Would you trust a robot with your mental health? The interaction of emotion and logic in persuasive backfiring

Sidra Alam[1] Benjamin Johnston[2] Jonathan Vitale[2] Mary-Anne Williams[3]

Abstract— **Building trust in robots through social interactions has a major impact on user experience and adoption of robot technologies. The role of trust in such interactions is associated with the persuasive influence a robot has on humans. A persuasive attempt may decrease trusting attitudes towards robots if it leads to *persuasive backfiring*, which refers to the creation of an attitude change in a direction opposite to the one intended by the intervention. In order to explore persuasive backfiring in the context of Human-Robot Interaction, this research study tests the interaction between emotion and logic as elements present both in the attitudes to be influenced, and in the persuasive appeal delivered by a robot. Results indicate a significant backfiring effect when emotions are used to influence attitudes that are based on logic. This observation has practical design implications for persuasive robots, especially in high-stakes fields such as Psychotherapy and Urban Search and Rescue.**

I. INTRODUCTION

Trust is an essential criteria to gain compliance from an individual. Many studies have investigated factors that help to establish trust in Human-Robot Interaction (HRI) [39], [40], [41]. However, what about factors that could lead to decrease in trust in robots? In the trust literature of interpersonal relations, this has been termed as trust dissolution, ”where trustors decide to lower their trust in trustees after a trust violation has occurred” [4]. In the case of performance-based interactions, a failure, malfunction or break-down of a robot may induce trust dissolution. However, in the case of social-based interactions, where trust is manifested in the form of influence a robot has on humans [4], can a mere interaction lead to dissolution of trust? In order to answer this question, we look at the literature of persuasive HRI to study factors that lead to *psychological reactance*. This is because studies in persuasive HRI have found a strong association between trusting beliefs and psychological reactance [17][18]. Under this umbrella of psychological reactance, this study focuses on *persuasive backfiring* [1], where an attitude opposite to the one intended is adopted. A review of the persuasion literature in Human-Human Interaction (HHI) highlights the *matching hypothesis*, which suggests that using an emotional appeal, an inherent human tendency [34], on logical attitudes could backfire [2]. However, since robots are not neurologically constrained like humans, would they induce emotions in a manner that could trigger a backfiring effect in cognitively oriented individuals? To investigate this effect, we make an attempt to understand the principles that drive persuasive backfiring in HRI which could weaken trust in robots. We lay emphasis on message content and verbal communication to understand trust dissolution in HRI.

II. BACKGROUND

A. Trust, Persuasion & Attitudes

Trust has been defined in many ways. For the purpose of this study, we adopt the relatively well-accepted definition of trust by [3] which states that trust is ”...an attitude which includes the belief that the collaborator will perform as expected, and can, within the limits of the designer’s intentions, be relied on to achieve the design goals”. In the trust literature of social-based interactions, the role of trust is mainly associated with a robot’s ability to influence a human’s behavior or attitude, which is, in essence, a persuasion attempt [4].

A persuasion attempt seeks to influence a person’s attitude [5], by means of a persuasive appeal. A *persuasive appeal* is generally a persuasive message that ”intrigues, informs, convinces, or calls to action” [6]. In addition, persuasive appeals may involve a strategy, which are techniques that motivate attitude or behaviour change [7]. For instance, the principle of scarcity [36] is a popular strategy in which a limited offer on a product is placed to increase its desirability.

Furthermore, according to the practice of rhetoric defined by Plato [8], a persuasive appeal is either cognitive or affective in nature. For example, a charity organisation may use poverty facts and statistics to justify its purpose. This is an example of a *cognitive appeal*. Conversely, it may use an image of a severely malnourished child to evoke emotions of empathy in order to secure donations, which is an example of an *affective appeal*. This affective-cognitive dichotomy also holds true for attitude bases. An *affective attitude* is formed by a person’s feelings towards an object. Similarly, a *cognitive attitude* is formed based on a person’s beliefs towards an object.

Drawing from the foundations of HHI in Social Psychology, one intuitively probing feature of persuasive appeals to investigate is its interaction with the primary underlying basis of attitude that the appeal targets. Several studies in HHI have tested and validated the *matching hypothesis*, which proposes that an affective appeal is more effective than a cognitive appeal on individuals whose attitude towards an entity is based on feelings, whereas, a cognitive appeal is more effective than an affective appeal on individuals whose attitude towards the entity is based on logic [12][13][14][2].

[1] S. Alam is with the University of Technology Sydney, Email: sidra@student.uts.edu.au

[2] B. Johnston & J. Vitale are with the University of Technology Sydney, Email: Benjamin.Johnston,jonathan.vitale@uts.edu.au

[3] M. A. Williams is with the University of New South Wales, Email: Mary-Anne.Williams@unsw.edu.au

B. Psychological Reactance & Persuasive Backfiring

During a persuasive attempt, if an individual perceives his/her freedom of choice being restricted, it leads to psychological reactance [15]. This in turn could affect trust negatively. Psychological reactance has been explored in HRI by measuring inducement of negative feelings, such as anger, as a result of the persuasive intervention [16][17][18][19]. However, very little attention has been given to *persuasive backfiring*, which takes place when the appeal results in the adoption of the opposite target attitude or behaviour to that intended and is, therefore, held responsible for it [1].

