Freudbot: An Investigation of Chatbot Technology in Distance Education

Bob Heller
Centre for Psychology, Athabasca University
Athabasca, Alberta, Canada
bobh@athabascau.ca

Mike Procter
Centre for Psychology, Athabasca University
Athabasca, Alberta, Canada
mikeprocter@shaw.ca

Dean Mah
Centre for Psychology, Athabasca University
Athabasca, Alberta, Canada
deanm@athabascau.ca

Lisa Jewell
Centre for Psychology, Athabasca University
Athabasca, Alberta, Canada
lisaj@athabascau.ca

Billy Cheung
Centre for Psychology, Athabasca University
Athabasca, Alberta, Canada
billyc@athabascau.ca

Abstract: A chatbot named Freudbot was constructed using the open source architecture of AIML to determine if a famous person application of chatbot technology could improve student-content interaction in distance education. Fifty-three students in psychology completed a study in which they chatted with Freudbot over the web for 10 minutes under one of two instructional sets. They then completed a questionnaire to provide information about their experience and demographic variables. The results from the questionnaire indicated a neutral evaluation of the chat experience although participants positively endorsed the expansion of chatbot technology and provided clear direction for future development and improvement. A basic analysis of the chatlogs indicated a high proportion of on-task behaviour. There was no effect of instructional set. Altogether, the findings indicate that famous person applications of chatbot technology may be promising as a teaching and learning tool in distance and online education.

Chatbots are agents programmed to mimic human conversationalists. The first and still quite successful chatbot was ELIZA (Weizenbaum, 1966), a computer program designed to emulate a Rogerian therapist, a type of self-directed therapy where the patient's discourse is redirected back to the patient by the therapist usually in the form of a question. "Its name was chosen to emphasize that it may be incrementally improved by its users, since its language abilities may be continually improved by a "teacher". Like the ELIZA of Pygmalion fame, it can be made to appear even more civilized, the relation of appearance to reality, however, remaining in the domain of the playwright." (Weizenbaum, 1966, p.2)

The playwright in this case is the programmer but instead of classic Artificial Intelligence, ELIZA was programmed with rules to give the illusion of understanding. Essentially, ELIZA was programmed to recognize keywords and choose an appropriate transformation based on the immediate linguist context. Weizenbaum used the term 'script' to refer to the collection of keywords and associated transformation rules. Even though ELIZA is easily exposed as a fraud in the Turing sense, the popularity of the Rogerian therapist script remains high and there are a number of sites that allow you access to ELIZA. It is interesting to note that of all the scripts planned and developed by Weisenbaum, the Rogerian therapist script was the most enduring.

Arguably the most successful chatbot today is ALICE (Artificial Linguistic Internet Chat Entity), 3 time winner of the Loebner Prize, the holy grail for chatbots. ALICE was written by Richard Wallace and although no chatbot has passed the Turing test in the Loebner competition, ALICE has been judged the most human-like chatbot in 2000, 2001, and 2004.

Like ELIZA, ALICE has no true understanding and is programmed to recognize templates and respond with patterns according to the context. Moreover, like ELIZA, ALICE is incrementally improved with the addition of new responses. Unlike ELIZA, ALICE is programmed to talk to people on the web for as long as possible on any topic. Compared to the ELIZA's knowledge of 200 keywords and rules, ALICE is embodied by approximately 41,000 templates and associated patterns.

Perhaps the most important difference between ALICE and ELIZA is that ALICE is written in AIML (Artificial Intelligence Markup Language), an XML-based open source language with a reasonably active development community. There are also a variety of AIML parsers available written in Java, Perl, PHP, and C++ that permit interaction through a variety of interfaces, from simple web pages to Flash-based (or other) animation, instant messaging, and even voice input/output. In addition, Pandorabots, a web service that promotes and supports the use of ALICE and AIML is reporting support for over 20,000 chatbots on their site (http://www.pandorabots.com). At Pandorabots, would-be botmasters can easily create their own chatbot by modifying the personality of ALICE or by starting from scratch.

