# Character Initiative in Dialogue Increases User Engagement and Rapport

**Usman Sohail** **Carla Gordon** **Ron Artstein** **David Traum**
USC Institute for Creative Technologies
12015 Waterfront Drive, Los Angeles CA 90094-2536, USA
usman.f.sohail@gmail.com {cgordon|artstein|traum}@ict.usc.edu

## Abstract

Two dialogue policies to support character initiative were added to the Digital Survivor of Sexual Assault, a conversational agent designed to answer questions about sexual harassment and assault in the U.S. Army: (1) asking questions of the user, and (2) suggesting conversation topics after a period of inactivity. Participants who interacted with a system that had these initiative policies reported that they felt higher engagement and rapport with the character, compared to participants who interacted with the baseline system. There was also a positive correlation between the number of instances of character initiative in a dialogue and the level of engagement and rapport reported by participants.

## 1 Introduction

There is a large body of work discussing the efficacy and benefits of using conversational agents as educational and assistive tools (Rickel, 2001; Kerly et al., 2009; Bickmore et al., 2013; Graesser et al., 2014; Gardiner et al., 2017). Some of these systems are designed more for formal educational learning, while others are designed to educate with the intent of changing the user's behavior (Bickmore et al., 2013; Gardiner et al., 2017).

The Digital Survivor of Sexual Assault (DS2A: Artstein et al., 2019) was created to educate U.S. Army soldiers on the topic of sexual assault, in an effort to change attitudes and behavior and help prevent future harassment and assault. Inspired by the New Dimensions in Testimony project of conversation with Holocaust survivors (Traum et al., 2015), the DS2A system allows users to engage in a natural conversation with audio-visual recordings of Specialist Jarett Wright, a U.S. Army soldier who was sexually assaulted by senior members of his company while stationed in Iraq in 2010. Through conversation, users learn about Jarett's experiences of assault, retaliation, litigation, and other aspects of his life that were shaped by the assault. The interactive conversation is intended to forge a personal bond between the users and Jarett. The system is presently deployed by the U.S. Army in Fort Leavenworth, Kansas, and is used in training career professionals who deal with educating, preventing, and reporting of sexual harassment and assault in the Army (Figure 1).

Figure 1: The Digital Survivor of Sexual Assault system at the SHARP Academy in Fort Leavenworth, Kansas. Photo Credit: Stephen P. Kretsinger Sr.

An example conversation with Jarett is shown in Figure 2. This excerpt is typical of conversations with the DS2A system, and it shows the reactive nature of the system's operation: the system answers the user's questions, but does not take the initiative to ask questions of the user. However, there is evidence that asking questions helps the connection between two parties in a conversation. Burbules and Rice (1991) state that communicative virtues such as showing interest in a person's thoughts and opinions promote serious and open discussion. Nystrand and Gamoran (1991) found that student engagement has a direct correlation with academic achievement, with students who were more engaged being better able to absorb and retain the information presented. Addi-

