Mitigating Boredom Using An Empathetic Conversational Agent

Samiha Samrose
University of Rochester
ssamrose@cs.rochester.edu

Kavya Anbarasu
Sharon High School
kavya.anbarasu@gmail.com

Ajjen Joshi
Affectiva
ajjen.joshi@affectiva.com

Taniya Mishra
Affectiva
taniya.mishra@affectiva.com

ABSTRACT

In spite of their ubiquity, our interactions with contemporary conversational agents (CA), such as Alexa, are still transactional in nature and lack the expressiveness of human-human communication. Conversational agents equipped with the ability to detect and address users' emotional and cognitive states could make our interactions with them more human. In this work, we investigate whether an empathetic CA can help mitigate boredom. We design a protocol in order to first elicit boredom in users, and explore strategies that attempt to mitigate their boredom with the help of two conversational agents, an empathetic agent and a non-empathetic agent, administered in a Wizard-of-Oz setting. We quantify their efficacy by measuring the effects on user mood and task performance. Our user study with 34 participants shows that the empathetic CA outperforms the non-empathetic CA with respect to modulating users' mood and performance.

CCS CONCEPTS

• **Human-centered computing** → **User studies**; **Natural language interfaces**.

KEYWORDS

Conversational Agent; Empathetic Agent; Boredom Mitigation

ACM Reference Format:
Samiha Samrose, Kavya Anbarasu, Ajjen Joshi, and Taniya Mishra. 2020. Mitigating Boredom Using An Empathetic Conversational Agent. In *IVA '20: Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents (IVA '20), October 19–23, 2020, Virtual Event, Scotland Uk*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3383652.3423905

1 INTRODUCTION

Conversational agents (CA), such as Amazon's Alexa and Apple's Siri, are becoming increasingly capable of fulfilling a complex variety of tasks, such as searching for information, controlling media, or even playing games. However, CA-human interaction still lacks the ease and expressiveness of human-human communication. In spite of recent technical advances, human interactions with CAs continue to be challenging, frustrating, and consequently distracting.

An essential component of human-human interactions is the ability of people to detect and address the emotional and cognitive states of the person they are interacting with. The ability of a CA to address emotional and cognitive states of the user can be beneficial in numerous applications, such as customer service, healthcare, etc. In this work, we envisage an in-car CA that detects and proactively mitigates a negative emotion - boredom. Boredom can be particularly dangerous in the context of driving. For example, boredom can lead to mind-wandering [40], fatigue [42] and drowsiness [2], and thus increase the risk of serious accidents. This risk can potentially be alleviated by an in-vehicle conversational agent that first detects drivers' boredom and mitigates it by engaging in a conversation with them.

In this paper, we answer two primary research questions. First, we investigate whether CAs can actually help mitigate user boredom. In order to answer this question, we design and implement a protocol that first elicits boredom in users. The protocol requires participants to watch a monotonous driving video on a screen and annotate road signs that appear in the video. After a certain amount of time, the CA administered in a Wizard-of-Oz setting provides an intervention by offering to play a game of "20 questions". Using self-reported surveys, automatically captured task logs, and subjective interviews with the participants, we measure how the CA affects the user's mood as well as task performance.

Second, we explore whether an empathetic CA can better resolve user boredom compared to a traditional, non-empathetic CA designed to mimic the abilities of currently available voice agents. We designed the "personality" of the empathetic CA such that it appears to detect and respond to certain emotional and cognitive states of the user, with the hypothesis that incorporating an agent with these qualities will encourage the user to engage more with the agent. Each participant in our user study interacts with both CAs, allowing us to compare their performance and efficacy in a within-subject study.

In summary, we (1) developed a protocol to elicit boredom and collected audio-video and self-reported data from 34 participants in order to capture the audio-visual signals of boredom; (2) designed demonstrations of a traditional, non-empathetic CA (which mimics the "transactional" interaction characteristics of currently available CAs) and an empathetic CA (which is capable of both understanding and expressing emotions) and applied them through a Wizard-of-Oz methodology to investigate our hypothesis that an empathetic CA can better mitigate boredom than a traditional CA while users perform monotonous tasks.

2 RELATED WORK

2.1 Conversational Agents & Empathy

Research in developing conversational agents that can interact with humans has a long history [44]. The proliferation of mobile phones and devices such as Amazon Echo and Google Home has increased a lot of interest on it within the HCI community [36]. While CAs have become a key mode of human-computer interaction, the difference between user expectation and actual interaction experience

IVA '20, October 19–23, 2020, Virtual Event, Scotland Uk

ACM ISBN 978-1-4503-7586-3/20/09…$15.00
https://doi.org/10.1145/3383652.3423905

is still quite large [32]. Current usage of CAs are limited to completing constrained tasks, such as checking the weather, sending or reading messages, playing music, and controlling Internet of Things (IoT) devices. Users' interactions with CAs are conceptualized in transactional terms, making it difficult for users to form emotional connections [8].

It has been argued that understanding users' affective experience is crucial in improving upon the current state of CAs [46]. In order to establish stronger emotional connections between a user and an agent, their interactions with people must contain social mechanisms, that people employ in their interactions with each other [30]. The mechanisms to make interactions more natural may include the capacity to engage in small talk [4], the ability to be humorous [8], and respond to users' affective states [46]. In order for humans to develop meaningful relationships with artificial agents, the agent may need to possess "empathy" [12, 34].

2.2 Boredom: Elicitation & Mitigation

An in-car conversational companion is an interesting use-case for an empathetic CA. In the Unites States, people spend an estimated average of 17600 minutes per year in a car [3], equivalent to more than 12 24-hour days. A significant portion of that time may be spent driving alone in monotonous, non-stimulating environments, increasing the chances of the driver reaching a state of boredom and cognitive underload. A CA that is able to detect the emotional state of the driver can help improve driving performance [23].

