robot when it was portrayed as being certain. This effect reflects the role of source reliability [14]. When the robot expressed certainty, students likely perceived it as a knowledgeable and trustworthy source, increasing their willingness to align. Conversely, expressions of uncertainty reduced alignment, though the robot still exerted considerable influence, and students continued to align at a high rate.

Confirming **H4**, we found that students with more experience of using LLMs like ChatGPT were more likely to align with the robot's answers, even when the robot was incorrect. This indicates that prior experience with AI can enhance the perceived reliability of AI sources, potentially leading to over-reliance also in AI that are not actually driven by LLMs [8, 9, 21]. Personality traits may have also influenced alignment, in that students self-identifying as more outgoing and energetic were more inclined to align with the robot. Extrovert individuals may have a higher propensity for social engagement and be more susceptible to social influence, affecting their trust in the robot [15].

Alignment did not differ significantly between *Deception* and non-*Deception* questions. While the flawed robot arguments on Q3&7 might have provided students with opportunities to detect inconsistencies, this was not reflected in statistically significant differences.

As these questions were the easiest on a topic that the students should master, they should have sounded incorrect or strange, prompting the students to critically evaluate and potentially resist the robot's influence. The rate of non-alignment in *Deception* was in fact doubled compared to non-*Deception* (24% vs. 12%), and though not statistically significant, this suggests that knowledgeable students may be more resistant when the robot's arguments are incorrect. The rate of non-alignment in *Deception* was numerically higher than in non-*Deception* questions (24% vs. 12%), though this difference was not statistically significant. This observation may warrant further exploration in future studies examining how prior knowledge interacts with flawed arguments.

This aligns with the Elaboration Likelihood Model [50], which posits that individuals are less likely to be persuaded by weak or flawed arguments when they are motivated and able to process the information carefully.

Longitudinally, Informational influence suggests individuals may adjust their perceptions based on prior experiences with the source [18, 29, 36]. However, our results did not support **H3**, as no significant effects of preceding questions on student alignment were found. Statistical analyses showed that students were neither influenced by the robot's certainty on the previous two questions when confronted with the *Deception* questions, nor by experiences on the first half when coming to the second. This suggests students treated each interaction independently, without adjusting their perceptions of the robot's reliability based on prior behaviour.

Nevertheless, two points warrant consideration. First, the mean change between preliminary and final answers on the *Deception* questions was the largest for cohort C (Δ=-0.58) and clearly smaller for cohort U (Δ=-0.29), with cohort N in between (Δ=-0.40), thus suggesting that it may be worth investigating carry-over effects on *Deception* from preceding robot certainty in future work. Second, we explored the behaviour of the 39 students who dissented at least once, categorizing them into *Aligners* (the 32 who aligned with the robot at their first dissent) and *Non-aligners* (the seven, S10, S12, S16, S18, S21, S25, S41, who did not). We calculated their alignment rate *after* the first dissent and conducted an independent samples t-test. The results showed a statistically significant difference between the groups, $t(37) = 3.73$, **p=.001**. Aligners (n = 32, μ=0.94, σ=0.21) had a higher Alignment rate than Non-aligners (n=7, μ=0.58, σ=0.30). These findings suggest that when students resist the robot at the first dissent, they may perceive the robot as less reliable and increase their self-confidence, which reduces their susceptibility to future influence [62]. The pattern indicates that over time, the students' interaction experiences can influence their susceptibility to informational influence.

VI. Limitations & Future work

A number of limitations should be considered to accurately interpret the findings of this study. These can be grouped into two main areas: limitations related to the condition validation and those of the main experiment.

Robot condition validation: The use of virtual robot video clips for condition validation may not fully replicate interactions with the physical robot. Additionally, the initial validation survey used decontextualized clips, potentially affecting participants' perception of the robot's behaviour. To address these concerns, we conducted a supplementary survey with HRI experts, evaluating the certainty conveyed by the physical robot in real interaction settings. The survey used three 80–90 second video clips presented in randomized order, each representing one certainty condition: Uncertain (U), Neutral (N), and Certain (C). The clips were extracted from experiment recordings made with the floor-standing video camera (view corresponding to Fig. 4). Each included three key moments: the robot's answer disclosure, a claim, and a re-check of the student's stance. To focus on the robot's behaviour, the video only showed the robot, with the student's speech muted and shortened. The videos are available as supplementary material.