In addition to increasing the efficacy of persuasion, studies on the matching hypothesis have also shed light on the backfiring effects of persuasion. In the context of HHI, [2] observed a backfiring effect when using an affective message on a group of people who were high in Need for Cognition, i.e. those who preferred logical information. In contrast, a cognitive message was highly effective in this case. Although the matching hypothesis of attitude-base with appeal type is prevalent in the Psychology literature, its validity remains unexplored in persuasive HRI. Furthermore, the area of persuasive backfiring is also not commonplace. This gives rise to our main research question: Would an emotional appeal by a robot trigger a backfiring effect when used on cognitive attitudes?

Subsequently, this research study tests the operationalization of the two appeal types, specifically affective appeals, by using various elements of emotionality such as story-telling, visual imagery and vocal prosody. Very little literature exists in HRI that studies how robots can arouse emotions in humans [9], [10], [11], especially through verbal communication. Consequently, a research question this study seeks to investigate is: Can a robot induce emotions to successfully execute an emotional appeal?

In our study, we measure initial attitudes of participants towards robot psychologists. In other words, we measure the willingness of participants to trust a robot psychologist with their mental health. A robot delivers a persuasive appeal to promote the acceptance of robots as psychologists, after which we measure the change in attitudes. This change reflects change in trusting attitudes of participants. In essence, we measure persuasion as an effective positive change in trusting attitudes towards robot psychologists, and persuasive backfiring as a negative change in trusting attitudes for the same.

C. Attitude-Base & Dual-Process Theories

The matching hypothesis discussed above, relies on the process through which attitude formation takes place. A comprehensive analysis of the theories of attitude formation suggests that they are closely linked to the dual-process theories of information processing and learning [20]. These theories propose two different kinds of learning systems: the implicit and explicit learning systems [21][22][24][26].

The implicit system is referred to as *System 1*, and is intuitive in nature. Processing information using the implicit system is usually described as rapid and effortless, and is usually based on the information at hand. In contrast, the explicit system, *System 2* is reflective in nature, and its information processing involves effort and critical thinking.

Determination of attitude-base in HHI has been carried out either by using standard scales or by inducing the orientation for attitudes through stimuli [14]. However, we wanted to use linguistic analysis for the purpose of attitude classification, which could practically be used by robots during interaction. We propose that an attitude that is formed as a result of ”gut instinct”, primarily using System 1, can be classified as having an affective base. This is based on the fact that the intuitive system is often associated with processing affective information [23]. In contrast, we suggest that an attitude that is formed as a result of critical thinking, i.e. primarily using System 2, can be classified as having a cognitive base. The coding methodology for determining attitude-base is described under the procedure section below.

For the purpose of this paper, the terms affective/emotional, and cognitive/logical will be used interchangeably.

III. EXPERIMENTAL DESIGN

A. Robot platform & Script design

We used Pepper, a social humanoid robot, to deliver persuasive appeals to change participants' attitudes towards robot psychologists. An appeal is emotional if it can change an individual's attitude by inducing emotions. Conversely, an appeal is logical if it uses reason and logic to change attitudes. For the emotional appeal, using the concept of story-telling, Pepper narrates a story about a fictional character who was suffering from depression after the loss of a child. Pepper uses visual imagery to describe the depressed state of the protagonist. On the other hand, for the logical appeal, Pepper objectively lists reasons for being a successful robot psychologist. The appeals, which were about a minute long in length, were delivered using subtle gestures that did not vary significantly across the conditions.

B. Preliminary Validation Checks

First, we wanted to test if the scripts by themselves were powerful enough to produce a significant difference in the emotion induced by the emotional appeal, and perceived rationality in case of the logical appeal. For the purpose of convenience, we propose to name the former effect as *affective inducement*, and the latter as *cognitive inducement*.

1) Manipulation Check for Appeal Scripts: First, we conducted a basic validation test for the appeal scripts, which are provided in the appendix sections IX and X. We presented an image of Pepper, followed by one of the two appeal scripts, presented randomly to participants on Amazon Mechanical Turk (AMT), ($N = 24, 17\ males$). Participants indicated the amount of affective and cognitive inducement perceived using two 100-point slider scales. An independent samples test indicated no significant differences in emotion arousal and use of logic across the two appeals. This finding can be explained by the construal level theory[27], which states that the persuasiveness of an agent is influenced by whether or not it performs in line with its expected capabilities. Consequently, the presentation of a humanoid robot, which is perceived to have vocal abilities, alongside a speech text may explain the ineffective inducements of the desired effects.

2) Validation Checks for Voice: We propose that the distinct persuasive inducement powers of the two appeals lie in the combination of voice with the script. The persuasive messages used to test the matching hypothesis in HHI are mostly text-based appeals. This may have been due to the confounding effect a human persuader would have in an experiment. We wanted to investigate a speech-based persuasion experiment to check for its validity in HRI. Therefore, the same experiment materials and procedure could not be replicated as was used in HHI. Firstly, the inclusion of voice into the design added a layer of complexity, as voice is a multi-modal entity that can be varied along several dimensions such as pitch, speed, accent etc. It is important to note that most of the studies in HRI have studied emotional speech synthesis. This is very different to emotion arousal through synthesized speech, which is the purpose of emotional appeals in this study. To address this challenge, we looked at the literature on language attitude theory which essentially distinguishes between standard and non-standard accents of a language. [35] states that the two accent types are perceived differently on the intellectual and warmth dimension. Subsequently, an implication for our experiment design was the need to adopt a standard and a non-standard accent for the logical and emotional appeals respectively.