An AIML chatbot is suitable for many educational applications but our interest was in the famous personality application. Specifically, we were interested in whether students would enjoy and benefit from chatting with famous historical figures in psychology. As a distance education provider, we are always looking for ways to improve the interaction between student and course content over the web. Chatting with an historical figure via the internet may be intrinsically more interesting than the same information presented in a standard third party format over the web.

In terms of a theoretical rationale, there are several bases for investigating a famous personality application of chatbot technology as learning tool in distance education. Social constructionist theories of learning emphasize collaboration and conversation as a natural and effective means of knowledge construction and elaboration. The work of Graesser and colleagues on AutoTutor is based largely on these theories (see Graesser,Wiemer-Hastings, Wiemer-Hastings, Kreuz, & Tutoring Research Group 1999). A second rationale is found in the work of Cassell and colleagues on Embodied Conversational Agents (ECA). Cassell indicates that motivation for their research is based on the primacy of conversation as a natural skill learned early and effortlessly in life (Cassell, Bickmore, Campbell, Vilhjalmsson, & Yan, 2000). A conversational interface to a famous psychologist should be engaging and intuitive. A third rationale is provided through cognitive resource theory that argues linguistic rules governing conversational exchanges are automatic in nature due to frequency of use and consequently, free up additional resources to devote to encoding, understanding, and learning. Finally, according to the media equation (Reeves & Nass, 1996), people are predisposed to treat computers, television and other instances of media as people. They describe a number of experimental studies that generally show no differences in how media is 'treated' in comparison to people. The social rules that govern human-human interactions appear to govern human-computer interactions as well. If this is the case, then people may be predisposed to interact with a famous person application on the computer given the close fit of the application to human and conversational characteristics.

Freudbot

Like most disciplines, psychology has its share of founding figures and perhaps none is more familiar than Sigmund Freud. The selection of Freud seemed to be an obvious choice if the project was to be successful. Based on the first author's experience in the classroom, the subject of Sigmund Freud generates more group discussion than any other topic in an introductory psychology course.

Unlike ELIZA, Freudbot was not programmed to analyze or help users. Rather, Freudbot was programmed to chat in the first person about his theories, concepts and biographical events. Content was developed based on an existing Athabasca University resource that contained dictionary-type definitions of Freudian terms, concepts, & theories. Information from a Freud biography was used to create content on Freud's autobiographical events (Breger, 2000).

Early on in the development of Freudbot, it was unclear regarding how much of the ALICE personality/content should be included. Based on some pilot work, it was decided that a Freud-focused chatbot would probably lead to better evaluations.

Although the content for Freudbot was developed in AIML, there were a number of ELIZA-like control features built into Freudbot. Like ELIZA, Freudbot would 'recognize' certain key words or combination of words and respond accordingly. In cases, where no input was recognized, Freudbot would default to one of several strategies at random; ask for clarification, suggest a new topic for discussion, indicate that he had no response, or ask the user for a new topic. Another type of conversational pickup strategy was to ask a question and then redirect to a known topic. For example, Freudbot could ask a question like "Are you happy?" and regardless of how the user responded, Freudbot would respond with a leading statement like "That reminds me of the pleasure principle." This feature was designed to lead the user back to a discussion of Freudian topics.

In addition to the ELIZA-like control features, Freudbot was also programmed to 'release' content in a conversational or sequential fashion. If a user asked about a particular concept or episode, Freudbot was designed to provide an answer with implicatives that would invite the user to request more information using conversational directives (i.e. go on, tell me more, is that all, why is that? etc.). We hypothesized that this would be consistent with the conversational rules related to turn-taking.

Study Description

The purpose of the investigation was to explore the experience of having students interact with a chatbot designed to emulate Sigmund Freud and programmed to discuss Freudian concepts, theories and biographical events. To measure the experience, chat participants completed a web-based questionnaire to assess their subjective impressions immediately after a 10 minute chat with Freudbot over the web. Seven of the questions covered impressions of the chat where participants provided responses on a 5-point bi-directional scale (e.g. 1 - not very engaging and 5 - very engaging). Participants were also asked if they would chat with Freudbot again, a measure that was use to partition the group into those with positive and negative chat experiences. Finally, participants were asked to rate the value of five other chatbot applications and provide direction on five ways of improving Freudbot using a 5-point importance scale where 1 was least important and 5 was most important.