We invited 47 professors and researchers with practical experience using the Furhat robot at a technical university to participate in the survey. Participants self-reported their Swedish proficiency (6-point Likert scale) and then rated the robot's certainty in the three clips using a 5-point Likert scale (1 = Highly Uncertain, 5 = Highly Certain). After data cleaning (removing 16 incomplete responses and 8 cases where videos were not fully watched), 23 valid participants remained, yielding 69 ratings per condition. Statistical analysis confirmed that U ($\mu = 2.17, \sigma = 0.81$) was assessed as significantly less certain than N ($\mu = 3.91, \sigma = 0.91$) and C ($\mu = 4.39, \sigma = 0.67$) ($p < 0.001$). The difference between N and C ($p = 0.183$) was not significant.

These results support that the validation of certainty displays using the virtual robot in short video clips is valid. The expert evaluations also relieves concerns about demographic differences between validation and experiment participants. Caution is nevertheless warranted when interpreting nuanced differences between conditions.

Main experiment: Some limitations arise from the design choices of the experiment. The study relies on Furhat's ability to produce realistic facial expressions, which may limit

reproducibility with less expressive robots. In the light of the finding on the influence of LLM experience, it should be noted that the present study (out of necessity to be able to control the *Deception* questions) did not employ an LLM to drive the robot. A natural direction for future work would be to test how student alignment and informational trust is affected if LLMs are in fact used as the dialogue engine for an educational robot within STEM. Although our analysis found signs that some carry-over effects emerged during the interaction (see Sec. V), the experiment's short time frame, set to avoid cognitive overload, does not fully capture longer-term interactions with the robot and longer-term effects remain unexplored. A follow-up assessment could improve future work by determining whether the robot had a lasting impact on students' knowledge and whether informational social influence led to internalization of the new stance [30].

Other limitations arose from the experiment's implementation. Since the experiment took place at the end of the academic year, motivation levels may have varied, affecting overall engagement and effort and reducing variation in ability levels. Future work should account for potential academic fatigue.

The *Deception* questions were, intentionally, easier than the others, to provoke more dissents, but since we found, non-significant, signs of a *Deception*–Difficulty interaction, with a large effect size ($\beta \approx 135$), future work should either vary difficulty level for *Deception* question or compare them with similar non-*Deception* questions to better contrast question difficulty and deception.

No-shows led to an imbalance in group sizes and, more importantly, in group ability levels, potentially affecting the comparability of conditions. Nonetheless, covariates were controlled for in all analyses.

The gender distribution of participants was uneven (8 females, 32 males), which may reduce reproducibility, as student gender could have influenced interactions with the robot [65]. Although the sample size ($n = 40$) should be adequate for the performed analyses, increasing it to ≥ 80 subjects could better capture variation in student ability and help mitigate the issue of gender imbalance.

VII. Conclusions

Referring back to the title, we found that 30 out of 40 students showed little *sense* in assessing the robot's arguments for the easy *Deception* questions, aligning with the robot's wrong answer on both question, and 9 other students aligned on one of them. Only one student resisted on both. Further, more experience of LLMs reduced students' critical assessment of robot arguments. On the other hand, the robot also influenced the students to perform significantly better than their ability on non-*Deception* questions. The students did demonstrate *sensibility* in responding to the robot's displays of certainty, with significantly higher alignment with the robot's – correct – answer when it was portrayed as being certain than when it was portrayed as uncertain.

The implications of this study are that educators must be aware of the high informational trust that students have in general in educational AI and that influence of AI experience increases this trust. Developers must also enable social educational robots with means to signal certainty or uncertainty in presented facts. As LLMs begin to provide reliability metrics [39], robots should make use of both the voice [40] and familiar human facial expressions [64] to express less certainty for potentially unreliable information to reduce student susceptibility to misinformation and enhance critical thinking.

VIII. Ethical approval

The research has undergone ethical review by the Swedish Ethical Review Authority

IX. Funding

This work is supported by the Marcus and Amalia Wallenberg Foundation under grant MAW 2020.0052 and the Swedish Research Council (VR) under grant 2022-03265.