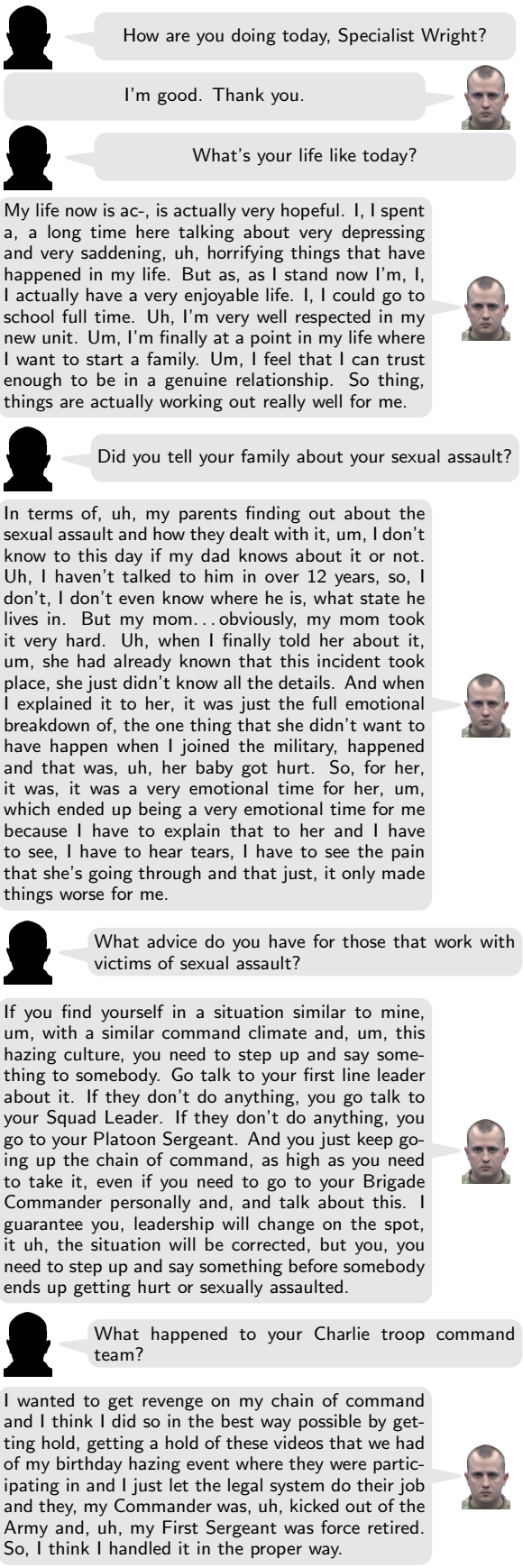


Figure 2: Sample conversation with Jarett Wright

tionally, Nystrand et al. (2003) underline the importance of maintaining student engagement with "authentic questions" which encourage them to share their own ideas and opinions. It is therefore reasonable to hypothesize that user engagement and rapport could be increased by giving the DS2A system an ability not only to respond to user questions, but also to engage the user by asking questions of its own.

Establishing rapport between users and conversational agents is an important component of creating engagement; in human-robot interaction it has been shown to increase customer comfort (Kanda et al., 2009) and social influence (Artstein et al., 2017). There has been a fair amount of work on using online measures to track user engagement in real time, using visual cues such as eye-gaze and head movement (see Sidner et al., 2005; Nakano and Ishii, 2010; Bohus and Horvitz, 2014); similar on-line measures have also been developed for assessing rapport (Zhao et al., 2016). However, such online measures were not available to us due to time and budget constraints, so we estimate user engagement and rapport using a post-interaction questionnaire.

This paper presents two main contributions: a set of policies for a reactive, question-answering character to take initiative and ask the user questions (section 2), and an experiment that shows that these policies increase user engagement and rapport, compared to interaction with a baseline system (section 3). We conclude by discussing some limitations of the experiment (section 4).

## 2 System Description

### 2.1 Baseline system

The baseline DS2A system is designed as an integrated application in the Unity game engine (`https://unity3d.com`); it incorporates several components from the USC ICT Virtual Human Toolkit (Hartholt et al., 2013), which is publicly available (`http://vhtoolkit.ict.usc.edu`). Input to the system is user speech, and the output is video clips of Jarett, recorded in our Light Stage studio. The overall system architecture is shown in Figure 3; a more detailed description is given in Artstein et al. (2019).

The baseline system uses a fairly standard pipeline for processing user utterances: audio from the user is sent to a speech recognizer, and the text output of the recognizer is sent to

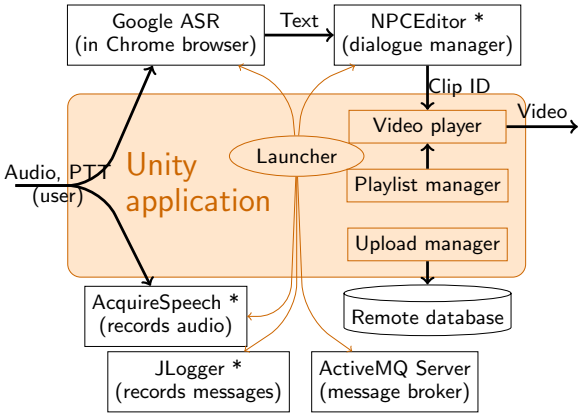


Figure 3: System architecture (* = Toolkit component)

