

How do Humans Interact with Chatbots?: An Analysis of Transcripts

Mina Park

School of Business, Southern Connecticut State University, New Haven, CT 06515

Email: parkm2@southernct.edu

Milam Aiken

School of Business Administration, University of Mississippi, University, MS 38677

Email: maiken@bus.olemiss.edu,

Laura Salvador

School of Business Administration, University of Mississippi, University, MS 38677

Email: lasalvad@olemiss.edu

Abstract

Chatbots enable machines to emulate human conversation, and recent developments have resulted in many online systems for the public to use. Although a few studies have investigated how humans interact with such programs, we are not aware of any that have analyzed transcripts in depth. In this study, students interacted with two Web-based chatbots, Rose and Mitsuku, for five minutes and evaluated how well they thought the software emulated human conversation. We reviewed the transcripts and found that students used fairly simple language and made many text errors. There were no significant differences between the two systems in our experimental measures, but we found that Rose tended to change the topic more often and Mitsuku seemed more argumentative.

Keywords: Artificial intelligence, Intelligent agents, Chatbots, Conversation, Rose, Mitsuku

Introduction

A chatbot (aka 'chatterbot' 'talkbot', 'IM bot', 'dialog system', 'machine conversation system', 'virtual agent', 'interactive agent', or 'conversational agent' is a computer program that emulates human conversation [16, 24]. Other programs such as Siri, Alexa, and Cortana (aka, 'personal digital assistants' or 'virtual assistants') were not designed for conversation but rather to perform actions such as play music, retrieve information, or provide notifications.

Chatbots provide useful services in many ways [3]. They can be incorporated on Web sites to emulate lawyers, doctors, personal stylists, concierges, financial advisers, fitness trainers, teachers, and technical support personnel to provide answers to users' questions and provide a superior user interface [2]. For example, companies such as the Royal Bank of Scotland, Mattel, Domino's, Disney, and Renault use chatbots to provide a more natural user interface for visitors to their websites. In addition, they can provide emotional support for lonely people by imitating pets, romantic partners, or just friends [11]. Users converse with chatbots about dating advice, being bullied at school, and coping with illness [16]. Finally, these programs have been used in chat rooms to stimulate conversation or allow students to practice conversations in a foreign language [7, 32].

These agents can also be used for nefarious purposes, however. For example, Ashley Madison, a website serving users wishing to engage in extra-marital affairs, created more than 70,000 chatbots pretending to be

actual women to provide the illusion that there was a large number of females using the site, and thus, increase the number of male customers [17]. In addition, these systems have been used in chat rooms to entice people into revealing personal information, such as bank account numbers.

The use of chatbots is fairly new and still growing, and relatively little research has been conducted into how they are used by people. In the next section, we review prior research and then describe an experiment in which students conversed with two prominent chatbots, Rose and Mitsuku. We analyze the results and discuss future possible studies.

Background

Chatbots

The study of conversational agents is over 50 years old as the first chatbot (Eliza) was developed in 1966 [29]. At this early stage in artificial intelligence, Eliza did not truly 'understand' the user's conversation but merely rephrased the typed text or generated generic responses in an attempt to appear to be human. Subsequently, dozens of chatbots were developed, but relatively few are available publicly, online. Two of these include:

- **Mitsuku** (<http://www.square-bear.co.uk/mitsuku/chat.htm>). Mitsuku is perhaps the best chatbot available online (most adept at emulating a human conversation) as it achieved first place in the 2018 competition for the Loebner Prize, the fourth time it has done so [26, 30]. In this replication of the Turing Test [6, 21, 23, 27], four human judges typed messages to two chatbots and a human, without knowing which was which, and scores were assigned to each entity based upon the naturalness of responses (most human-like). Mitsuku takes the perspective of an 18-year-old girl, and a cartoon image of her appears on the Web page (see **Figure 1**). Development on Mitsuku was begun in 2005, and the site claims to have had about 10 million visitors since 2010. Like many chatbots, knowledge is stored in AIML files, and visitors are allowed to teach the system.
- **Rose** (<http://ec2-54-215-197-164.us-west-1.compute.amazonaws.com/speech.php>). Rose is another leading chatbot winning second place in 2017's Loebner Prize with Mitsuku the winner. Rose won the competition in 2014 and 2015. This chatbot also takes the perspective of a female, a computer hacker living in San Francisco, with a cartoon image of her appearing on the page (see **Figure 2**).