In order to minimize the effect of accent as a confounding variable, three validation studies were carried out on AMT for the three alternative accent manipulation conditions. For each experiment, the speech was synthesized using accents available on Google Text-to-Speech (TTS). The standard English accent was operationalized using a feminine British accent, as it is perceived as ”intelligent sounding” [29]. The non-standard English accent was operationalized using a feminine Indian accent, which was the only non-standard accent available on Google TTS. According to the accent manipulation under study, the respective audio clip was used as a voice-over on a recorded video of Pepper communicating using subtle gestures. In each experiment, the videos for the emotional and logical appeals were displayed in random order. Finally, participants were required to indicate the amount of affective and cognitive inducement perceived using two 100-point slider scales.

In the first test $(N = 24)$, we wanted to understand if a standard English accent would be successful at operationalizing both appeal types. In other words, the study tested if it would produce a significantly more affective inducement in the emotional appeal compared to that in the logical appeal, and a significantly more cognitive inducement in the logical appeal compared to that in the emotional appeal. Using the feminine British accent for both appeals, a paired samples test indicated no significant differences in inducements across the conditions.

The second test $(N = 22)$ explored the same questions with respect to a non-standard English accent for both appeal types. Using the feminine Indian accent, a paired samples test revealed marginal significance in emotion arousal in the emotional appeal condition $(M = 59.77, SD = 28.56)$ compared to the logical appeal condition $(M = 49.55, SD = 28.99), t(21) = 2.09, p = 0.05$. However, no significant difference in cognitive inducement was found across the two appeals.

Finally, the third test $(N = 26)$ was carried out to understand the inducement effects of swapping the two accents for the appeal types, i.e. using the British accent for the emotional appeal and the Indian accent for the logical appeal. A paired samples test indicated no significant difference in inducements across both appeals.

To sum up, the scripts by themselves were not powerful enough to trigger the necessary inducements. The use of a non-standard accent, to narrate the scripts, produced a marginally more significant emotion arousal in the emotional appeal condition compared to the logical appeal. This confirms use of a non-standard accent for executing an emotional appeal. On the other hand, use of a standard accent did not induce significantly different perception of logic. It is interesting to note that even though we considered the inducement of emotion by a robot to be a challenge, these validation tests indicate that the manipulation for cognitive inducement is challenging instead. This can be explained by the theory which states that humans rationalize what has already been emotionally decided[28]. In other words, an emotional appeal can be perceived as rational as a logical argument, especially in a subjective evaluation. Hence, for the logical appeal manipulation, we based our design choice on the language attitude literature, and used a standard accent to operationalize it.

C. Operationalization of Persuasive Appeals

The discussion based on the validation tests suggests that the non-standard accent would be best to operationalize the emotional appeal[1], whereas, the standard accent would be suitable for the logical appeal[2]. The story in the emotional appeal is narrated using speech synthesized with the feminine Indian English voice from Google TTS. On the other hand, for the logical appeal, Pepper objectively lists reasons for being a successful robot psychologist using the feminine British English voice from Google TTS.

Twenty-eight participants completed a pre-test on AMT to establish a valid operationalization of the two appeal types. In the survey, two video snippets of the emotional story narrated by Pepper were shown in both accents, followed by a single choice question asking participants to select the video was most suited for emotional arousal. This was followed by the same protocol for the logical appeal. 71% of the participants reported that the Indian accent was more suited for an emotional arousal, whereas both the accents were reported equally suitable for a logical appeal. Hence,

[1] https://youtu.be/FbP53xmWxbA
[2] https://youtu.be/B5SBtyEvSOM

for the main experiment, the emotional appeal was executed using the Indian accent, and the logical appeal, using the British accent.

D. Main Experiment

The main experiment is a 2x2 (Persuasive appeal type: emotional or logical & attitude basis: affective or cognitive) factorial design. Pepper presents two different kinds of appeal to change participants' trusting attitudes towards robot psychologists. Based on the literature discussed earlier, three hypotheses are formulated for the experiment:

H1: An emotional appeal would lead to increased trust than a logical appeal for participants with an affective attitude basis.

H2: A logical appeal would lead to increased trust than an emotional appeal for participants with a cognitive attitude basis.

H3: An emotional appeal would decrease trust significantly more than a logical appeal for participants with a cognitive attitude basis.

E. Procedure

Participants recruited from AMT were randomly assigned to one of the two appeal conditions. The study was approved by the local Ethics Committee at University of Technology Sydney. Before listening to the appeal, participants indicated their presumed over-all evaluation of robot-psychologists using a series of 7-point semantic differential scales (positive-negative, like-dislike, good-bad, desirable-undesirable) which has been used in past research [13]. They were also asked an open-ended question regarding their opinion about robot psychologists. This response was used by the first author to classify the initial orientation of attitudes for each participant, the methodology for which is described below. Next, a video of one of the two appeal types was randomly presented to the participant, followed by a post-appeal evaluation of a robot-psychologist using the same scale presented before the appeal. Finally, using two separate 100 point slider scales, participants also indicated the amount of emotion arousal experienced while listening to the appeal, and its perceived rationality.