Fisher defined boredom as "an unpleasant, transient affective state in which the individual feels a pervasive lack of interest and difficulty concentrating on the current activity" [15]. Boredom has been shown to cause mind-wandering [40], increase fatigue [42] and lead to distraction [37], all of which are dangerous in the context of driving [11]. Steinberger et al. discussed technology interventions, such as performance feedback, increased challenge, and gamification, to increase task engagement and therefore offer safety benefits [41]. Similarly, an empathetic CA that engages with the driver when it detects a heightened state of boredom could provide a potentially life-saving service.

Advances in computer vision, speech and signal processing as well as machine learning have accelerated the development of automated emotion recognition from facial [31], vocal [43], physiological [47], or multimodal signals [24]. While most of the work has been done to identify basic prototypic emotions, there have been some work in building automated systems to detect more complex emotional and cognitive states. For example, researchers in the education domain have attempted to build automatic models to detect student engagement, a phenomenon inversely related to boredom, from face [6, 45] and speech [21]. Researchers have also attempted to model mind-wandering, another phenomenon related to boredom, using eye-gaze behavior [20, 29] and physiological signals [5]. However, computational models of boredom and approaches to mitigate it, especially in the automotive context, is under-researched.

Modern, data-driven models of boredom require datasets of subjects in varying states of "boredom", thereby engendering the need to build protocols capable of inducing boredom. Also, in order to experiment with strategies that attempt to mitigate boredom, boredom must first be elicited in participants. Markey et al. [33] tested and compared the effectiveness of various boredom induction methods. Boredom can be induced from the lack of a desired level of stimulation. Stimulation in users can be modulated based on the characteristic of the ongoing task, the personality of the task performer, or both [14]. We utilized these findings to inform the design of our boredom elicitation and mitigation protocol.

3 PROTOCOL TO ELICIT & MITIGATE BOREDOM

We designed a protocol with two phases that first elicits and then attempts to mitigate a user's boredom: first, participants perform a task that is conducive to boredom elicitation; second, a conversational agent attempts to mitigate their boredom by engaging the participants in a conversation while they continue to perform the boredom-inducing task. We designed two conversational agents: one that possesses empathetic qualities and one that doesn't, to determine which one is more effective in mitigating boredom.

3.1 Boredom Inducing Task Design

Guest et al. [18] illustrated three characteristics of boredom inducing tasks: (a) repetitive [25, 26], (b) long-winded [28], and (c) unchangeable but predictable [9]. A solo long-drive on an empty highway at night possesses all three characteristics. The driver cannot do anything else with his/her hands besides steering the car (unchangeable state due to inability to seek other stimuli). The driver also has to keep his/her eyes on road-signs, which are infrequent on highways (long-winded and non-happening), and continue checking the road-signs as other visual stimuli may not be visible during night-time (repetitive and non-happening).

Based on these criteria, we designed a similar task in order to induce boredom in task performers. The task requires a user to watch a first-person driving simulation video on a computer screen and annotate the road-signs, which appear infrequently in the video. Watching such videos for a long time with minimal stimuli has proven to have negative impact on focus, attention, and mood [7, 17]. In the video, which ran for approximately 21 minutes, the car moves along the highway at night, at a slow speed with very few other cars. Even though long drives may last more than 21 minutes, Geden et al. [16] shows drivers' minds start wandering after the first few minutes of starting to drive, giving support to our experimental design. Participants were instructed to annotate whether a road-sign appears to the left or to the right side of the road. As during actual driving where the driver's hands remain pre-occupied, the task performer was asked to press keys on the keyboard for annotation. The keys were selected in a way so that s/he was primed to use both hands: 'z' for 'left' and 'm' for 'right', which are far apart on a regular QWERTY keyboard.

3.2 Conversational Agent Design

The next stage of our protocol involves the application of a boredom-mitigating intervention administered via conversational agents. Modern AI-enabled cars have started introducing in-car CAs to assist drivers with a variety of tasks. Due to the challenges of running experiments while participants are actually driving, we use our task as a proxy for real-world driving scenarios. Because we wish to imitate interactions with an intelligent agent inside a vehicle, a conversational agent is best-fitted (as opposed to chatbots or avatar-based agents).

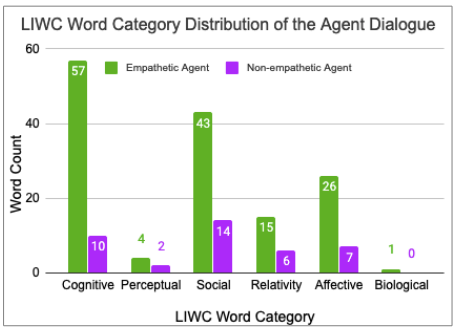


Figure 1: Distribution of words in the various LIWC Categories in Agent Dialogue Formation for the two CAs.

As existing CAs do not possess the ability to detect boredom from a user's facial or vocal expressions and respond accordingly, we applied a Wizard-of-Oz (WoZ) approach to operate the agents [10]. In our protocol, operating CAs with a WoZ approach involved two challenges. First, in designing an empathetic CA to have a distinct conversational style, different from a traditional non-empathetic CA, the dialogues generated by the WoZ setup needed to be fully structured. Second, substantial delay in delivering the agent's dialogue may exacerbate the negative mental state of the task performer. Therefore, the agent's dialogues were mostly pre-recorded instead of improvising the agent's responses ad-hoc.

3.2.1 Dialogue Topic. The empathetic agent we designed, named *Emma*, attempts to mitigate boredom by engaging the user in a conversation. In order to compare the effects of *Emma* against a non-empathetic agent like Alexa, the non-empathetic agent must also engage in a conversation, as opposed to, say, merely play music. Thus, we designed the dialogues of a non-empathetic CA, named *Nina*, to mimic the conversational style of currently available transactional CAs. Both CAs inform the user that they are equipped with the ability to detect the user's affective state.