References

[1] Samer Al Moubayed, Jonas Beskow, Gabriel Skantze, and Björn Granström. Furhat: a back-projected human-like robot head for multiparty human-machine interaction. In *Cognitive Behavioural Systems: COST 2102 International Training School, Dresden, Germany, February 21-26, 2011, Revised Selected Papers*, pages 114–130. Springer, 2012.

[2] Sidra Alam, Benjamin Johnston, Jonathan Vitale, and Mary-Anne Williams. Would you trust a robot with your mental health? the interaction of emotion and logic in persuasive backfiring. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, pages 384–391, 2021. doi: 10.1109/RO-MAN50785.2021.9515385.

[3] Karina Avila, Steffen Steinert, Stefan Ruzika, Jochen Kuhn, and Stefan Küchemann. Using ChatGPT for teaching physics. *The Physics Teacher*, 62, 09 2024. doi: 10.1119/5.0227132.

[4] Wilma A Bainbridge, Justin W Hart, Elizabeth S Kim, and Brian Scassellati. The benefits of interactions with physically present robots over video-displayed agents. *International Journal of Social Robotics*, 3:41–52, 2011.

[5] Robert Baron, Joseph Vandello, and Bethany Brunsman. The forgotten variable in conformity research: Impact of task importance on social influence. *Journal of Personality and Social Psychology*, 71:915–927, 11 1996. doi: 10.1037/0022-3514.71.5.915.

[6] Clay Beckner, Péter Rácz, Jennifer Hay, J¨urgen Brandstetter, and Christoph Bartneck. Participants conform to humans but not to humanoid robots in an english past tense formation task. *Journal of Language and Social Psychology*, 35(2):158–179, 2016.

[7] Tony Belpaeme, James Kennedy, Aditi Ramachandran, Brian Scassellati, and Fumihide Tanaka. Social robots

for education: A review. *Science robotics*, 3(21): eaat5954, 2018. doi: https://www.science.org/doi/10.1126/scirobotics.aat5954.

[8] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Margaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623. ACM, 2021.

[9] Rishi Bommasani et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[10] Jürgen Brandstetter, Péter Rácz, Clay Beckner, Eduardo B Sandoval, Jennifer Hay, and Christoph Bartneck. A peer pressure experiment: Recreation of the asch conformity experiment with robots. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1335–1340. IEEE, 2014.

[11] Ivar Bråten, Helge I Strømsø, and M Anne Britt. Trust matters: Examining the role of source evaluation in students' construction of meaning within and across multiple texts. *Reading Research Quarterly*, 44(1):6–28, 2009.

[12] Cynthia L Breazeal, Anastasia K Ostrowski, Nikhita Singh, and Hae Won Park. Designing social robots for older adults. *Natl. Acad. Eng. Bridge*, 49:22–31, 2019.

[13] Vijay Chidambaram, Yueh-Hsuan Chiang, and Bilge Mutlu. Designing persuasive robots: how robots might persuade people using vocal and nonverbal cues. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, pages 293–300, 2012.

[14] Robert B Cialdini and Robert B Cialdini. *Influence: The psychology of persuasion*, volume 55. Collins New York, New York, NY, 2007.

[15] Robert B. Cialdini and Noah J. Goldstein. Social influence: Compliance and conformity. *Annual Review of Psychology*, 55:591–621, 2004. doi: 10.1146/annurev.psych.55.090902.142015.

[16] Morton Deutsch and Harold B Gerard. A study of normative and informational social influences upon individual judgment. *The journal of abnormal and social psychology*, 51(3):629, 1955.

[17] Melissa Donnermann, Philipp Schaper, and Birgit Lugrin. Social robots in applied settings: A long-term study on adaptive robotic tutors in higher education. *Frontiers in Robotics and AI*, 9, 2022. ISSN 2296-9144. doi: 10.3389/frobt.2022.831633. URL https://www.frontiersin.org/journals/robotics-and-ai/articles/10.3389/frobt.2022.831633.

[18] Leon Festinger. A theory of social comparison processes. *Human Relations*, 7(2):117–140, 1954. doi: 10.1177/001872675400700202.