The chatlogs were also examined to quantify the extent of on-task student participation and the judged success of Freudbot in relation to conversational exchanges. Participant exchanges were categorized as either on or off task relative to the study instructions and then subcategorized as either a statement, a question, a one word sentence, or uncategorized. To control for differences in output, the percentage of each type of exchange was calculated for each individual.

In addition, the study was also designed to examine whether the instructional set provided (i.e. instructions on how to talk with Freudbot) would affect chat behaviour and/or evaluation. Based on some pilot work it was suggested that instructions on how to chat may improve ratings for those participants who are unfamiliar with chatbot technology or the task in general. Alternatively, since the task is natural and intuitive, instructional detail may have not effect on overall ratings. To examine these hypotheses, participants were randomly assigned to one of two sets. In the brief set, participants were simply directed to chat with Freudbot and given general tips on input formation. In the elaborate set, students were given elaborate instructions on how to interact with the chatbot and areas of discussion in addition to the general tips.

There were 53 participants in the present study (10 men and 43 women). The participants were Athabasca University students currently or recently registered in one or more psychology courses. Approximately half of the participants listed themselves as full-time (45%) students and half as part-time (53%). Participants were recruited into the study primarily through an email request from the first author. A monetary incentive to participate was also provided (1/30 chance in a draw for $300).

After providing informed consent, participants were randomly directed to a web page containing either brief instructions on how to chat or elaborate instructions. After reading the instructions, they advanced to the chat page that displayed an image of Freud on the right hand side and a user text box and Freudbot response on the left hand

side. Once connected to the chat page, Freudbot would respond with "Hello, my name is Sigmund Freud. What would you like to talk about?". After 10 minutes of unstructured chat behaviour, participants were automatically directed to a web questionnaire survey where they completed questions on their chat experience and relevant demographic variables.

Results

Table 1 displays the average ratings for each of the questions used to measure chat experience on a 5-point scale where higher values were associated with positive experiences. Across the entire sample, participants were generally neutral towards Freudbot as the average rating for 6 of the questions ranged from 2.83 to 3.08. The exception was for a question about expansion to other areas where participants were much more positive about the role of a chatbot. Ratings on this question were significantly higher than the ratings on next highest rated question, level of engagement, t (52) = 5.31, $p < .001$. Of the participants who were asked whether they would be willing to chat again with Freudbot, 68% (n=36) of the participants indicated that they would chat again compared to 32% who said no.
To examine whether elements of the chat experience were rated differently between the two groups, t tests were used to compare ratings. Levine's test of homogeneity of variance was also calculated and the degrees of freedom were adjusted in cases where the assumption of homogeneity was violated. The ratings for the two groups are shown in the right half of Table 1. Interestingly, the group of participants who said yes to another chat rated their chat experience significantly higher across all questions than did the group who said no (all p's < .01).

Chat Experience	Entire Sample n=53		Chat again? Yes n=36		Chat again? No n=17	
	M	SD	M	SD	M	SD
Would you recommend that this activity be expanded to include other topics?*	4.00	1.13	4.39	.90	3.18	1.13
How engaging did you find this activity?*	3.08	1.12	3.50	.85	2.18	1.07
How memorable do you think this activity will be for learning about Freud?*	3.04	1.06	3.53	.81	2.00	.71
Overall, how would you rate this activity?*	3.02	.93	3.42	.65	2.18	.88
How useful did you find this activity for learning about Freud?*	2.96	.94	3.22	.90	2.41	.80
How enjoyable did you find this activity?*	2.92	1.05	3.22	.76	2.29	1.31
Would you recommend this activity to others?*	2.83	1.12	3.36	.80	1.63	.72

Table 1: Freudbot chat experience ratings based on a 5-point scale where higher numbers reflect positive experiences. Note: * - statistically significant difference ($p < .01$) between the group that indicated they would chat again (Chat again? yes) and the group that would not (Chat again? no).