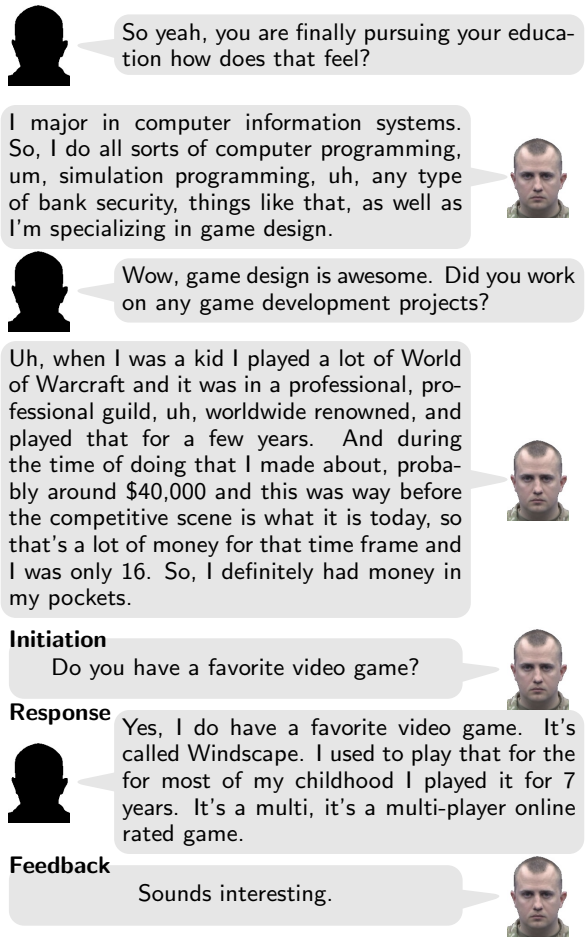


Figure 4: Follow-up question

the dialogue manager component, which selects a video response to be played back to the user. DS2A uses the NPCEditor component from the Virtual Human Toolkit (Leuski and Traum, 2011), which combines statistical Natural Language Understanding (NLU) with rule-based dialogue management. The NLU functionality is trained on question-answer pairs, and for each user question it returns a ranked list with zero or more appropriate responses (an empty list means that the classifier wasn't able to find an appropriate response). The dialogue manager functionality uses this ranked list to choose a response. The default action is to pick the top-ranked response; additional policies for avoiding repetition and handling non-understanding are described in section 2.3.

The baseline system is completely reactive: it acts in response to each user utterance, and it acts only in response to user utterances.

### 2.2 Mixed-Initiative Dialogue

We define character initiative as any character utterance which is not a reaction to a user utterance. To create a mixed-initiative dialogue experience, we implemented two types of character initiative by adding rules to NPCEditor's rule-based dialogue manager: *follow-up questions* and *timeout suggestions*.

**Follow-up questions** are designed to build rapport and encourage the user to engage more deeply with a topic; they are directly tied to the character utterances selected by the default dialogue manager. For example, if in response to a user question Jarett talks about his affinity for video games, he may follow up by asking the user about their video game preferences (Figure 4). If the user responds to the question, then Jarett will reply with a feedback utterance, completing an Initiation-Response-Feedback (IRF) exchange unit (Sinclair and Coulthard, 1975).

We annotated some of Jarett's responses with potential follow-up questions (see section 2.4 below); whenever Jarett issues a response, the dialogue manager checks to see if the response is associated with a potential follow-up question. In order to avoid asking too many questions, Jarett will ask a follow-up question only if an initiative utterance has not been delivered in any of the last three system utterances. Also, to avoid repeating himself, Jarett will not ask a follow-up question that he has already used in the same session, except for a few generic follow-ups which may be repeated, such as "How about you" or "If you were in my situation, what do you think you would have done?"

After Jarett asks a follow-up question, the system needs to determine whether the following user utterance is a response to the question. In principle this could be achieved by the NLU functional-

ity; however, at this point we do not have enough data to reliably train the NLU to identify user responses to follow-up questions. We therefore use a simple time-based heuristic, and assume that a user's utterance is a response to Jarett's question if a substantial portion of the question was actually played; in this case, Jarett will react to the user's utterance with a feedback utterance. However, if the user interrupts the initiative question more than two seconds before it is scheduled to end, it is assumed the user did not hear the initiative question, and the system will process the user's utterance using its default policy.

**Timeout questions** are designed to re-engage a participant after a period of inactivity. The time interval required to trigger a timeout varies between installations. The instructional system in Fort Leavenworth is typically used in front of a class, and we found that a threshold of 40 seconds provided a good balance between prompting the instructor and not being too disruptive. However, in piloting for the experiment reported below we found this threshold to be too long for one-on-one conversations, so we reduced it to 15 seconds for the experiment.

The timeout question can be a follow-up question to Jarett's previous utterance, if such a follow-up is available but wasn't asked previously due to the restriction of not asking too many follow-ups in succession. If a follow-up question is not available, then the timeout question utilizes the *topic suggestion* concept from NPCEditor's default dialogue manager (Leuski et al., 2006; Leuski and Traum, 2011). Originally designed to bring the user back into the domain of conversation after several consecutive instances of non-understanding (see section 2.3.2), we added topic suggestions as timeout questions, which serve not only to re-engage participants, but also to inform them of what the system can talk about in the event they cannot think of anything to ask. The system has 49 topic suggestion utterances covering 20 varied topics such as sexual assault prevention, reporting, retaliation, and bystander intervention. At the beginning of each session the system generates a list of all the topic suggestions, and then goes through the list throughout the session.

## 2.3 Additional dialogue policies

In addition to the policies above, the initiative system retains the reactive dialogue management policies of the baseline system. The policies below are all default policies that come with NPCEditor and are described in Leuski and Traum (2011). These policies are triggered by an incoming user utterance and its NLU interpretation, which takes the form of a (possibly empty) ranked list of appropriate responses by Jarett; they handle responding to user utterances, avoiding repetition, dealing with non-understanding, and special cases.

### 2.3.1 Responses and repetition avoidance

If the NLU functionality returns a non-empty list of responses, the dialogue manager will choose one of those responses to play to the user. The choice balances rank (for best match) and recency (to avoid repetition): if the list contains utterances that were not used by Jarett in the last ten turns, it will use the highest ranked of these; otherwise, it will use the least recently used utterance.