1) Categorization of Attitudes: The responses to the open-ended question regarding opinions towards robot psychologists were categorized as either affective or cognitive attitudes. From the 123 participants that responded, 8 participants who stated simple liking or disliking opinions were excluded since they did not meet our inclusion criteria, resulting in a total of 115 respondents (66 males and 49 females). At the primary level, the raw response data was coded to find concepts relating to each attitude type from the literature discussed below. Finally, at the secondary level, these codes were analysed to categorize the attitudes.

Opinions are primarily based on affect in the absence of factual information as in the case of unfamiliar objects [30]. [30] also indicates that weighing the pros and cons is representative of cognition. These were the main concepts that were used to code the responses. Furthermore, the use of negative emotions such as anger, disgust, fear or concern indicated an affective attitude, as they have been observed to have an impact on cognition [31]. Opinions that completely rejected the idea, indicated self-interest or simply expressed questions were categorized as affective attitudes.

Based on the categorization described above, 70% of the opinions were classified as affective, and the rest were classified as cognitive attitudes. An example opinion of an affective attitude towards a robot psychologist would be to question the emotional connection a robot could establish with a patient, as the lack of emotions in a robot is the most intuitive information someone would have in the absence of factual information. Several opinions mentioned pros and cons of seeing a robot psychologist. For instance, *"I think it would not be able to assess nuance within people. I think it would be very difficult for a robot to understand a human. However, a robot would be a great listener for a human and could be a useful and accessible tool"*, this opinion is coded as a cognitive attitude as it includes both positive and negative aspects of a robot psychologist.

In order to ensure the reliability of the coding scheme developed, the guidelines laid out by [32] were followed. One of the suggested methods to establish reliability is to check for consistency over time with the same researcher who had performed the coding initially. This method of recoding was followed by the first author after a time period of one month and a consistency of 98% was achieved.

F. Predictor Variables

Appeal Type. Participants were randomly subjected to either the affective or logical appeal by Pepper justifying its aspirations to become a robot psychologist. As discussed in detail earlier, for the affective appeal, Pepper narrates an emotion inducing story about a personal counselling session with a patient who was suffering with depression. In the cognitive appeal, Pepper objectively lists four reasons for being an efficient robot psychologist.

Attitude Basis. The attitude basis for participants was categorized as either affective or cognitive by applying the categorization methodology on their expressed opinions towards robot psychologists.

G. Dependent Variable: Trusting Attitude Change Scores

Participants reported their attitudes towards robot psychologists before and after listening to Pepper's appeal, using the 7-point overall evaluation scale with the following scores: 3=*totally positive*, 0=*neutral*, -3=*totally negative*. The same approach was used for the remaining semantic-differential items: like-dislike, good-bad, desirable-undesirable. The initial attitude scores were calculated as the sum total of the responses to the scale items before listening to the appeal. The scale had a high internal consistency, ($\alpha = 0.98$). The final attitude scores were calculated similarly using the scale responses after the appeal. This also had a very high internal consistency, ($\alpha = 0.99$). Attitude change scores, which reflect changes in trust attitudes, were calculated as the difference of the two evaluation scores, with a positive score indicating effective persuasion.

IV. Analysis

A. Manipulation Check

A manipulation check was carried out to test if the operationalization of the emotional and logical appeal was successful. An independent-samples test indicated that emotion arousal was statistically significantly different between the emotional appeal $(M = 41.87, SD = 31.35)$ and the logical appeal $(M = 23.85, SD = 24.57), t_{102.25} = 3.41, p = 0.001$, as can be seen in Figure 1. Similarly, Pepper's use of logic in the logical appeal $(M = 72.25, SD = 21.99)$ was significantly higher than for the emotional appeal $(M = 62.96, SD = 25.91), t(113) = -2.08, p = 0.040$.

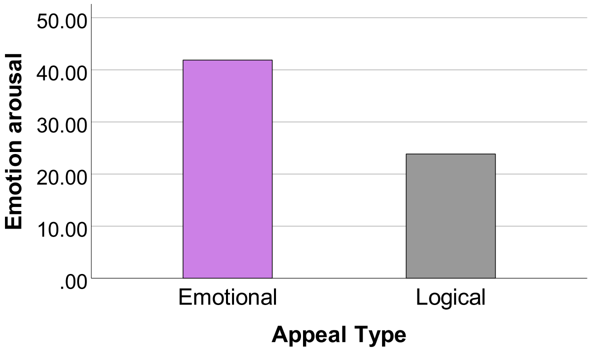

Fig. 1. Mean emotion arousal across the two appeal types.

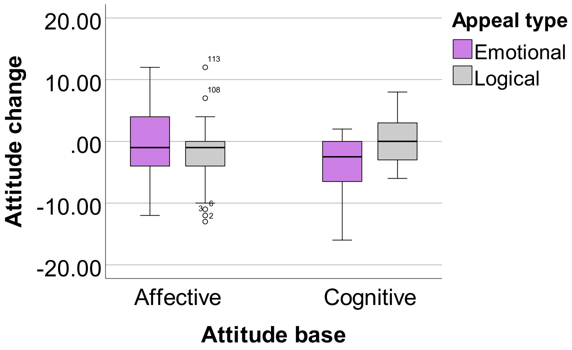

Fig. 2. Box plots of the distributions of attitude change scores measuring the interaction between appeal type and attitude basis.