In order for the user to interact and engage with each of the CAs, we picked the game '20 Questions'[1]. In 20 questions, one player thinks about a person, place or object, which the other player has to guess by asking up to 20 questions, each of which can only be answered with a "Yes" or "No". Currently, Alexa is capable of playing 20 questions, but only as the "guesser". In our protocol however, letting the participant think about an item for the CA to guess would make the experimental setting too open-ended for the CA (through the WoZ) to respond appropriately and in a timely manner. Instead, we reversed the roles and asked the user to guess a person which the CA has in mind. This opens up more opportunities for the user to engage (as opposed to answering with just a "Yes" or "No") and also helps constrain the experimental setup.

3.2.2 Dialogue Property. The dialogue structure of present-day CAs is mostly transactional [38, 39]. That is, they only provide a response when a user asserts a query. CAs today do not possess any empathetic qualities. For example, they are unable to initiate an assistive interaction when a person is in a negative mental state. We attempt to imbue empathy in *Emma* through careful dialogue formation, which we describe below:

(1) Word Level: In order for *Emma* to appear to be cognizant of emotion and express sufficient empathy, her dialogue incorporates person-based (e.g., 'I', 'you'), emotion-related (e.g., 'feel', 'happy'), and cognitive (e.g., 'believe', 'think') words. We followed Linguistic Inquiry and Word Count (LIWC) [35] categorization to include specific words in the dialogues of the two agents. As shown in Fig 1, the empathetic agent engages more with the participant and therefore has more wordcount and uses significantly more words from *cognitive, social and affective* LIWC categories.

(2) Sentence Level: Emma's dialogue was designed to support empathy-enabled rapport building, which can be established through compassion (e.g., '*I can sense that you are feeling bored. I can understand because I also feel bored counting numbers all day. Is your task similarly tedious?*'), small-talk (e.g., '*How is your task going?*'), self-disclosure (e.g., '*I like long drives. I was in a car that drove to Disneyland, it was fun.*'), rolling conversation through questions (e.g., '*Just wondering, what was making your task tedious?*'), assistance (e.g., '*I'm here if you need me for anything*'), and appreciation (e.g., '*You won! You are really good at this game!*'). To construct self-disclosure, the agent was provided with a general background, covering favorite place, food, color, etc. (e.g., '*My favorite color is blue*'). To handle any unforeseen queries from the user, generic responses (e.g., '*I feel the same way*', '*I have never thought about it*') were also stored.

(3) Dialogue Level: To form congruent flow between subsequent sentences, *glue sentences* were incorporated (e.g., '*Oh, I see*', '*Got it*'). Steering properties (such as, asking questions about specific domains) were included to guide the conversation in the desired direction (e.g., '*Interesting! When I am bored, I try to find a game to play! We can play 20 questions, if you want. Do you know how to play?*'). The dialogues for *Nina*, on the other hand, were designed to mimic the transactional property of currently available CAs (e.g., *'Yes', 'No', 'Thank you', 'Sorry, I do not understand'*, etc.). After testing different voices, we purchased a premium version of the speech-to-text reader developed by NaturalSoft[2], and chose the 'English(US) - Jennifer' voice at speed:1 for recording the dialogues.

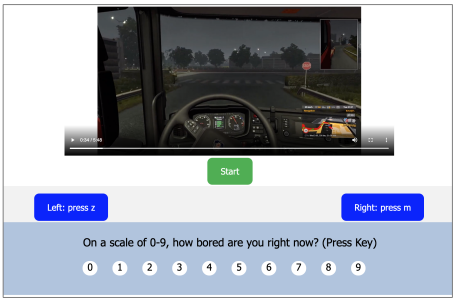
3.3 Platform Design

To implement the protocol in a user study, we built two interfaces: (1) a task platform for participants, and (2) a Wizard-of-Oz platform for the experiment administrator. The task platform (Fig 2(a)) lets a user perform the annotation task, records audio-video feeds, and logs all button interactions. It contains (1) a video player that starts once the button 'Start' is pressed; (2) two signs 'Left: press z' and 'Right: press m' which can only be interacted by pressing the corresponding keyboard keys; and (3) a survey, that pops up every 3 minutes asking the annotator's current boredom level.

The system back-end logs the time when the audio-video capture started, which keys are pressed and when. The system also records the webcam and the microphone feeds. Once the video ends, the system automatically opens up a new tab for the participants to fill out a survey. As the dialogues of our CAs are pre-recorded, we built an interface shown in Fig 2(b). Pressing a dialogue button plays that particular sound clip and records the time when that particular button is clicked.

[1] https://en.wikipedia.org/wiki/Twenty_Questions

[2] naturalreaders.com



(a) Task UI

(b) WoZ UI

Figure 2: Task UI for participants and WoZ UI for experiment administration.

4 USER STUDY

We conducted an experiment to compare the effectiveness of *Emma* and *Nina*, in modulating the participant's mood and task performance. To assess which agent a person would choose in the third session after experiencing both, we adopted a within-subject design. Between-subject design could have given users' scores for agents to compare, but users would not have been able to choose a final agent which is a direct preference. Therefore, to get the same user's direct comparative preference, within-subject design was adopted for this user study.

4.1 Participants

Participants were recruited by circulating an open email. A total of 39 participants attended the user study. Data of 5 participants were excluded because of technical and environmental errors during the study. In our study, 62% of participants were female and 38% were male. The age distribution for age ranges 19-29, 30-39, 40-49, 50-59, 60-69, 70-79 were 26%, 12%, 20%, 12%, 24%, and 6%, respectively. As boredom can be related to people's personality, we collected their likelihood of being bored by having them fill a survey measuring the Boredom Proneness Scale [13]. The BPI score range was 78-138. Each participant was provided with a $75 honorarium in exchange for their participation.