### 2.3.2 Handling of non-understanding

If the NLU returns an empty list, the dialogue manager uses a strategy designed to gradually move the speaker to saying something the system can understand (Artstein et al., 2009). With each successive non-understanding, Jarett will go further down the following list.

1. Say again: Jarett asks the user to repeat their question.
2. Don't understand: Jarett tells the user he doesn't understand their question.
3. Can't answer: Jarett informs the user he can't answer their question.
4. Topic suggestion: Jarett suggests a new topic.
5. Topic story: Jarett tells a story based on the topic suggested in step 4.

If at any point the user says something that is understandable (that is, the NLU returns a non-empty list), then the policy goes back to that in section 2.3.1 and the non-understanding counter resets to zero.

### 2.3.3 Special utterances

While choosing a response typically means playing the video clip associated with the response, NPCEditor also allows for special response tokens that do some other action. One such token is used in the DS2A system: replaying the previous response. This token is selected by the NLU as the

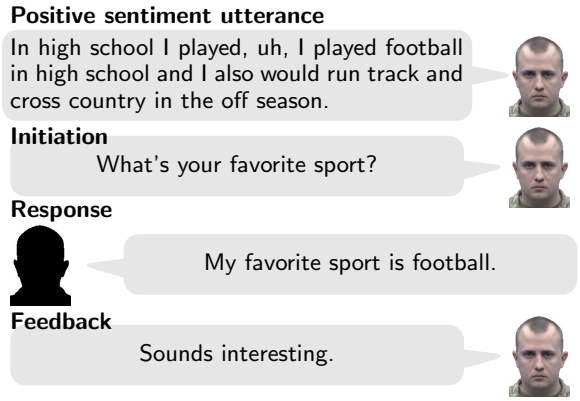


Figure 5: Follow-up to positive sentiment

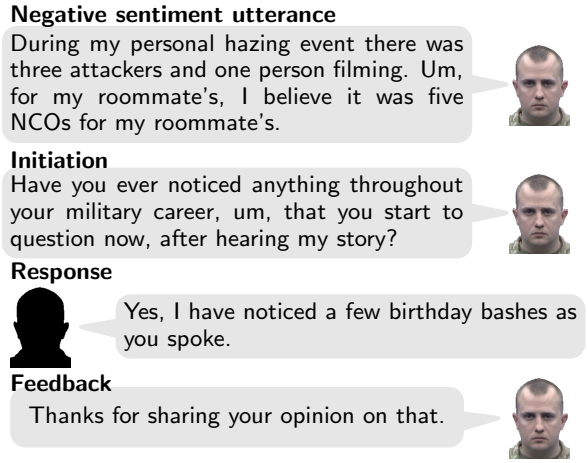


Figure 6: Follow-up to negative sentiment

interpretation of user utterances like "Could you repeat that?" If this token is selected by the dialogue manager (according to the policy in section 2.3.1), then Jarett will repeat his most recent utterance.

### 2.4 Annotations

The initiative policies require annotations that are not part of the default set-up of NPCEditor; these annotations were achieved by defining several new fields, detailed below (NPCEditor allows an unlimited number of user-defined fields).

**Follow-up Questions.** Each of Jarett's utterances is annotated with a field that lists (zero or more) potential good follow-up questions.

**Follow-up Sentiment.** Character utterances with follow-up questions are annotated with a sentiment label (positive, negative, or neutral). Jarett's feedback to a user's response matches the sentiment of the utterance that triggered the follow-up question (Figures 5 and 6).

**Utterance Length.** Each of Jarett's follow-up questions is annotated with its length, so that in case the question is interrupted by the user, the dialogue manager will know whether to issue a feedback utterance.

## 3 Experiment

In order to determine whether the mixed-initiative dialogue strategy has an effect on user engagement and rapport, we conducted an experiment comparing interactions with the baseline system and the initiative system. Each participant interacted with one version of the system, and we measured participant engagement and rapport using a post-interaction questionnaire. This section describes the experimental design and results, demonstrating that mixed initiative interaction does lead to increased engagement and rapport.

### 3.1 Method

**Materials.** We compared two versions of the DS2A system: a baseline system with the default dialogue management policies, and an initiative system with the default and initiative policies. The content was the same as in the system used by the U.S. Army, except that we removed some of Jarett's utterances which included "colorful" language. Participants interacted with the system on a MacBook Pro laptop running the Windows 10 operating system, using the laptop's built in display, speakers, microphone and trackpad.

**Participants.** A total of 58 participants were recruited through an ad on Craigslist (`https://www.craigslist.org`), paper flyers posted on a university campus, and a mass email sent to the Computer Science department. Participants were divided equally between the two experimental conditions. One participant was excluded from analysis because they chose to end the interaction early, resulting in a total of 29 participants in the baseline condition, and 28 in the initiative condition. The participants were 25 female, 32 male; most were between the ages of 18–27; the most common ethnic background was Asian; and the majority of participants had no affiliation or relation with the military (Table 1). Participants were given $10 as compensation for their participation in the study.