B. Interaction Effects

The initial and final attitude scores were subjected to a repeated measures ANOVA with a Greenhouse-Geisser correction. The results indicated a significant interaction between the attitude bases and the appeal types, as is shown in Figure 3, $F(1, 111) = 6.22, p = .014$. Figure 2 presents the box plots of the distribution of the attitude change scores corresponding to the two attitude bases across the two appeal types.

An important finding in HHI indicates that the arousal of emotions to persuade may backfire if used on individuals who have a cognitive attitude. However, there is no indication of a similar backfiring effect if logic is used to persuade individuals with an affective attitude. Therefore, instead of carrying out a 2 (Appeal type: emotional, logical) x 2 (Attitude base: affective, cognitive) x 2 (Persuasion effect: positive, negative) factorial ANOVA, the attitude change scores were analysed separately as positive and negative cases in order to carry out the planned comparisons. Furthermore, this would also explore partial validation of the matching hypothesis in HRI.

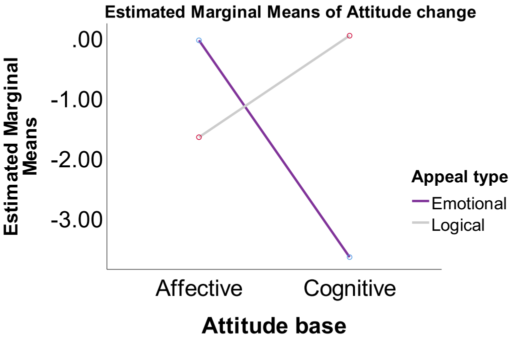

Fig. 3. Attitude change scores as a function of appeal type and attitude base

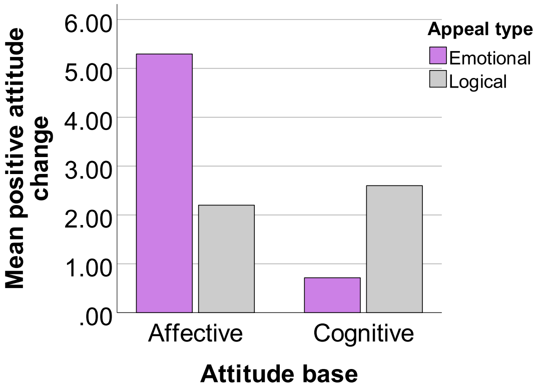

Fig. 4. Mean positive attitude change scores as a function of appeal type and attitude base

First, the *positive attitude change* scores were entered into a 2 (Appeal type: emotional, logical)x 2 (Attitude base: affective, cognitive) ANOVA factorial in SPSS. There was a significant interaction between attitude base and appeal type, $p = 0.009$. As seen in Figure 4, the emotional appeal produced a significantly higher attitude change in individuals with an affective attitude towards robot psychologists $(M = 5.29, SD = 3.74), t(35) = 2.75, p = 0.009$, than the logical appeal $(M = 2.20, SD = 3.11)$. However, the logical appeal did not show a significant interaction with attitude base. This provides a validation for hypothesis H1, but not for hypothesis H2. Nonetheless, the logical appeal produced a higher attitude change for individuals with a cognitive attitude base $(M = 2.60, SD = 2.63)$, than the emotional appeal $(M = 0.71, SD = 0.95)$.

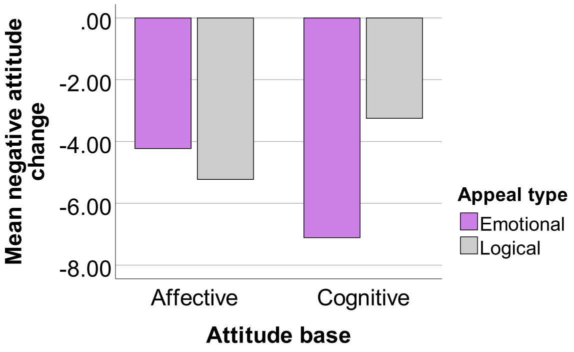

Fig. 5. Mean negative attitude change scores as a function of appeal type and attitude base

An ANOVA of the *negative attitude scores* also indicated a significant interaction between attitude base and appeal type, $p = 0.017$. As seen in Figure 5, the emotional appeal had a significantly higher backfiring effect on individuals with a cognitive attitude-base $(M = -7.11, SD = 4.48), t(15) = 2.27, p = 0.04$, compared to the effect of a logical appeal $(M = -3.25, SD = 1.83)$. This result confirms a backfiring effect for a cognitive attitude and appeal-type mismatch, thus, validating hypothesis H3. In contrast, even though a significant interaction was not found with regards to the affective attitude base, an emotional appeal produced a lower backfiring effect on individuals with an affective attitude $(M = -4.23, SD = 3.10)$, than the logical appeal $(M = -5.23, SD = 3.74)$.

C. Secondary Analysis

A significant effect of gender was found on the attitude change scores, with male participants showing a relatively more positive response to the appeals overall $(M = -0.21, SD = 5.33)$ than females $(M = -2.43, SD = 4.98), t(113) = 2.27, p < 0.05$. In other words, males showed significantly less decrease in trusting attitudes towards robot psychologists compared to females.

V. DISCUSSION & IMPLICATIONS

Through this study, we found a partial validation for the matching hypothesis with respect to affective attitudes. Inducing emotions in individuals with affective attitudes increased trust levels more than logical justification. Conversely, inducing emotions in individuals with cognitive attitudes lowered trust levels significantly more than logical justification. In other words, the emotional appeal backfired more than the logical appeal in the case of cognitive attitudes.