Participants were also asked to rate the importance of planned improvements to Freudbot and the application of chatbots into other areas. The top half of Table 2 displays the rated improvements to Freudbot in descending order of importance. As can be seen in the first column, Chat Behaviour was clearly the highest rated improvement and significantly higher than Audio Response, the next highest rated improvement, t (52) = 4.38, $p < .001$. In addition, Audio Response was rated significantly higher than Voice Recognition, the next highest rated improvement, t (52) = 3.6, $p < .001$. The ratings for the 3 remaining improvements were not statistically different from each other. The last 4 columns in the top half of Table 2 show the means for the two groups that indicated they would or would not chat again with Freudbot. There were no major differences in the order of rated improvements within each group. The group that said yes to another chat had significantly higher ratings on the Animation/movement improvement than the group that said no, t (44.6) = 2.3, $p < .05$.

	Entire Sample n=53		Chat again? Yes n=36		Chat again? No n=17	
Freudbot Improvements	M	SD	M	SD	M	SD

Chat Behaviour	4.15	1.35	4.22	1.20	4.00	1.66
Audio Response	3.15	1.41	3.08	1.40	3.29	1.45
Voice Recognition	2.53	1.51	2.44	1.52	2.71	1.53
Synchronization	2.47	1.40	2.64	1.46	2.12	1.22
Animation/movement*	2.26	1.26	2.50	1.34	1.76	.90
Chatbot Applications	M	SD	M	SD	M	SD
Practice Quizbot*	4.08	1.22	4.42	.84	3.35	1.58
Famous Personality*	4.06	1.07	4.33	.89	3.44	1.21
Course Content	3.32	1.27	3.53	1.18	2.88	1.36
Chatroom	3.23	1.25	3.42	1.23	2.81	1.22
Course Administration	3.17	1.22	3.19	1.22	3.12	1.27

Table 2: Freudbot improvement and chatbot application ratings based on a 5-point scale where higher numbers are positive. Note: * - statistically significant difference ($p < .01$) between the group that indicated they would chat again (Chat again? yes) and the group that would not (Chat again? no).

In the bottom half of Table 2, the first column displays the average rating across the entire sample for each applications in descending order of importance. Practice Quizbots and Famous Personality applications were the highest rated applications and both were rated significantly higher than the next application, a Course Content chatbot, t (52) = 3.93, $p < .001$ and t (51) = 3.58, $p < .001$, respectively. The ratings for the remaining applications were not statistically different from each other. In terms of the two groups who indicated they would or would not chat again, there were no major differences in the order of rated importance the rated applications. In addition, Practice Quizbots and Famous Personality applications were rated significantly higher by the chat-again group compared to the group that said no, t (20.4) = 2.61, $p < .05$ and t (50) = 2.98, $p < .01$, respectively.

Although the duration of the chat was limited to 10 minutes, there was considerable variability in the number of exchanges as the average number of exchanges was 27.7 with a range of 6 to 115 and a standard deviation of 20.3. In general, participants were quite cooperative in the task as close to 90% of their input was judged to be on-task in the form of a statement (36.5%), a question (46.7%), a one word response (4.7%) or some other combination (1.8%). Less cooperative was Freudbot as approximately 61% of Freudbot's responses were judged to be appropriate. There was also a significant positive correlation between the proportion of on-task user responses and a global measure of chat evaluation ($r = .27$, $p < .05$)

To examine the manipulation of instructional set, the mean ratings from the seven measures of chat experience were analyzed in a Multivariate Analysis of Variance with Instructional set as a between groups variable. There was no effect of instructional set, F (7, 44) = 1.47, $p > .05$. Nor was there any significant differences in the number of exchanges between groups, t (51) = .59, $p > .05$, or in the proportion of on-task chat behaviour, t (50) = .10, $p > .05$.

Discussion

Although far from definitive, the present investigation offers mildly positive evidence regarding the utility of a famous personality chatbot application in on-line education. The endorsements of Freudbot were mostly neutral but, participants did note the potential of such chatbots and recognized that improvements could be made. As a learning activity, chatting with a famous personality is natural and intuitive and borne out by the observations that 90% of the participant's chat behaviour was classified as on-task and the proportion of on-task behaviour was related to overall evaluations. In addition, there were no differences in chat behaviour as a result of instructional set, a null effect that is consistent with the natural and intuitive nature of the interface.