**Procedure.** Each participant interacted with either the baseline system or the initiative version of

| Age | |
|---|---|
| 18–27 | 50 |
| 28–37 | 4 |
| 38–47 | 1 |
| 48–57 | 2 |

| Gender | |
|---|---|
| Male | 32 |
| Female | 25 |

| Military Affiliation | |
|---|---|
| None | 53 |
| Close friend or family | 4 |

| Race | |
|---|---|
| Asian | 45 |
| Black/African American | 2 |
| White | 3 |
| Other | 7 |

Table 1: Demographics of study participants

**[Engagement questions]**

I was interested in hearing what Jarett had to say.
Jarett seemed interested in what I had to say.
During the interaction, I lost track of time.
**R** I found my mind wandering during the interaction.

**[Rapport questions]**

I felt Jarett created a sense of closeness or camaraderie between us.
**R** Jarett created a sense of distance between us.
I think Jarett and I understood each other.
**R** Jarett communicated coldness rather than warmth.
Jarett was warm and caring.
**R** I wanted to maintain a sense of distance between us.
I felt I had a connection with Jarett.
Jarett was respectful to me.
**R** I felt I had no connection with Jarett.
I tried to create a sense of closeness or camaraderie between us.
**R** I tried to communicate coldness rather than warmth.

Figure 7: Post-interaction questionnaire. Each question is rated on a 5-point scale. The label **R** indicates reverse-coded items.

the system for 20 minutes. Interaction took place in a quiet room on a university campus, and no experimenters or other personnel were present in the room during the interaction. This was done to ensure participants did not have any distractions in the room which might affect their overall engagement with the system, or their ability to build rapport. Participants were seated at a table in front of a laptop which displayed the video responses, and interacted with the system by pressing on the trackpad when asking their questions and releasing when their question was finished. At the end of the 20-minute interaction the experimenter reentered the room and administered two questionnaires: one with questions about demographic information, and one designed to quantify the level of engagement and rapport felt by the user.

| | Baseline | Initiative | Hi Init |
|---|---|---|---|
| Participants | 29 | 27/28 | 21/22 |
| Engagement | 3.60 | 3.84 | 3.95 |
| Rapport | 3.66 | 3.83 | 3.94 |

Table 2: Means of questionnaire responses

**Measures.** The engagement and rapport questionnaire was given on two sheets of paper, the first with the engagement questions and the second with the rapport questions (Figure 7). The questions about engagement were devised specifically for this study, while the questions about rapport were adapted from von der Pütten et al. (2010) and Artstein et al. (2017). Each question was rated on a 5-point scale: 1 Strongly Disagree, 2 Disagree, 3 Neither Agree nor Disagree, 4 Agree, 5 Strongly Agree. We devised two measures, one for engagement and one for rapport, by summing the responses to the positive questions, subtracting the reverse-coded questions, and normalizing the result to the interval 1–5.

We compared the baseline and initiative groups using t-tests, and used ANOVAs to test for interactions with gender. Since initiative behavior is dependent on the course of the dialogue, some participants in the initiative condition experienced very little initiative behavior by the system. It is not immediately clear how to treat these low-initiative dialogues in the initiative group, because the user experience in these dialogues is similar to that of the baseline group. We therefore tested comparisons both between the baseline group and the full initiative group, and also between the baseline group and a "high initiative" subset, defined (somewhat arbitrarily) as those members of the initiative group who experienced at least two initiative utterances in their conversation with Jarett.

For the initiative group, we also measured the correlation between the number of initiative utterances and the level of engagement or rapport.

### 3.2 Results

The mean values of engagement and rapport for the various groups are shown in Table 2. One participant in the initiative group and high initiative subset did not answer the questions on engagements, so they were excluded from the analysis of engagement. The engagement and rapport scores are highly correlated, both for the baseline group ($r = 0.54$, $df = 27$, $p < 0.005$) and for the initia-

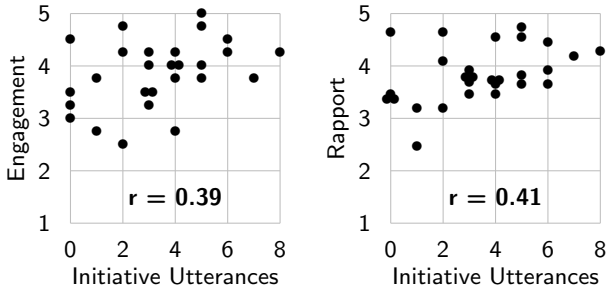

Figure 8: Positive correlation between number of initiative utterances and engagement/rapport

tive group ($r = 0.67$, $df = 25$, $p < 0.001$). This correlation could be interpreted as showing that the notions of engagement and rapport go hand in hand, or that the two instruments are actually tapping into a common notion, for example general satisfaction with the interaction, with the system, or with Jarett.