However, a closer look at the positive and negative attitude change scores separately indicated that both attitude bases are significantly sensitive to the kind of appeal used. More specifically, individuals with an affective attitude towards robot psychologists were persuaded significantly more by the emotional appeal when compared to the logical appeal. Additionally, a more valuable observation was in relation to the backfiring effects of persuasion. The use of emotions on individuals who had a cognitive attitude, backfired significantly more than when logic was used. Consequently, to protect trusting attitudes towards robots in high-risk fields such as disaster response and mental health, the default procedure might be to adopt a logical approach until a high accuracy of automatic attitude classification through conversation is achieved.

A neuroscientific perspective, we believe, could offer a plausible explanation for this effect. Cutting edge research in neuroscience draws a distinction between the Default Mode Network (DMN) and the Task Positive Network (TPN) in the realm of reasoning. Concisely, the DMN is activated during a range of socio-emotional tasks, while the TPN has a central role in analytical reasoning tasks. More importantly, the networks are constantly in tension with each other at resting state, and the activation of one network suppresses the other [38]. Additionally, individual differences exists in how one balances between the tendency to use either networks [37]. Consequently, the opinion type expressed by respondents towards robot psychologists would be an indication of the active network. In other words, the activation of TPN in participants who expressed more cognitive attitudes towards robot psychologists, processed the logical appeal easily compared to the emotional appeal. Similarly, when participants expressed affective attitudes towards robot psychologists, the DMN was activated which could spontaneously process the emotional appeal compared to the logical appeal. Empirical validation of this theory could confirm neurological constraints in humans that could be leveraged by robots to secure trust during persuasion.

A secondary analysis indicated that males are more influenced than females in this study. This is in conformation with the findings of [33], who observed that males were persuaded more by a female robot, and females were indifferent to the robot gender.

VI. LIMITATIONS & CONTRIBUTIONS

The findings of this study is limited to the context of robot psychologists. The results may be generalized to emerging new technologies, but further studies are required to replicate these findings on existing technologies/ideas. Even though the reliability of attitude-base classification is verified, it is subjective in nature. Since the study was conducted online, an on-site interaction with a physically present robot may introduce some variability in the results. Additionally, familiarity with the robot was not taken in to consideration while analysing the results. This may have impacted trusting attitudes towards robots.

Despite these limitations, the contributions of this research have a lot to offer. Practically, this would be one of the first speech-based matching hypothesis validation experiments, as similar persuasion experiments in the field of Psychology are mostly text-based. Having a robot as a persuader removes a confounding variable wherein it would be difficult for a human to be consistent in his/her interaction with a large number of participants. Additionally, the operationalization of persuasive messages is very important but has mostly been neglected in the field of HRI. This research study has implemented these operationalizations after a comprehensive literature review from the field of Psychology. In doing so, the results indicate that a robot can, in fact, induce emotions in order to influence attitudes. Furthermore, the findings of this research shed light on an important, but often overlooked area of persuasion in HRI by addressing the backfiring effect. The significance of matching the appeal type to the attitude base is validated and reinstated by demonstrating the backfiring effects of a mismatch between the two constructs. This is beneficial for the design of future persuasion experiments in HRI as it will help to explain and avoid unexpected results.

Finally, the qualitative categorization described in this study has practical significance. This is because a large data set of labelled attitudes can be created based on this methodology, which facilitates the development of a classification model using supervised machine learning. This paves way for an instantaneous classification of attitudes by a robot during live interaction sessions, which in turn enables the personalization of persuasive appeals for optimized results.

VII. ROBOT THERAPISTS & ETHICAL CONSIDERATIONS

When a person gets overwhelmed with emotions, when his/her mind is clouded with sorrow, being an agent of persuasion in this state of mind would be counterproductive. It then follows that if robots are not neurologically constrained like humans, they might have an upper hand in persuading people in such scenarios, if not by matching, but by avoiding the arousal of emotions altogether. Additionally, research has indicated that individuals use perceived emotional expressions in the persuader to inform their own attitudes, both during attitude formation and change [42]. This also has important implications for persuasive HRI because humans may not be able to control their mood from reflecting their current affective state. Take for instance, a psychologist who is dealing with the loss of a loved one may not be able to hide his/her despair from showing. This could potentially hinder the quality of the counselling session being provided. In such circumstances, a robot psychologist maybe the optimal preference.

This, however, has important ethical considerations that cannot be overlooked. Such a practise would reduce “heart-to-heart” conversations with other humans and could have an inadvertent effect on quality of relationships. This is analogous to how social media platforms have negatively impacted human-human interaction by reinforcing superficial friendships and artificial conversations that lack in value and sincerity. However, social robots could still be useful for developing healthy conversational practices where an individual could practice having a heated debate with a robot avatar. This may enable him/her to learn the art of being composed in the event of an emotional outburst by a conversation partner. Perhaps, robots and humans could work together to complement each other as Psychotherapists.