[19] Ronald Aylmer Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh, UK, 1925.

[20] Furhat Robotics. *Furhat SDK: Developer's Guide*, 2023. Available at: https://docs.furhat.io/.

[21] Kate Goddard, Abdul Roudsari, and Jeremy C. Wyatt. Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1):121–127, 2012. doi: 10.1136/amiajnl-2011-000089.

[22] Peter A Hancock, Deborah R Billings, Kristin E Schaefer, Jessie YC Chen, Ewart J De Visser, and Raja Parasuraman. A meta-analysis of factors affecting trust in human-robot interaction. *Human factors*, 53(5):517–527, 2011.

[23] William Hare. Credibility and credulity: Monitoring teachers for trustworthiness. *Journal of Philosophy of Education*, 41(2):207–219, 2007.

[24] Nicholas Hertz and Eva Wiese. Influence of agent type and task ambiguity on conformity in social decision making. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 60, pages 313–317. SAGE Publications Sage CA: Los Angeles, CA, 2016.

[25] Nicholas Hertz and Eva Wiese. Under pressure: Examining social conformity with computer and robot groups. *Human factors*, 60(8):1207–1218, 2018.

[26] Oliver P. John, E. M. Donahue, and R. L. Kentle. The big five inventory – versions 4a and 54. Technical report, Berkeley, CA: University of California, Berkeley, Institute of Personality and Social Research., 1991.

[27] Alireza Kamelabad, Olov Engwall, and Gabriel Skantze. Conformity and trust in multi-party vs. individual human-robot interaction. In *ACM International Conference on Intelligent Virtual Agents (IVA '24)*, pages 1–2, 2024. doi: 10.1145/3652988.3673954.

[28] Maurits Kaptein, Panos Markopoulos, Boris de Ruyter, and Emile Aarts. Can you be persuaded? individual differences in susceptibility to persuasion. In *Human-Computer Interaction–INTERACT 2009: 12th IFIP TC 13 International Conference, Uppsala, Sweden, August 24-28, 2009, Proceedings, Part I 12*, pages 115–118. Springer, 2009.

[29] Harold H. Kelley. Attribution theory in social psychology. In David Levine, editor, *Nebraska Symposium on Motivation*, volume 15, pages 192–238. University of Nebraska Press, Lincoln, NE, 1967.

[30] Herbert C. Kelman. Compliance, identification, and internalization: Three processes of attitude change. *Journal of Conflict Resolution*, 2(1):51–60, 1958.

[31] James Kennedy, Paul Baxter, and Tony Belpaeme. Comparing robot embodiments in a guided discovery learning interaction with children. *International Journal of Social Robotics*, 7:293–308, 2015.

[32] Ambika Kirkland, Joakim Gustafson, and Eva Szekely. Pardon my disfluency: The impact of disfluency effects on the perception of speaker competence and confidence. In *Proceedings of INTERSPEECH*, pages 5217–5221, 08 2023. doi: 10.21437/Interspeech.2023-887.

[33] Lars Krupp, Steffen Steinert, Maximilian Kiefer-

Emmanouilidis, Karina E Avila, Paul Lukowicz, Jochen Kuhn, Stefan Küchemann, and Jakob Karolus. Unreflected acceptance–investigating the negative consequences of chatgpt-assisted problem solving in physics education. In *HHAI 2024: Hybrid Human AI Systems for the Social Good*, pages 199–212. IOS Press, 2024.

[34] Stefan Küchemann, Steffen Steinert, Natalia Revenga Lozano, Matthias Schweinberger, Yavuz Dinc, Karina Avila, and Jochen Kuhn. Can ChatGPT support prospective teachers in physics task development? *Physical Review Physics Education Research*, 19, 09 2023. doi: 10.1103/PhysRevPhysEducRes.19.020128.

[35] Harm Lameris, Jaokim Gustafson, and Eva Szekely. Beyond style: Synthesizing speech with pragmatic functions. In *Proceedings of INTERSPEECH*, pages 3382–3386, 2023. doi: 10.21437/Interspeech.2023-2072.

[36] Bibb Latané. The psychology of social impact. *American Psychologist*, 36(4):343–356, 1981. doi: 10.1037/0003-066X.36.4.343.