The difference in means between the baseline and initiative groups is not significant but shows a trend for engagement ($t(54) = 1.36$, $p < 0.1$) and a weak trend for rapport ($t(53) = 1.24$, $p = 0.11$). Between the baseline group to the high-initiative subset of the initiative group, the difference is significant for both engagement ($t(45) = 1.89$, $p < 0.05$) and rapport ($t(47) = 2.22$, $p < 0.05$). The above tests are one-tailed Welch's two-sample t-tests. ANOVAs found no effect of gender nor any interactions with gender.

For the initiative group, we also calculated the correlation between the number of initiative utterances in the dialogue and the participant's perceived engagement and rapport (Figure 8). Pearson's correlation is positive and significant for both engagement ($r = 0.39$, $df = 25$, $p < 0.05$) and rapport ($r = 0.41$, $df = 26$, $p < 0.05$).

### 3.3 Discussion

Our results suggest that the mixed-initiative dialogue strategy we employed increases the user's perception of their engagement and rapport with the DS2A system. The correlation of engagement and rapport scores with the number of system initiative utterances suggests that it is the character initiative that is responsible for the improved perception of the system.

We should make two notes about the observed correlations with system initiative utterances. First, since the length of the dialogues was fixed by the experiment at 20 minutes, we cannot tell whether the effect is due to the *number* of initiative utterances or the *rate* of initiative utterances. Second, recall that the rate of initiative utterances was explicitly limited by the dialogue policies (section 2.2): by design, the system is not capable of asking initiative utterances all the time, reflecting our belief that such behavior is undesirable. Therefore, the observed correlation should not be extrapolated into making conclusions about very high rates of initiative utterances.

## 4 Conclusion

Our experiment shows that the mixed-initiative dialogue strategy increases the levels of perceived user engagement and rapport. In conclusion, we wish to discuss several limitations of the experiment.

**Measuring engagement.** Much previous research has noted the difficulty of measuring engagement (see e.g., Nystrand and Gamoran, 1991; Cerrato and Campbell, 2017). While many studies have had success in measuring engagement online through the analysis of eye-gaze behaviors, affective postures, and other auditory and visual cues (e.g., Nakano and Ishii, 2010; Sanghvi et al., 2011; Yu et al., 2004; Huang et al., 2016), this was not a strategy available to us for this study. As noted above, our off-line engagement measure was highly correlated with our rapport measure; a better, more direct way of measuring participant engagement would be helpful.

**Repetitive feedback.** The system contains a total of 16 feedback utterances, and many of these are fairly similar (for example, all 8 instances of feedback to negative sentiment are variations on "thank you for sharing"). In informal discussion, some participants mentioned that the feedback they were receiving seemed repetitive; this may have negatively impacted engagement or rapport. We suspect that a system with more varied feedback utterances could have a more pronounced effect on engagement and rapport.

**Removed content.** Since this experiment was conducted on a non-military population, the U.S. Army requested that we remove some of Jarett's utterances that the Army felt were unsuitable for a civilian population (primarily utterances involving "colorful" language). The effect of removing these utterances was a slight reduction in overall coherence, since some user questions which could have

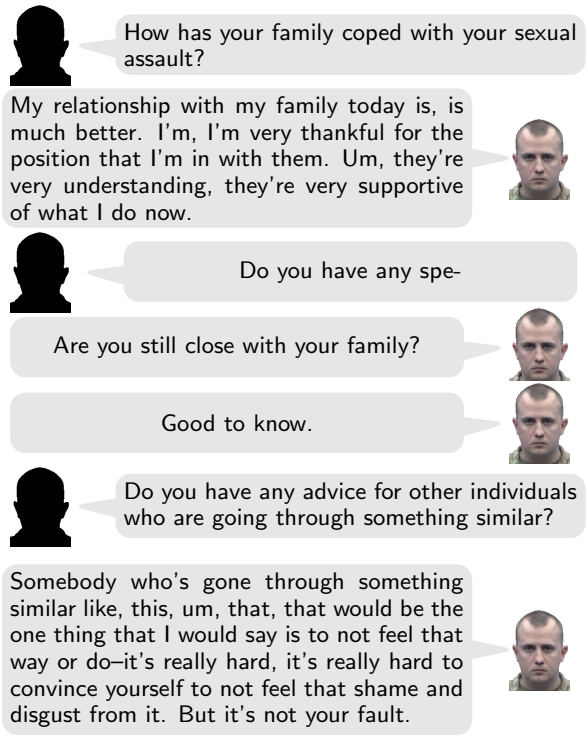


Figure 9: Timing issues

received a "colorful" response in the military system had to be treated as non-understanding in the experiment system. The lower coherence of civilian Jarett could have the effect of reducing participant engagement and rapport (though it would have a similar effect on both baseline and initiative conditions).