4.2 Study Flow

The user study took place in an in-lab setting. Upon arrival, participants were briefed that the general purpose of the study was a "conversational agent interaction study" with no reference to boredom so as not to precondition them to expect that emotion.

First, participants were seated, assigned to a computer and a headphone, and asked to fill out a survey which assessed their

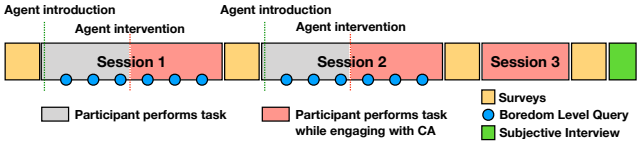


Figure 3: Schema of our protocol that illustrates the flow of sessions and when agents act within each session.

emotional state right before starting the experiment. Next, the experiment administrator explained the annotation task and the task interface. The actual experiment consisted of three sessions. The first session began by the agent introducing itself to the participant, informing its ability to detect emotion and expression, and asking for permission to enable the feature. This phase of calibration was designed to provide an autonomous vibe and minimize the suspicion about the actual WoZ mechanism. Following this, the annotation task was started. That is, the 21-minute first-person driving simulation video started playing on the computer interface, and the participant labeled which side of the road a road-sign appeared. Every 3-minutes, an automatic survey prompt appeared at the bottom of the screen, asking the current boredom level on a scale of 0 (not bored) to 9 (very bored).

The first 10 minutes of the annotation task took place without any agent intervention. At the 10th minute, the agent initiated a conversation with the user by first saying it detected that the participant was bored and then offering to play the game '20 Questions'. After the driving video ended, a survey window automatically popped up, asking the participant several questions about his/her perception of the task and the agent. After the participant completed the survey, the second session was started. The second session was identical to the first, except that the agent was switched. The ordering of the agents in the two sessions was counterbalanced, so that half of all participants engaged first with *Emma* and the other half with *Nina*.

Finally after the completion of the two sessions, participants were asked to choose one of the two for the final annotation session spanning for 6 minutes. The purpose of including a third session was to determine which agent the participants preferred, given their experiences in the first two sessions. A final survey was provided, asking the participant for justifications behind choosing the particular agent. In total, the experiment lasted about 45 minutes.

Once the sessions began, the participant remained alone in the user study room, so that they felt more comfortable in authentically expressing their emotions. The study administrator operated from a different room, monitoring the participant's interactions with the agent and responding to their questions using the WoZ interface.

5 RESULTS

Before verifying our research hypotheses, we first validated the assumptions set in the protocol. Once the assumptions were validated, we analyzed the outcome of our empirical exploration.

5.1 Protocol Verification

Our protocol bears the assumption the formulated task can successfully elicits boredom. The surveys after the first two annotation sessions included questions from the NASA Task Load Index questionnaires [19]. We administered this in order to measure the cognitive load of the participant associated with the performed task. The NASA Task Load Index calculates how demanding a task is on

its provided scale. Participants provided their responses in a 5-scale likert scale which were then converted to a 0-100 point scale as per the standard conversion technique[3]. Final scores for the two task sessions were 29.6 and 27.5.

We conducted a *two-tailed paired student t-test* on the responses across each question. The result shows no significance difference between the two tasks, which implies that characteristically the tasks had similar effect. On a 0-100 point scale, the relatively low scores ($val = 29.6, 27.5$) for the two tasks imply that both tasks were not very mentally demanding and therefore, had low cognitive load. The relative closeness of the two scores ($diff = 2.1$) concludes that both induced similar levels of mental demand, and hence, are comparable tasks. In addition to this, intermittent prompts during the experiment captured the participants' boredom level at a regular 3 minute interval. From the individual curves shown in , it is evident that participants experienced a heightened level of boredom during the labeling task. During the first 10 minutes, the average peak boredom reached 5.03 on a scale of 0 (not bored) to 9 (very bored), which implies that on average the protocol successfully induced moderate boredom within participants within just 10 minutes. Alongside, the trend in Fig 4 shows a non-zero strictly increasing pattern during the boredom elicitation period confirming. Therefore, we established that the protocol was successful in eliciting boredom among participants in our study.

5.2 Agent Effect Evaluation

Having established that the protocol was indeed successful in eliciting boredom in participants, we now explore our primary research questions related to boredom mitigation. We evaluated the agents based on two parameters: (1) Effect on the participant's mood: the target was to mitigate or improve the negative emotional state in a way that is noticed and appreciated by the participants, (2) Effect on the participant's task performance: as the agent engages with the participant during an ongoing task, it is crucial that the participant's focus on the task is not drastically altered by the interaction.

5.2.1 Effect on User's Mood. To observe this effect, we collected both intermittent and the overall boredom levels of the participants during the course of the experiment. During the annotation task, the participants provided self-reports on their perceived boredom level at every 3-minute intervals. In each survey administered at the end of a session, the participants reported a combined score for their boredom level.

Temporal Modulation of Boredom: Fig 4 illustrates the temporal patterns of all participants' boredom levels during an experimental session. In a 21-minute video labeling session, the interface prompts the user to report a boredom level every 3 minutes, except at the end of the video. The prompt asks "*On a scale of 0-9, how bored are you right now?*", and the user presses the corresponding number key. Therefore, we consider a total of 6 data points in each session for each participant.