Famous personality applications were also rated more highly than 3 other course-related chat agents. In the words of one participant; "It was pretty cool the way it felt like I was actually interacting with Freud... he's deceased though, yeah, but the picture, the fast answers... made me pay attention to the answers alot more than if I had been simply reading a text written by someone else. Plus it was cool to feel like I could voice my own opinion with the most well-known psychoanalyst of all time."

In terms of improvement, it is clear that Freudbot has lots to learn. Participants rated improvement in chat behaviour higher than any of the other possible improvements. An informal analysis of Freudbot's sequential conversational output indicated that conversational depth was generally quite shallow. Similarly, Freudbot's attempt to redirect conversation met with varying degrees of success. Asking for clarification or indicating no response appeared to be the least successful strategy whereas suggesting a new topic for discussion or asking a question in order to redirect was generally quite successful. The room for improvement is generally consistent with the development process for AIML chatbots and ELIZA where the creation of a successful chatbot is an incremental process. The chatlogs from this study are currently being used to refine Freudbot and improve overall performance and a formal process for improving performance is being developed.

The present study is similar to another investigation involving AIML carried out at the University of Huddersfield. Gibbs and colleagues (2003) created Emile, an AIML chatbot designed to discuss several leading social theorists (i.e. Marx Mead, Weber, and Foucoult) in either the first person, if asked, or in the third person. According to the project report, the 15 students who participated in the study expressed difficulty in attaining sought-after information but they were optimistic about the potential of Emile and felt that Emile was helpful as a tool. Interestingly, none of the students indicated that the first person feature of Emile should be disabled when asked to rate improvements and 87% of the students indicated that the ability to emulate a social theorist should be enhanced. Unlike Freudbot, Emile did not display an image of the theorist during the chat.

The differences in visual presentation between Freudbot and Emile draw attention to the research around ECA and the role of animacy in computer interfaces. Beun, de Vos, & Witteman (2003) defines ECAs as "electronic agents that visually presented in the computer interface with some kind of embodiment – human animal or fantasy figure. " (p315). Beun et al. found that the presence of a visual image improved performance on a learning and memory task The case for social agency as means of fostering deeper learning in interactions with animated pedagogical agents is also made by Moreno, Mayer, Spires, & Lester (2001). In future work, we plan to explore the role of animacy in famous personality applications of chatbots.

References

Breger, L. (2000). *Freud: Darkness in the midst of vison*. New York: Wiley.

Cassell, J., Bickmore, T., Campbell, L., Vilhjalmsson, H., & Yan, H. (2000). Human conversation as a system framework: Designing embodied conversational agents. In J. Cassell, J. Sullivan, S. Prevost, and E. Churchhill (Eds.), *Embodied Conversational Agents* (pp. 29 – 63). Cambridge: MIT Press

Graesser, A. C., Wiemer-Hastings, K., Wiemer-Hastings, P., Kreuz, R. & Tutoring Research Group (1999). Auto Tutor: A simulation of a human tutor. *Journal of Cognitive Systems Research, 1*, 35-51.

Moreno, R., Mayer, R. E., Spires, H. A., & Lester, J. C. (2001). The case for social agency in computer-based teaching: Do students learn more deeply when they interact with animated pedagogical agents. *Cognition and Instruction, 19(2),* 177-213.

Reeves, B. & Nass, C. (1996). *The Media Equation: how people treat computers, television and new media like real people and places*. CLSI Publications

Beun, R. J., Vos, E. De & Witteman, C. L. M. (2003). Embodied conversational agents: Effects on memory performance and anthropomorphisation. *Springer Lecture Notes in AI*, IVA, 315-319.

Gibbs G. R. (2003). Emile: Using a chatbot conversation to enhance the learning of social theory. (Tech. Rep. No. 05/S/01) Queensgate, Huddersfield: University of Huddersfield, Centre for Learning and Teaching – Sociology, Anthropology, and Politics.

Weizenbaum, J. (1966). ELIZA--A computer program for the study of natural language communication between man and machine. *Communications of the ACM, 9(1),* 36-35.