**Timing of follow-up questions.** As discussed in section 2.2, follow-up questions are triggered by a preceding utterance by Jarett, and are implemented by playing the question video clip directly after the conclusion of the video clip of the previous utterance. However, since each of Jarett's video clips is a self-contained conversational turn, the appearance is as if Jarett is yielding the turn to the participant, and then immediately claims the turn and asks a question. This somewhat unnatural sequence of cues often led participants to believe that Jarett was done talking at the end of his original response, so the participant would ask another question immediately; in some cases, participants interrupted early enough that they never heard Jarett's follow-up question, while in other cases they heard part of the follow-up question, which was cut short by their interruption.

Figure 9 shows a case in which Jarett's follow-up question cut off the participant's next question. Jarett's follow-up question started while the participant was mid-utterance, and an examination of the audio files reveals that the question was played in its entirety. Consequently, the system treated the participant's cut-off utterance as a response to the follow-up question, and Jarett immediately replied with the feedback "Good to know" even though this was not conversationally appropriate. This is an example of the unnatural conversational cues causing a communicative breakdown, and incidents like this may have had a negative effect on the participants' engagement and rapport. A better implementation of follow-up questions would be for the dialogue manager to somehow modify the ending of the trigger utterance clip, so that it does not give the impression of yielding the turn to the participant; however, this is not possible using the current tools.

Despite the above limitations, our study shows that a mixed-initiative dialogue strategy can lead to higher levels of perceived user rapport and engagement compared to a fully reactive strategy, when talking to agents designed to educate and inform users. As mentioned in the introduction, higher levels of engagement can lead to better retention of information, and higher levels of rapport can lead to increased social influence. This suggests that agents designed to educate for the purposes of effecting behavioral change would benefit greatly from implementing a mixed-initiative dialogue strategy.

## Acknowledgments

The work depicted here was sponsored by the U.S. Army Research Laboratory (ARL) under contract number W911NF-14-D-0005. Statements and opinions expressed and content included do not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

## References

Ron Artstein, Sudeep Gandhe, Jillian Gerten, Anton Leuski, and David Traum. 2009. Semi-formal evaluation of conversational characters. In Orna Grumberg, Michael Kaminski, Shmuel Katz, and Shuly Wintner, editors, *Languages: From Formal to Natural. Essays Dedicated to Nissim Francez on the Occasion of His 65th Birthday*, volume 5533 of *Lecture Notes in Computer Science*, pages 22–35. Springer, Heidelberg.

Ron Artstein, Carla Gordon, Usman Sohail, Chirag Merchant, Andrew Jones, Julia Campbell, Matthew Trimmer, Jeffrey Bevington, COL Christopher Engen, and David Traum. 2019. Digital survivor of sexual assault. In *IUI '19: Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 417–425, Marina del Rey, California. ACM.

Ron Artstein, David Traum, Jill Boberg, Alesia Gainer, Jonathan Gratch, Emmanuel Johnson, Anton Leuski, and Mikio Nakano. 2017. Listen to my body: Does making friends help influence people? In *Proceedings of the Thirtieth International Florida Artificial Intelligence Research Society Conference*, pages 430–435, Marco Island, Florida. AAAI Press.

Timothy W. Bickmore, Daniel Schulman, and Candace Sidner. 2013. Automated interventions for multiple health behaviors using conversational agents. *Patient Education and Counseling*, 94:142–148.

Dan Bohus and Eric Horvitz. 2014. Managing human-robot engagement with forecasts and... um... hesitations. In *Proceedings of the 16th International Conference on Multimodal Interaction (ICMI)*, pages 2–9, Istanbul, Turkey. ACM.

Nicholas Burbules and Suzanne Rice. 1991. Dialogue across differences: Continuing the conversation. *Harvard Educational Review*, 61(4):393–417.

Loredana Cerrato and Nick Campbell. 2017. Engagement in dialogue with social robots. In Kristiina Jokinen and Graham Wilcock, editors, *Dialogues with Social Robots: Enablements, Analyses, and Evaluation*, volume 427 of *Lecture Notes in Electrical Engineering*, pages 313–319. Springer, Singapore.

Paula M. Gardiner, Kelly D. McCue, Lily M. Negasha, Teresa Cheng, Laura F. White, Leanne Yinusa-Nyahkoon, Brian W. Jack, and Timothy W. Bickmore. 2017. Engaging women with an embodied conversational agent to deliver mindfulness and lifestyle recommendations: A feasibility randomized control trial. *Patient Education and Counseling*, 100:1720–1729.

Arthur C. Graesser, Haiying Li, and Carol Forsyth. 2014. Learning by communicating in natural language with conversational agents. *Current Directions in Psychological Science*, 23:374–380.

Arno Hartholt, David Traum, Stacy C. Marsella, Ari Shapiro, Giota Stratou, Anton Leuski, Louis-Philippe Morency, and Jonathan Gratch. 2013. All together now: Introducing the virtual human toolkit. In Ruth Aylett, Brigitte Krenn, Catherine Pelachaud, and Hiroshi Shimodaira, editors, *Intelligent Virtual Agents: 13th International Conference, IVA 2013, Edinburgh, UK, August 29–31, 2013 Proceedings*, volume 8108 of *Lecture Notes in Computer Science*, pages 368–381. Springer, Heidelberg.