Fig 4 shows the self-reported boredom scores across all participants revealing how the reported level of boredom evolved over the course of a session. The average patterns for sessions having empathetic and non-empathetic traditional agents are shown using green and purple curves, respectively. We apply a *one-tailed paired t-test*

[3]https://measuringu.com/nasa-tlx

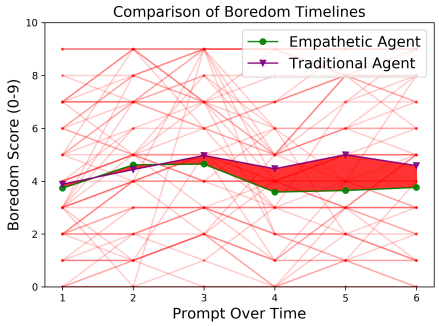


Figure 4: Participants' boredom modulation as illustrated by plotting their self-reported boredom scores for all sessions.

on boredom scores between with and without agent parts of the session. For empathetic agent, there was a statistically significant difference ($M = -0.39, SD = 4.01$) with $t = -1.82, p = 0.036 < 0.05$. It confirms that in comparison with no-agent portion, boredom level significantly decreased after the empathetic agent was introduced in the session. For non-empathetic agent, there was no significant difference between no-agent vs agent portions.

After the 3rd prompt at around the 10 minute mark, the agent initiates interaction with the participant. After that, even though the average self-reported boredom initially decreases in both sessions, the magnitude of the decrease is more prominent during the presence of the empathetic agent. *Emma* decreased boredom in users by a range of 21.87% to 37.10% in comparison with that of *Nina*. To measure whether reduced boredom with the presence of *Emma* is significantly different than that of *Nina*, we conducted a *one-tailed paired t-test* which shows statistically significant difference ($M = 0.79, SD = 9.1$) with $t = 2.45, p = 0.008 < 0.05$. This confirms that the empathetic agent significantly reduced boredom in comparison with the non-empathetic agent.

Overall Boredom Score for Each Session: After each session, the participants answered survey questions specific to agent (Q1:"Overall, how was the response of the agent", Q2:"Overall, what was the characteristic of the agent", Q3:"Overall, how was your interaction with the agent"), self-performance (Q4:"How do you think you performed in the labeling task", Q5:"Overall, how would you rate the labeling task"), boredom (Q6:"How would you rate labeling task when the agent was present", Q7:"How would you rate the labeling task when the agent was present"), and agent impact (Q8:"How did the agent influence the task experience", Q9:"How did the agent influence your mood", Q10:"How did the agent influence your attention level to the task"). A *paired samples t-test* on the

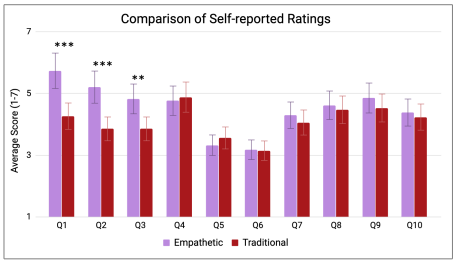


Figure 5: Mean response values of participants' ratings of the two agents, based on filling survey questions.

self-reports shows that Q1, Q2, Q3 were statistically significant with $p - val < 0.001, < 0.001, 0.00156 < 0.01$, respectively. This reveals that during a boredom inducing task, the empathetic agent's response, characteristic, and interaction are more preferable in comparison with that of the non-empathetic agent. Fig 5 shows that participants rated the overall session on a 7-point boredom scale (Q5 range: 1-very boring, 7-very exciting. $full\_emma_{mean}$ = 3.32, $full\_nina_{mean}$ = 3.56). They also reported the first halves of the annotation sessions without any interaction with the agents to be of similar level of boredom (Q6: $firsthalf\_emma_{mean}$ = 3.18 and $firsthalf\_nina_{mean}$ = 3.15). In the second halves of the sessions, the task was reported to be less boring when participants interacted with *Emma* in comparison with that of *Nina* (Q7: $secondhalf\_emma_{mean}$ = 4.29,$secondhalf\_nina_{mean}$ = 3.56).

Automated Affect Analysis: Parallel to analyzing self-reported affect, we applied automated affect analysis on the recorded study videos. We processed videos of participants with the Affdex SDK [1] in order to automatically capture the facial expressions and emotion signals. Fig 6 shows the average affective responses over the duration of a session. Fig 6(a) shows that the overall valence remained negative throughout each session, which supports the previous self-reported finding that overall the task was boredom-inducing. However, the higher spikes in the empathetic agent session indicate that with *Emma* the participants experienced moments of "relief" bringing their valence to less negative state. *Joy* feature (Fig 6(b)) supports this, showing more smile markers with higher magnitude during the session with *Emma*. Fig 6(c) reveals that the participants indeed engaged with the empathetic agent more. *Attention* analysis in Fig 6(d) shows that the attention variation did not change before and after agent interventions.

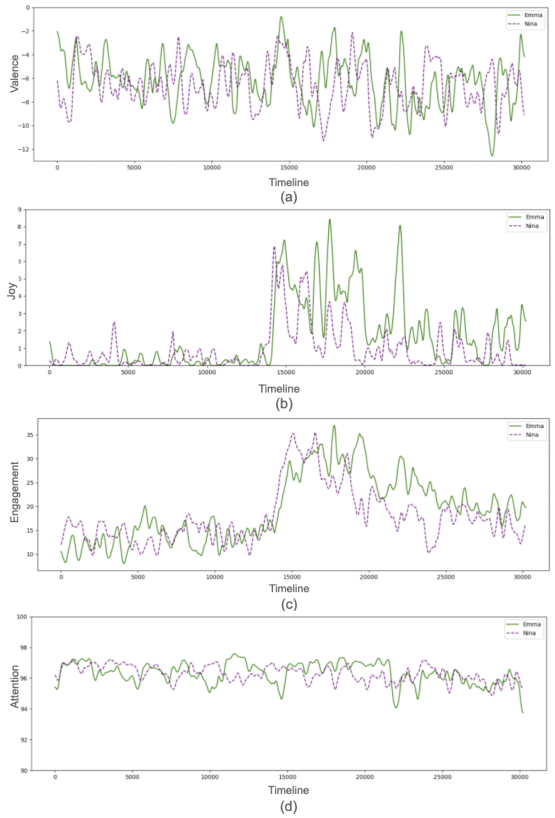


Figure 6: Session Comparison of Participants' Facial Affect.