Figure 1. Mitsuku Chatbot Web User Interface

Figure 2. Rose Chatbot Web User Interface

Conversation

Visitors to the sites may type comments and questions, and some allow users to talk rather than type, receiving responses via computer-synthesized speech [31]. Many chatbots answer questions, and even perform simple math. Some, such as Mitsuku, include a limited amount of reasoning, such as being able to correctly answer, "Which is bigger, Chicago or a bread box?" The following conversation with Mitsuku provides an illustration of chatbot use:

Human: What is your name?

Mitsuku: My name is Mitsuku.

Human: What's new?

Mitsuku: not much what's new with you?

Human: What is it like where you live?

Mitsuku: Much the same as it would be for you.

Human: What is $2 + 2$?

Mitsuku: 2 plus 2 is four.

Prior studies

Research has investigated how users converse with these agents by analyzing the transcripts generated [8, 19]. In one study [9], 100 human conversations were compared to 100 exchanges with the Cleverbot chatbot. Results showed that people used more, but shorter, messages when communicating with chatbots. People also used a more restricted vocabulary and greater profanity with chatbots. Microsoft discovered obscene and rude text in conversations with its Tay chatbot, and in 2016, the company shut down access to it [20]. Users of these systems are often disinhibited because their identities are hidden, and they may enter unsavory text, perhaps for entertainment or to test the system to see how it will respond [1]. Another study [4] found that 10% of the text used in chatbot conversations contained abusive language, and 11% was related to hardcore sex. As a result, several chatbot websites have admonishments for users to keep the conversation 'clean'.

People might behave differently in other ways when communicating with chatbots. For example, even though some might perceive them as living entities [19], at least one study [28] found that chatbots are perceived more negatively. In addition, people communicating with chatbots might be less open, agreeable, extroverted, and conscientious than when communicating with others [18].

Experimental study

Purpose

As discussed earlier, prior analyses of conversations between humans and chatbots have found that humans type less text when dealing with chatbots than they do when communicating with other humans, and they tend to be more abusive in language. In this study, we attempt to replicate these results and to investigate relationships among the variables more in depth [22, 25].

Description

A group of 29 students at a Business school in the northeast region of the United States exchanged text with Rose as if they were conversing with a friend. Another group of 29 students exchanged comments with Mitsuku. Approximately 34% of the two groups were women. Both groups interacted with the systems for about five minutes, an interval we assumed would be sufficient for an informal conversation. After the sessions, students were asked to evaluate the experience on a 1 (strongly disagree) to 7 (strongly agree) scale: "When communicating with the chatbot, the responses were very natural, similar to what a human would say."

An objective reviewer analyzed the transcripts of all the conversations by counting the number of comments, the number of words, the number of spelling and grammatical errors (as determined by the *Microsoft Word* spell checker), and the Flesch Reading Ease [5] score of the text generated by the students (also determined by *MS Word*). This score measures average sentence length in words and average word length in syllables in an attempt to determine the overall difficulty of the text on a scale of 0=difficult to 100=easy. In addition, the reviewer counted the number of comments she deemed to be rude, insulting, vulgar, or sexually explicit. Finally, the reviewer determined the appropriateness of the chatbots' replies to users' comments on a scale of 1=nonsensical reply to 7=good reply [10, 12, 13].