Yuyun Huang, Emer Gilmartin, and Nick Campbell. 2016. Conversational engagement recognition using auditory and visual cues. In *Proceedings of Interspeech 2016*, pages 590–594, San Fransisco. ISCA.

Takayuki Kanda, Masahiro Shiomi, Zenta Miyashita, Hiroshi Ishiguro, and Norihiro Hagita. 2009. An affective guide robot in a shopping mall. In *HRI '09: Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction*, pages 173–180, La Jolla, California, USA.

Alice Kerly, Richard Ellis, and Susan Bull. 2009. Conversational agents in e-learning. In *Applications and Innovations in Intelligent Systems XVI: Proceedings of AI-2008, the Twenty-eighth SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence*, pages 169–182. Springer.

Anton Leuski, Jarrell Pair, David Traum, Peter J. McNerney, Panayiotis Georgiou, and Ronakkumar Patel. 2006. How to talk to a hologram. In *IUI '06: Proceedings of the 11th International Conference on Intelligent User Interfaces*, pages 360–362, Sydney, Australia. ACM.

Anton Leuski and David Traum. 2011. NPCEditor: Creating virtual human dialogue using information retrieval techniques. *AI Magazine*, 32(2):42–56.

Yukiko I. Nakano and Ryo Ishii. 2010. Estimating user's engagement from eye-gaze behaviors in human-agent conversations. In *Proceedings of the 15th International Conference on Intelligent User Interfaces (IUI)*, pages 139–148, Hong Kong. ACM.

Martin Nystrand and Adam Gamoran. 1991. Instructional discourse, student engagement, and literature achievement. *Research in the Teaching of English*, 25:261–290.

Martin Nystrand, Lawrence L. Wu, Adam Gamoran, Susie Zeiser, and Daniel A. Long. 2003. Questions in time: Investigating the structure and dynamics of unfolding classroom discourse. *Discourse Processes*, 35:135–198.

Astrid M. von der Pütten, Niole C. Krämer, Jonathan Gratch, and Sin-Hwa Kang. 2010. "It doesn't matter what you are!" Explaining social effects of agents and avatars. *Computers in Human Behavior*, 26(6):1641–1650.

Jeff Rickel. 2001. Intelligent virtual agents for education and training: Opportunities and challenges. In Angélica de Antonio, Ruth Aylett, and Daniel Ballin, editors, *Intelligent Virtual Agents: Third International Workshop, IVA 2001, Madrid, Spain, September 10–11, 2001 Proceedings*, volume 2190 of *Lecture Notes in Computer Science*, pages 15–22. Springer.

Jyotirmay Sanghvi, Ginevra Castellano, Iolanda Leite, Andre Pereira, Peter W. McOwan, and Ana Paiva. 2011. Automatic analysis of affective postures

and body motion to detect engagement with a game companion. In *Proceedings of the 6th International Conference on Human-robot Interaction (HRI)*, pages 305–312, Lausanne, Switzerland. ACM.

Candace L. Sidner, Christopher Lee, Corry D. Kidd, Neal Lesh, and Charles Rich. 2005. Explorations in engagement for humans and robots. *Artificial Intelligence*, 166:140–164.

John McHardy Sinclair and Richard Malcolm Coulthard. 1975. *Towards an Analysis of Discourse: The English Used by Teachers and Pupils*. Oxford University Press, London.

David Traum, Andrew Jones, Kia Hays, Heather Maio, Oleg Alexander, Ron Artstein, Paul Debevec, Alesia Gainer, Kallirroi Georgila, Kathleen Haase, Karen Jungblut, Anton Leuski, Stephen Smith, and William Swartout. 2015. New Dimensions in Testimony: Digitally preserving a Holocaust survivor's interactive storytelling. In Henrik Schoenau-Fog, Luis Emilio Bruni, Sandy Louchart, and Sarune Baceviciute, editors, *Interactive Storytelling: 8th International Conference on Interactive Digital Storytelling*, volume 9445 of *Lecture Notes in Computer Science*, pages 269–281. Springer, Heidelberg.

Chen Yu, Paul M. Aoki, and Allison Woodruff. 2004. Detecting user engagement in everyday conversations. In *Proceedings of the 8th International Conference on Spoken Language Processing (ICSLP)*, pages 1329–1332, Jeju Island, Korea. ISCA.

Ran Zhao, Tanmay Sinha, Alan W. Black, and Justine Cassell. 2016. Socially-aware virtual agents: Automatically assessing dyadic rapport from temporal patterns of behavior. In David Traum, William Swartout, Peter Khooshabeh, Stefan Kopp, Stefan Scherer, and Anton Leuski, editors, *Intelligent Virtual Agents: 16th International Conference, IVA 2016, Los Angeles, CA, USA, September 20–23, 2016 Proceedings*, volume 10011 of *Lecture Notes in Computer Science*, pages 218–233. Springer, Cham.