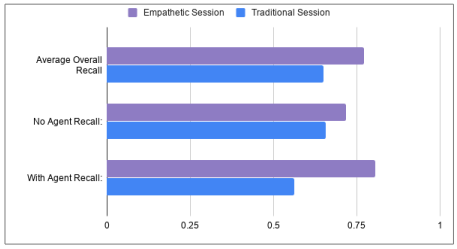


Figure 7: Average recall values for task accuracy in empathetic agent session versus non-empathetic agent session.

5.2.2 Effect on User's Task Performance. To analyze the task performance of the participants, we measure the task accuracy across sessions (number of road signs correctly annotated). Accuracy, precision and recall were all measured to be higher in the session with the empathetic agent. For empathetic and traditional agent sessions, accuracy was 35%, 26%; precision was 38%, 32%; recall was 80%, 56%; respectively. In both sessions accuracy and precision were low - the reason for which was revealed through further investigation. We found out that there was a high volume of false positive annotations, due to the participants misinterpreting what traffic signs to mark. Even though the labeling instruction asked to annotate road signs with direction or speed limit on them, participants interpreted colorful road-side poles to be parts of road signs. This caused the accuracy and precision measures to be skewed. Recall here is more insightful as it is representative of the target signs being correctly annotated. We therefore report results using recall (shown in Fig 7). The results show that during the time when the participant was not interacting with the agent, recall was similar between sessions. However, recall increased significantly (over 20%), when the participant was interacting with the empathetic agent as opposed to the non-empathetic one.

In detail, averaged over all participants, precision scores for empathetic and traditional agent sessions were 0.50 and 0.54, respectively; and the difference was not statistically significant ($p - val$ = 0.22). Recall scores for the same were 0.76 & 0.69, respectively with no statistically significant difference ($p - val$ = 0.056).

5.2.3 Agent Preference. After the second annotation session, participants were asked to report their preferred agent. The responses in Fig. 8(a) show that 73% of the participants preferred *Emma*, 12% *Nina*, 6% both agents, and 9% did not prefer any of the agents. At the beginning of the third session, participants were made to choose between *Emma* and *Nina*. Fig 8(b) illustrates the preference of participants when forced to make a choice. Understanding the reasons behind their preference is important because they reveal important information about the CA's usability. From survey and subjective interviews, we adopt a case-study based approach to understand the following phenomenons:

Case Study 1: Preference for Emma/Nina: Clearly most participants expressed preference for *Emma* over *Nina*, as illustrated by the results in the survey as well as their choice of agent for the third session. Here we include some comments made by participants in the survey and during the subjective interview:

> P28: "*Emma seems like a cute, harmless, transparent kind of voice; obviously intended to be friendly, and made me feel like I'm in the presence of someone/something more or less 'safe.' Agent Nina gave me a 1984 vibe -*

> *kind of opaque, very 'tech,' as if it wasn't designed to sound human at all... I would purchase Agent Emma for my car if it wasn't too pricey honestly*"

P7, who preferred *Emma*, explained the characteristics of the two agents from her perspective-

> P7: "*I found Nina was very straightforward - A, B, and C, and that would be enough. And then Emma, I felt like she cared about me.*"

Participants found both agents as stimuli to reduce their boredom. However, they found their interactions with *Emma* to be was more pleasant because she gave a sense of support and empathy. Few participants who preferred *Nina* expressed the reason to be its familiar and concise nature.

Case Study 2: When participants chose neither agent: As reported in the survey, 3 participants preferred none of the agents. They wrote down their reasoning behind this decision:

> P5: *"The agent (Nina) is distracting and annoying"*
> P20: *"I don't like talking to things like Siri or Alexa"*

During the post experiment interview session, P20 justified her answer by elaborating:

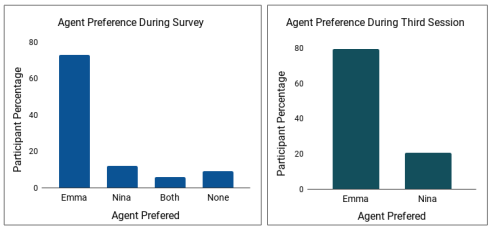
> P20: *"I said (to the agent) 'Wish I had some music', and she said 'That's not available', and I'm like 'Dude!'. It's like the lady in the grocery store in the machine that says there's nothing in the bagging area."*

This highlights the importance of improving performance range and capacity of any conversational agent. In our setup, the constrained use case did not allow for additional CA skills, such as playing music. But from the user's perspective, any agent not able to fulfill the desired assistance may cause the user to lose interest.

Case Study 3: When participant indicated preference for one agent but chose another for the 3rd session: One participant indicated preference for *Emma* in the survey but chose *Nina* as her CA in the third session. During the subjective interview, she revealed that the choice was made based on how she was brought up to behave in the car:

> P10: "*I never play 20 questions in the car, not even when I was little. We were told to sit and shut up in the car. That's how my parents were, I'm old school. It was engraved in me.*"

This sheds light on the importance of understanding a user's background and overall mental state to fully comprehend what type of CA would be effective for that particular user. Personalization can enable the same CA to be empathetic towards different users in different manners.



(a) Free Preference (b) Enforced Preference

Figure 8: Participant's preferred agent when a) given options to choose both or none, and b) forced to choose one agent.

6 DISCUSSION

6.1 CA in Boredom Modulation

Even with the presence of an agent, the boredom level does not become — nor can it be expected to become — zero as the boredom inducing annotation task stays co-present and thus, continues to impose boredom elicitation properties. Our goal is therefore to mitigate boredom, as complete elimination may not be possible.