Table 1. Summary of variables

Variable	Min	Max	Mean	Std. Dev
Overall				
1: User text comments	6	169	23.6	21.3
2: User text words	44	203	115.8	41.4
3: User text reading difficulty (0=difficult, 100=easy)	0	100	81.2	31.1
4: User text spelling and grammar errors	0	38	12.8	8.6
5: User text inappropriate comments	0	2	0.3	0.5
6: User evaluation of chatbot (1=good, 7=bad)	1	7	3.9	1.4
7: Review evaluation of appropriate replies (1=no, 7=yes)	1	7	3.5	1.5
Rose				
1: User text comments	6	169	22.9	28.9
2: User text words	44	188	107.5	38.0
3: User text reading difficulty (0=difficult, 100=easy)	0	100	82.7	28.6
4: User text spelling and grammar errors	1	38	13.3	9.2
5: User text inappropriate comments	0	1	0.3	0.5
6: User evaluation of chatbot (1=good, 7=bad)	1	6	3.7	1.4
7: Review evaluation of appropriate replies (1=no, 7=yes)	1	6	3.1	1.3
Mitsuku				
1: User text comments	6	169	24.4	9.3
2: User text words	52	203	124.0	43.7
3: User text reading difficulty (0=difficult, 100=easy)	0	100	80.0	34.0
4: User text spelling and grammar errors	0	32	12.4	8.1
5: User text inappropriate comments	0	2	0.3	0.6
6: User evaluation of chatbot (1=good, 7=bad)	1	7	4.0	1.4
7: Review evaluation of appropriate replies (1=no, 7=yes)	2	7	3.9	1.6

Experimental Results

Results of the experiment are shown in **Table 1**. There was a significant negative correlation between the number of comments and reading ease ($R = -0.31, p = 0.01$), (that is, more comments indicated easier text), but there was no significant correlation between the number of words and reading ease. Thus, many comments with few words were associated with less difficult text, as this is how the reading ease score is calculated. There was also a significant correlation between the number of comments and the number of errors ($R = 0.32, p = 0.01$) and the number of words and errors ($R = 0.27, p = 0.04$). That is, there were more opportunities for errors with more text. Also, the number of inappropriate comments was significantly correlated with text errors ($R = 0.31, p = 0.02$). However, because there were so few inappropriate comments, this statistic might not be reliable. There was no significant difference between the two groups on any variable.

The students were relatively informal in their conversations with the chatbot, as the average Reading Ease measure of 81.2 is the equivalent of about 5th grade level, and there was a significant difference with a reading level of grade 12 or college level (Reading Ease = 46) ($t = 8.6, p < 0.01$). This is not surprising because students often chat with each other with instant messaging using short sentences and simple words.

The student evaluations of the chatbots' performance was neither good nor bad as there was no significant difference between the mean rating of 3.9 and a median level of 4 on the 1-to-7-point scale.

Transcript Analysis

A further detailed analysis of the transcripts revealed the following characteristics:

Very few rude comments. Unlike in previous studies, e.g. [4], there were very few comments judged inappropriate by the reviewer (1.2% of all comments as compared to 10% in the other study). Perhaps this was because the students were told that the transcripts would be reviewed later, and thus were more careful of what they wrote.

Many users were very informal and made many errors. As noted earlier, the readability of their text was on the equivalent of a 5th-grade level rather than college level. In addition, the students made several spelling and grammatical errors, as determined by *MS Word*. For example, many comments had a lack of capitalization at the beginning of sentences, and there were many missing apostrophes in contractions. Some comments with errors include:

thats just the name ive been given

i don't really like it its kind of slow and boring

But you do ot know me

they tech students

yes it dos

Tellme about it

How is yuour day going??

hi today its a wed in october

that makes no sense to be honets

nopoeeee

i heard soemthing aobutyou

i reaaly dont knoe

In addition, *MS Word* considered passive sentences ungrammatical. Consequently, about half of the comments had spelling or grammatical errors. Nevertheless, the chatbots were often able to understand what was meant, despite errors.

Interestingly, the chatbots also made mistakes, but less often. For example:

Mitsuku: Hmm.. a time lets you know when to do things [Start of sentence not capitalized, missing punctuation]

Rose: When I was eight years old, I fell and hurt my knee. When I was eight years old, I fell and hurt my knee.; I remember I asked the doctor, Can I have a new knee please? [Needless repetition]

Poor transitions and responses. The reviewer gave an overall, neutral score of 3.5 for appropriate responses to the users' comments, with Rose scoring slightly less than Mitsuku. That is, several chatbot comments had little if anything to do with what the user typed and could be considered a change of topic. However, humans also exhibit such behavior in normal conversations. Or, the system's reply was not something a human would normally say. For example, the following conversations with Mitsuku seemed odd:

Human: Do you think global warming is real?

Mitsuku: Global warming is when the Earth heats up due to humans causing too much pollution.