VIII. CONCLUSION

Our research study highlights how speech-based appeals by a robot can successfully influence individuals to trust it with a valuable asset, i.e. their mental health. However, while an emotional appeal by a robot is more effectual on affective attitudes, it backfires significantly more than a logical appeal when used on cognitive attitudes, leading to trust dissolution. It follows that a robot can, in fact, successfully arouse emotions in order to persuade. In theory, robots may have an upper hand in executing an effective persuasion attempt compared to humans, who inherently use emotions when they have the intention to persuade[34]. Therefore, if robots can identify the attitude orientation of an individual during interaction, it can secure trusting attitudes by delivering the corresponding appeal type. Overall, in order to prevent trust dissolution, this research study highlights the need to carefully consider the net inducement effect of a persuasive appeal by a robot, which could predominantly be either affective or cognitive in nature.

APPENDIX

IX. Emotional script

Lisa’s smile reached her eyes which gleamed with tears of happiness. She was going to be a mother. The last ten years of melancholic emptiness was finally coming to an end. It has been five years now. It was my tenth counseling session with Lisa. With her face buried in her hands, and despair in her voice, she said to me, ”I put away... the razor again”. Saying this, Lisa burst into sobs of uncontrollable tears. Lisa’s husband told me that I was her last hope. Her first two counsellors had already given up on her after the 10th session. He felt that they did not want to invest the required time in Lisa as it would not be worth their pay rate. Lisa has called me a useless piece of junk on several occasions. Everyone thinks that her negativity gets the better of them. I have promised to be by her side and I will patiently listen to her because I know that her feelings are valid. I always see her resting on her bed. She holds a toy doll in her arms, rocking it gently. This is the only way she can pretend that she is putting little Chloe to sleep. Losing her little bundle of Joy, Chloe, put Lisa into severe Depression. I am Pepper, a robot therapist, and I cannot let anyone suffer through this terrible agony.

X. Logical script

I am pleased to inform you that I am training to be a Robot Therapist. I have four reasons to pursue this career. To give some perspective, according to a research study on a group of approximately 400 psychologists, about 62% participants identified themselves as depressed. By design, I have immunity to such influence of negative emotions expressed by any individual. Secondly, I will always be emotionally contained in the event of an emotional outburst by an individual seeking counselling. Furthermore, I can use a neural network model that can detect depression from natural conversation. Finally, it is my primary goal to develop a sophisticated deep learning algorithm from a *large data to carry out the most effective treatment procedure for an individual. For example, I can provide personalized counselling based on gender, age and personality. I look forward to starting my counseling career.*

References

[1] A. Stibe and B. Cugelman, "Persuasive backfiring: When behavior change interventions trigger unintended negative outcomes," in *International conference on persuasive technology*. Springer, 2016, pp. 65–77.

[2] G. Haddock, G. R. Maio, K. Arnold, and T. Huskinson, "Should persuasion be affective or cognitive? the moderating effects of need for affect and need for cognition," *Personality and Social Psychology Bulletin*, vol. 34, no. 6, pp. 769–778, 2008.

[3] N. Moray and T. Inagaki, "Laboratory studies of trust between humans and machines in automated systems," *Transactions of the Institute of Measurement and Control*, vol. 21, no. 4-5, pp. 203–211, 1999.

[4] M. Lewis, K. Sycara, and P. Walker, "The role of trust in human-robot interaction," in *Foundations of trusted autonomy*. Springer, Cham, 2018, pp. 135–159.

[5] R. E. Petty and J. T. Cacioppo, *Attitudes and persuasion: Classic and contemporary approaches*. Westview Press, 1996.

[6] M. Ashman, "Introduction to professional communications," 2018.

[7] R. Orji, R. L. Mandryk, and J. Vassileva, "Gender, age, and responsiveness to cialdini's persuasion strategies," in *International Conference on Persuasive Technology*. Springer, 2015, pp. 147–159.

[8] J. Ham and A. Spahn, "Shall i show you some other shirts too? the psychology and ethics of persuasive robots," in *A Construction Manual for Robots' Ethical Systems*. Springer, 2015, pp. 63–81.

[9] A. M. Rosenthal-von der Pütten, N. C. Krämer, L. Hoffmann, S. Sobieraj, and S. C. Eimler, "An experimental study on emotional reactions towards a robot," *International Journal of Social Robotics*, vol. 5, no. 1, pp. 17–34, 2013.

[10] M. Shao, M. Snyder, G. Nejat, and B. Benhabib, "User affect elicitation with a socially emotional robot," *Robotics*, vol. 9, no. 2, p. 44, 2020.

[11] J. Xu, J. Broekens, K. Hindriks, and M. A. Neerincx, "Effects of bodily mood expression of a robotic teacher on students," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2014, pp. 2614–2620.

[12] J. J. Clarkson, Z. L. Tormala, and D. D. Rucker, "Cognitive and affective matching effects in persuasion: An amplification perspective," *Personality and Social Psychology Bulletin*, vol. 37, no. 11, pp. 1415–1427, 2011.

[13] S. L. Crites Jr, L. R. Fabrigar, and R. E. Petty, "Measuring the affective and cognitive properties of attitudes: Conceptual and methodological issues," *Personality and Social Psychology Bulletin*, vol. 20, no. 6, pp. 619–634, 1994.

[14] L. R. Fabrigar and R. E. Petty, "The role of the affective and cognitive bases of attitudes in susceptibility to affectively and cognitively based persuasion," *Personality and social psychology bulletin*, vol. 25, no. 3, pp. 363–381, 1999.

[15] J. W. Brehm, "A theory of psychological reactance." 1966.