The tightly-structured nature of the dialogue used by the traditional agent compared to the less structured, conversational dialogue of the empathetic agent poses the difference in interactions. Our empathetic agent has more speech length than the traditional agent, which can be a confound. However, narrative empathy [27] especially depends on timing, consistency, and context going beyond just the narration length. In our study, the participants had the option to not engage with the agents at all. However, with *Emma* they chose to continue conversing with it, unlike how they engaged with *Nina*. Once the game of 20 questions finishes, even though re-starting games is an option the participants do not initiate any further conversations with the traditional agent and thus boredom increases, whereas they continue conversing with the empathetic agent across multiple turns and the boredom lowers down.

6.2 Effect of Boredom Mitigation on Task

The computed NASA task load index for the Empathetic Session was 0.426 while that of the Traditional Session received a score of 0.401. The relative closeness of the two scores indicates that tasks performed in both sessions induced comparable levels of mental demand. This was further bolstered by participants' answer to *Q5: "Overall, how would you rate the labeling task?"* in Fig 5, indicating that the tasks were similarly boring in sessions with both the traditional and empathetic agents. The participants perceived that interacting with the agents positively influenced attention level to the task and task experience, we expected that task experience would be positively influenced by boredom mitigation. However, Figure 7 shows that this only bore out in the case of the empathetic session, where task recall improved when participants began interacting with the agent compared to the no agent session. In contrast, for the non-empathetic standard agent session, task recall became worse when participants began interacting with the agent compared to the no agent session.

6.3 Recommendations for In-car Agent Design

The two main findings of this work — that an empathetic agent can substantially mitigate boredom *and* can positively impact task performance — provide strong support to the idea of developing empathetic in-car agents. Such an agent could reduce distraction and mind-wandering that result from boredom, and thus alleviate potentially dangerous driving conditions. It also is clear from our findings that these benefits will not be realized by standard agents because while they did temporarily reduce boredom — though to a smaller extent compared to empathetic agent, they negatively impacted task performance.

6.4 Future Work

Our user study enabled us to capture a rich data set of interactions with frequently self-reported boredom annotations. This provides an opportunity to build multimodal machine learning boredom prediction models based on facial and vocal signals of participants.

In future experiments, we can use the predictions of the boredom model to trigger agent interactions with the participants. As culturally aligned agents can be more relatable and effective [22], designing agents incorporated with personalized features would be another interesting research direction to pursue in the future.

7 CONCLUSION

Conversational agents equipped with the ability to detect and address users' emotional and cognitive states could make our interactions with them more humane. In this work, we designed a protocol to elicit boredom in users and explored strategies attempting to mitigate boredom with the help of an empathetic CA and a non-empathetic CA, administered in a Wizard-of-Oz setting. We showed that the empathetic CA outperformed the non-empathetic CA with respect to modulating users' mood and performance. Based on the results of our study, we made recommendations for the design of boredom-mitigating, in-car conversational agents.

REFERENCES

[1] 2016. Affectiva. http://www.affectiva.com/. Accessed: 2016-10-30.

[2] Clare Anderson and James A Horne. 2006. Sleepiness enhances distraction during a monotonous task. *Sleep* (2006).

[3] American Automobile Association. 2016. *Americans Spend an Average of 17,600 Minutes Driving Each Year.* https://newsroom.aaa.com/2016/09/americans-spend-average-17600-minutes-driving-year/

[4] Timothy Bickmore and Justine Cassell. [n.d.]. Small talk and conversational storytelling in embodied conversational interface agents. In *AAAI fall symposium on narrative intelligence.*

[5] Nathaniel Blanchard, Robert Bixler, Tera Joyce, and Sidney D'Mello. 2014. Automated physiological-based detection of mind wandering during learning. In *International Conference on Intelligent Tutoring Systems.*

[6] Nigel Bosch, Sidney K D'Mello, Ryan S Baker, Jaclyn Ocumpaugh, Valerie Shute, Matthew Ventura, Lubin Wang, and Weinan Zhao. 2016. Detecting student emotions in computer-enabled classrooms. In *IJCAI.*

[7] Chih-Ming Chen and Chung-Hsin Wu. 2015. Effects of different video lecture types on sustained attention, emotion, cognitive load, and learning performance. *Computers & Education* (2015).

[8] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, Vincent Wade, and Benjamin R. Cowan. 2019. What Makes a Good Conversation?: Challenges in Designing Truly Conversational Agents. In *Proceedings of CHI.*

[9] David Cox. 1970. Organization of repetitive tasks: Some shop floor experiments recalled. *Occupational Psychology* (1970).

[10] Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg. 1993. Wizard of Oz studies why and how. *Knowledge-based systems* (1993).

[11] Eric R Dahlen, Ryan C Martin, Katie Ragan, and Myndi M Kuhlman. 2005. Driving anger, sensation seeking, impulsiveness, and boredom proneness in the prediction of unsafe driving. *Accident Analysis & Prevention* (2005).

[12] Fiorella De Rosis, Addolorata Cavalluzzi, Irene Mazzotta, and Nicole Novielli. 2005. Can embodied conversational agents induce empathy in users? *Virtual Social Agents* (2005).

[13] Richard Farmer and Norman D Sundberg. 1986. Boredom proneness–the development and correlates of a new scale. *Journal of personality assessment* (1986).

[14] Cynthia D Fisher. 1987. *Boredom: Construct, causes and consequences.* Technical Report. Texas A AND M Univ College Station Dept Of Management.

[15] Cynthia D Fisherl. 1993. Boredom at work: A neglected concept. *Human Relations* (1993).

[16] Michael Geden, Ana-Maria Staicu, and Jing Feng. 2018. The impacts of perceptual load and driving duration on mind wandering in driving. *Transportation research part F: traffic psychology and behaviour* (2018).

[17] Nitza Geri, Amir Winer, and Beni Zaks. 2017. Challenging the six-minute myth of online video lectures: Can interactivity expand the attention span of learners. *Online Journal of Applied Knowledge Management* (2017).

[18] David Guest, Roger Williams, and Philip Dewe. 1978. *Job design and the psychology of boredom.* Work Research Unit.