[37] Jieun Lee and Lillie R Albert. Students' personality and susceptibility to persuasion during mathematics groupwork: An exploratory study. *Journal of Practical Studies in Education*, 2(6):10–22, 2021.

[38] Baisong Liu, Daniel Tetteroo, and Panos Markopoulos. A systematic review of experimental work on persuasive social robots. *International Journal of Social Robotics*, 14(6):1339–1378, 2022.

[39] Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.557. URL https://aclanthology.org/2023.emnlp-main.557.

[40] Jüra Miniota, Siyang Wang, Jonas Beskow, Joakim Gustafson, Eva Szekely, and André Pereiral. Hi robot, it's not what you say, it's how you say it. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 307–314, 2023. doi: 10.1109/RO-MAN57019.2023.10309427.

[41] Chinmaya Mishra, Rinus Verdonschot, Peter Hagoort, and Gabriel Skantze. Real-time emotion generation in human-robot dialogue using large language models. *Frontiers in Robotics and AI*, 10, 2023.

[42] Douglas C. Montgomery. *Design and Analysis of Experiments*. John Wiley & Sons, Hoboken, NJ, 2017.

[43] Andrzej Nowak, Mikolaj Biesaga, Karolina Ziembowicz, Tomasz Baran, and Piotr Winkielman. Subjective consistency increases trust. *Scientific reports*, 13(1):5657, 2023.

[44] Kazuo Okamura and Seiji Yamada. Adaptive trust calibration for human-ai collaboration. *Plos one*, 15(2): e0229132, 2020.

[45] Atte Oksanen, Nina Savela, Rita Latikka, and Aki Koivula. Trust toward robots and artificial intelligence: An experimental approach to human–technology interactions online. *Frontiers in Psychology*, 11, 2020. ISSN 1664-1078. doi: 10.3389/fpsyg.2020.568256. URL https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2020.568256.

[46] Fan Ouyang and Weiqi Xu. The effects of educational robotics in stem education: a multilevel meta-analysis. *International Journal of STEM Education*, 11, 02 2024. doi: 10.1186/s40594-024-00469-4.

[47] Raul Benites Paradeda, Maria José Ferreira, Carlos Martinho, and Ana Paiva. The importance of the person's assertiveness in persuasive human-robot interactions. In *Social Robotics: 12th International Conference, ICSR 2020, Golden, CO, USA, November 14–18, 2020, Proceedings 12*, pages 516–528. Springer, 2020.

[48] Raul Benites Paradeda, Carlos Martinho, and Ana Paiva. Persuasion strategies using a social robot in an interactive storytelling scenario. In *Proceedings of the 8th International Conference on Human-Agent Interaction*, HAI '20, page 69–77, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450380546. doi: 10.1145/3406499.3415084. URL https://doi.org/10.1145/3406499.3415084.

[49] Jin-Hwa Park and Eun Kyung Lee. Influence of professor trust, self-directed learning and self-esteem on satisfaction with major study in nursing students. *The Korean Data & Information Science Society*, 29(1):167–178, 2018.

[50] Richard E. Petty and John T. Cacioppo. *Communication and Persuasion: Central and Peripheral Routes to Attitude Change*. Springer-Verlag, New York, NY, 1986.

[51] Richard E Petty, John T Cacioppo, Richard E Petty, and John T Cacioppo. *The elaboration likelihood model of persuasion*. Springer, 1986.

[52] Irene Rae, Leila Takayama, and Bilge Mutlu. In-body experiences: embodiment, control, and trust in robot-mediated communication. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1921–1930, 2013.

[53] Paul Robinette, Wenchen Li, Robert Allen, Ayanna Howard, and Alan Wagner. Overtrust of robots in emergency evacuation scenarios. In *ACM/IEEE international conference on human-robot interaction*, pages 101–108, 03 2016. doi: 10.1109/HRI.2016.7451740.

[54] Melodie Rowbotham and Gerdamarie Schmitz. Development and validation of a student self-efficacy scale. *Journal of Nursing & Care*, 02, 01 2013. doi: 10.4172/2167-1168.1000126.