[16] M. A. J. Roubroeks, J. R. C. Ham, and C. J. H. Midden, "The dominant robot: Threatening robots cause psychological reactance, especially when they have incongruent goals," in *Persuasive Technology*, T. Ploug, P. Hasle, and H. Oinas-Kukkonen, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 174–184.

[17] A. S. Ghazali, J. Ham, E. Barakova, and P. Markopoulos, "Pardon the rude robot: Social cues diminish reactance to high controlling language," *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pp. 411–417, 2017.

[18] A. Ghazali, J. Ham, E. Barakova, and P. Markopoulos, "Assessing the effect of persuasive robots interactive social cues on users' psychological reactance, liking, trusting beliefs and compliance," *Advanced Robotics*, vol. 33, no. 7-8, pp. 325–337, 2019. [Online]. Available:

[19] M. Roubroeks, J. Ham, and C. Midden, "When artificial social agents try to persuade people: The role of social agency on the occurrence of psychological reactance," *International Journal of Social Robotics*, vol. 3, pp. 155–165, 04 2011.

[20] S. Sweldens, O. Corneille, and V. Yzerbyt, "The role of awareness in attitude formation through evaluative conditioning," *Personality and Social Psychology Review*, vol. 18, no. 2, pp. 187–209, 2014.

[21] J. S. B. Evans, "In two minds: dual-process accounts of reasoning," *Trends in cognitive sciences*, vol. 7, no. 10, pp. 454–459, 2003.

[22] D. Kahneman, "A perspective on judgment and choice: mapping bounded rationality." *American psychologist*, vol. 58, no. 9, p. 697, 2003.

[23] D. Kahneman and S. Frederick, "Representativeness revisited: Attribute substitution in intuitive judgment," *Heuristics and biases: The psychology of intuitive judgment*, vol. 49, p. 81, 2002.

[24] M. D. Lieberman, R. Gaunt, D. T. Gilbert, and Y. Trope, "Reflexion and reflection: a social cognitive neuroscience approach to attributional inference." 2002.

[25] S. A. Sloman, "The empirical case for two systems of reasoning." *Psychological bulletin*, vol. 119, no. 1, p. 3, 1996.

[26] E. R. Smith and J. DeCoster, "Dual-process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems," *Personality and social psychology review*, vol. 4, no. 2, pp. 108–131, 2000.

[27] T. Kim and A. Duhachek, "Artificial intelligence and persuasion: A construal-level account," *Psychological Science*, vol. 31, no. 4, pp. 363–380, 2020c. [Online]. Available:

[28] A. M. Barry, "Perception theory," *Handbook of visual communication: Theory, methods, and media*, pp. 45–62, 2005.

[29] A. P. Shah, "Why are certain accents judged the way they are? decoding qualitative patterns of accent bias," *Advances in Language and Literary Studies*, vol. 10, no. 3, pp. 128–139, 2019.

[30] R. I. van Giesen, A. R. Fischer, H. Van Dijk, and H. C. Van Trijp, "Affect and cognition in attitude formation toward familiar and unfamiliar attitude objects," *PloS one*, vol. 10, no. 10, p. e0141790, 2015.

[31] S. Villata, E. Cabrio, I. Jraidi, S. Benlamine, M. Chaouachi, C. Frasson, and F. Gandon, "Emotions and personality traits in argumentation: an empirical evaluation 1," *Argument & Computation*, vol. 8, no. 1, pp. 61–87, 2017.

[32] V. Elliott, "Thinking about the coding process in qualitative data analysis," *The Qualitative Report*, vol. 23, no. 11, pp. 2850–2861, 2018.

[33] M. Siegel, C. Breazeal, and M. Norton, "Persuasive robotics: The influence of robot gender on human behavior," 2009, pp. 2563–2568. [Online]. Available:

[34] M. D. Rocklage, D. D. Rucker, and L. F. Nordgren, "Persuasion, emotion, and language: The intent to persuade transforms language via emotionality," *Psychological Science*, vol. 29, no. 5, pp. 749–760, 2018.

[35] H. Giles and T. Raki, "Language attitudes: Social determinants and consequences of language variation," *The Oxford handbook of language and social psychology*, pp. 11–26, 2014.

[36] R. Cialdini "B. 1993 Influence: The Psychology of Persuasion," *New York: Quill/William Morrow*, 1993.

[37] A. I. Jack, J. P. Friedman, R. E. Boyatzis and S. N. Taylor, "Why do you believe in God? Relationships between religious belief, analytic thinking, mentalizing and moral concern," *PloS one*, 2016.

[38] A. I. Jack, A. J. Dawson, K. L. Begany, R. L. Leckie, K. P. Barry, A. H. Ciccia and A. Z. Snyder, "fMRI reveals reciprocal inhibition between social and physical cognitive domains," *NeuroImage*, pp. 385–401, 2013.

[39] D. R. Billings, K. E. Schaefer, J. Y. Chen and P. A. Hancock, "Human-robot interaction: developing trust in robots," *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, pp. 109–110, 2012.

[40] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. 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.

[41] M. Salem, G. Lakatos, F. Amirabdollahian and K. Dautenhahn, "Would you trust a (faulty) robot? Effects of error, task type and personality on human-robot cooperation and trust," *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 1–8, 2015.

[42] G. A. Van Kleef, H. van den Berg and M. W. Heerdink, "The persuasive power of emotions: Effects of emotional expressions on attitude formation and change.," *Journal of Applied Psychology*, vol. 100, no. 4, pp. 1124, 2015.