[19] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX: Results of empirical and theoretical research. In *Advances in psychology.*

[20] Stephen Hutt, Caitlin Mills, Shelby White, Patrick J Donnelly, and Sidney K D'Mello. 2016. The Eyes Have It: Gaze-Based Detection of Mind Wandering during Learning with an Intelligent Tutoring System. *International Educational Data Mining Society* (2016).

[21] Koji Inoue, Divesh Lala, Katsuya Takanashi, and Tatsuya Kawahara. 2018. Engagement Recognition in Spoken Dialogue via Neural Network by Aggregating Different Annotators' Models. In *Interspeech.*

[22] Melissa-Sue John, Ivon Arroyo, Imran Zualkernan, and Beverly P Woolf. 2014. Culturally aligned pedagogical agents for mathematics education. In *Fifth International Workshop on Culturally-Aware Tutoring Systems.*

[23] Christian Martyn Jones and Marie Jonsson. 2005. Detecting emotions in conversations between driver and in-car information systems. In *ACII.*

[24] Samira E Kahou, Xavier Bouthillier, Pascal Lamblin, Caglar Gulcehre, Vincent Michalski, Kishore Konda, Sébastien Jean, Pierre Froumenty, Yann Dauphin, and Nicolas Boulanger-Lewandowski. 2016. Emonets: Multimodal deep learning approaches for emotion recognition in video. *Journal on Multimodal User Interfaces* (2016).

[25] Steven J Kass and Stephen J Vodanovich. 1990. Boredom proneness: Its relationship to Type A behavior pattern and sensation seeking. *Psychology: A Journal of Human Behavior* (1990).

[26] Steven J Kass, Stephen J Vodanovich, and Anne Callender. 2001. State-trait boredom: Relationship to absenteeism, tenure, and job satisfaction. *Journal of business and psychology* (2001).

[27] Suzanne Keen. 2006. A theory of narrative empathy. *Narrative* 14, 3 (2006), 207–236.

[28] Willard A Kerr and Rudolph C Keil. 1963. A theory and factory experiment on the time-drag concept of boredom. *Journal of Applied Psychology* (1963).

[29] Kristina Krasich, Robert McManus, Stephen Hutt, Myrthe Faber, Sidney K D'Mello, and James R Brockmole. 2018. Gaze-based signatures of mind wandering during real-world scene processing. *Journal of Experimental Psychology: General* (2018).

[30] Iolanda Leite, André Pereira, Samuel Mascarenhas, Carlos Martinho, Rui Prada, and Ana Paiva. 2013. The influence of empathy in human–robot relations. *International journal of human-computer studies* (2013).

[31] Gil Levi and Tal Hassner. 2015. Emotion recognition in the wild via convolutional neural networks and mapped binary patterns. In *Proceedings of the 2015 ACM on international conference on multimodal interaction.*

[32] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf Between User Expectation and Experience of Conversational Agents. In *Proceedings of the CHI Conference on Human Factors in Computing Systems.*

[33] Amanda Markey, Alycia Chin, Eric M Vanepps, and George Loewenstein. 2014. Identifying a reliable boredom induction. *Perceptual and motor skills* (2014).

[34] Ana Paiva, Joao Dias, Daniel Sobral, Ruth Aylett, Polly Sobreperez, Sarah Woods, Carsten Zoll, and Lynne Hall. 2004. Caring for agents and agents that care: Building empathic relations with synthetic agents. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems.*

[35] James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: LIWC. *Mahway: Lawrence Erlbaum Associates* (2001).

[36] Martin Porcheron, Joel E Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice interfaces in everyday life. In *Proceedings of CHI.*

[37] Ronald Schroeter, Jim Oxtoby, Daniel Johnson, and Fabius Steinberger. 2015. Exploring boredom proneness as a predictor of mobile phone use in the car. In *Proceedings of the Annual Meeting of the Australian Special Interest Group for Computer Human Interaction.*

[38] Joseph Seering, Juan Pablo Flores, Saiph Savage, and Jessica Hammer. 2018. The social roles of bots: evaluating impact of bots on discussions in online communities. *Proceedings of the ACM on Human-Computer Interaction* CSCW (2018).

[39] Joseph Seering, Michal Luria, Geoff Kaufman, and Jessica Hammer. 2019. Beyond dyadic interactions: Considering chatbots as community members. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.*

[40] Jonathan Smallwood and Jonathan W Schooler. 2013. The restless mind. (2013).

[41] Fabius Steinberger, Ronald Schroeter, and Christopher N Watling. 2017. From road distraction to safe driving: Evaluating the effects of boredom and gamification on driving behaviour, physiological arousal, and subjective experience. *Computers in Human Behavior* (2017).

[42] Pierre Thiffault and Jacques Bergeron. 2003. Monotony of road environment and driver fatigue: a simulator study. *Accident Analysis & Prevention* (2003).

[43] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A Nicolaou, Björn Schuller, and Stefanos Zafeiriou. 2016. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *2016 IEEE international conference on acoustics, speech and signal processing.*

[44] Joseph Weizenbaum. 1966. ELIZA—a computer program for the study of natural language communication between man and machine. *Commun. ACM* (1966).

[45] Jacob Whitehill, Zewelanji Serpell, Yi-Ching Lin, Aysha Foster, and Javier R Movellan. 2014. Faces of engagement: Automatic recognition of student engagement from facial expressions. *IEEE Transactions on Affective Computing* (2014).

[46] Xi Yang, Marco Aurisicchio, and Weston Baxter. 2019. Understanding Affective Experiences with Conversational Agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.*

[47] Zhong Yin, Mengyuan Zhao, Yongxiong Wang, Jingdong Yang, and Jianhua Zhang. 2017. Recognition of emotions using multimodal physiological signals and an ensemble deep learning model. *Computer methods and programs in biomedicine* (2017).