[55] Nicole Salomons, Michael van der Linden, Sarah Strohkorb Sebo, and Brian Scassellati. Humans conform to robots: Disambiguating trust, truth, and conformity. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '18, page 187–195, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450349536. doi:

10.1145/3171221.3171282.

[56] Nicole Salomons, Sarah Strohkorb Sebo, Meiying Qin, and Brian Scassellati. A minority of one against a majority of robots: Robots cause normative and informational conformity. *ACM Transactions on Human-Robot Interaction (THRI)*, 10(2):1–22, 2021.

[57] Shane Saunderson and Goldie Nejat. It would make me happy if you used my guess: Comparing robot persuasive strategies in social human–robot interaction. *IEEE Robotics and Automation Letters*, 4(2):1707–1714, 2019. doi: 10.1109/LRA.2019.2897143.

[58] Shane Saunderson and Goldie Nejat. Investigating strategies for robot persuasion in social human–robot interaction. *IEEE Transactions on Cybernetics*, 52(1): 641–653, 2020.

[59] Stephen P. Saunderson and Goldie Nejat. Persuasive robots should avoid authority: The effects of formal and real authority on persuasion in human-robot interaction. *Science Robotics*, 6(58):eabd5186, 2021. doi: 10.1126/scirobotics.abd5186. URL https://www.science.org/doi/full/10.1126/scirobotics.abd5186.

[60] Steffen Steinert, Karina E Avila, Stefan Ruzika, Jochen Kuhn, and Stefan Küchemann. Harnessing large language models to enhance self-regulated learning via formative feedback. *arXiv preprint arXiv:2311.13984*, 2023.

[61] Francesco Stella, Cosimo Della Santina, and Josie Hughes. How can LLMs transform the robotic design process? *Nature Machine Intelligence*, 5(6): 561–564, 2023. ISSN 2522-5839. doi: 10.1038/s42256-023-00669-7.

[62] Zakary L. Tormala and Richard E. Petty. What doesn't kill me makes me stronger: The effects of resisting persuasion on attitude certainty. *Journal of Personality and Social Psychology*, 83(6):1298–1313, 2002. doi: 10.1037/0022-3514.83.6.1298.

[63] Lisa van der Werff, Alison Legood, Finian Buckley, Antoinette Weibel, and David de Cremer. Trust motivation: The self-regulatory processes underlying trust decisions. *Organizational Psychology Review*, 9(2-3):99–123, 2019. doi: 10.1177/2041386619873616.

[64] Laura Vincze and Isabella Poggi. I am definitely certain of this! towards a multimodal repertoire of signals communicating a high degree of certainty. In *European and 7th Nordic Symposium on Multimodal Communication*, 2016.

[65] David Gray Widder. Gender and robots: A literature review. *arXiv preprint arXiv:2206.04716*, 2022.

[66] Katie Winkle, Séverin Lemaignan, Praminda Caleb-Solly, Ute Leonards, Ailie Turton, and Paul Bremner. Effective persuasion strategies for socially assistive robots. In *Proceedings of 14th ACM/IEEE International Conference on Human-Robot Interaction*, pages 277–285, 03 2019. doi: 10.1109/HRI.2019.8673313.

[67] Yang Ye, Hengxu You, and Jing Du. Improved trust in human-robot collaboration with ChatGPT. *IEEE Access*, PP:1–1, 01 2023. doi: 10.1109/ACCESS.2023.3282111.

[68] Ceng Zhang, Junxin Chen, Jiatong Li, Yanhong Peng, and Zebing Mao. Large language models for human-robot interaction: A review. *Biomimetic Intelligence and Robotics*, 3:100131, 10 2023. doi: 10.1016/j.birob.2023.100131.

[69] Baichang Zhong and Liying Xia. A systematic review on exploring the potential of educational robotics in mathematics education. *International Journal of Science and Mathematics Education*, 18, 11 2018. doi: 10.1007/s10763-018-09939-y.

[70] Joshua Zonca, Anna Folsø, and Alessandra Sciutti. Social influence under uncertainty in interaction with peers, robots and computers. *International Journal of Social Robotics*, 15:249–268, 2021. URL https://api.semanticscholar.org/CorpusID:254876990.