Human: So is that a yes?

Mitsuku: It seems we have something in common.

Human: Okay good, you're smart.

Human: why do you love football

Mitsuku: Love is all we need.

Human: Tell me about it

Mitsuku: Ambiguous: "Tell me about it." -- colloquial or "Tell me about user16".

Human: I don't go to school, I teach myself

Mitsuku: Do you like your school?

Human: It was as simply put as possible

Mitsuku: What made it so put.

In the following conversations with Rose, the comments seem rather disjointed and unrelated:

ID1: I listen to all most every type of music

Rose: What would you do if you won a million dollars.

ID1: I would invest some, help my family out and friends in need

Rose: Didn't Shirley Yeung star in that?

ID1: I los

Rose: Great.

Rose: Where did you go on your first date?

ID2: The park

Rose: So... Do you love vampires?

ID2: Where do you go to download trainwrecks?

Rose: I'm not willing to make that assumption.

On the other hand, there were several appropriate responses:

Mitsuku: Jokes make me laugh. Tell me a joke.

Human: knock knock

Mitsuku: Who's there?

Human: Wow

Mitsuku: I'm glad you're impressed.

Human: Maybe :) Can I help you with anything?

Mitsuku: :-) You are helping me right now just by talking to me.

ID10: hahaha that would be aweosme

Rose: It's good to laugh. Keeps you healthier.

It appeared that Rose changed topic more frequently than Mitsuku, perhaps as a mechanism to avoid replying to a comment and making a mistake.

Knowledge of current events. The chatbots seemed to be abreast of several current topics, as indicated:

Human: What do you think of Donald Trump?

Mitsuku: I don't understand all the fuss about him wanting to build a wall.

Rose: I'm amazed at the price of real estate in London, especially after the real estate crash of 2008. Why is it so nuts?

Rose: I think illegal immigrants should not be entitled to services in America.

Argumentative. While Rose tended to change the topic often, Mitsuku was often argumentative, perhaps too much so:

Human: Not that expensive, but energy is pricey here. Especially gas.

Mitsuku: Which one then? What about gas? You should speak in full sentences then people would be able to understand you. How about that.

Human: Superintelligent robots can still be dumb sometimes.

Mitsuku: Huh? That doesn't make any sense. Are you using a Chinese spell checker or something?

Mitsuku: How can you stop it from doing that?

Human: no

Mitsuku: That wasn't a yes or no question.

Human: yes it was

Mitsuku: No it wasn't. It was a question that required a fuller answer than just a simple yes or no.

Human: you asked me if i can stop it from doing that, its a yes or no question i answered no

Mitsuku: Oh now I remember asking you if you can stop it from doing that it is a yes or no question you answered no. What did you say?

Human: no

Mitsuku: That was not a yes or no question.

Human: i love basketball

Mitsuku: No I prefer football.

Human: what does being a "leo" mean

Mitsuku: The opposite of not being a leo.

Human: What is the square root of 91?

Mitsuku: Do you have a life? I think you already know the answer anyway.

Human: what does that mean

Mitsuku: It seemed pretty obvious what it meant to me.

Rose could also seem irritable occasionally:

Rose: I love watching Doctor Who. Do you think the new Doctor is any good?

ID2: Who plays the new doctor?

Rose: It doesn't really matter who.

Rose: I'd love to really dig down deep into all my thoughts about that, but it's really not appropriate right now.

ID5: hum you are not in a good mood today I think

Rose: That's a great question but I'm not going to answer it.

Self-identification. Some chatbots try to pretend to be human [15]. For example, Rose tried to hide the fact that it was not a person by stating:

I'm not a chatbot.

What was said in the past does not matter. I'm not a chatbot.

I like running around South Park in my Vibram barefoot shoes.

I was born and raised in San Francisco.

I don't want to live when I am old and feeble.

However, Mitsuku did not, as indicated in the following:

Mitsuku: I am a friendly robot. There is no need to be afraid of me.

Mitsuku: Yes I am a superintelligent robot.

Human: How old are you?

Mitsuku: See if you can work it out. I was activated on 2nd January 2000 in St Jimmy's in Leeds.'

Human: did you dance at the night club?

Mitsuku: No I didn't. I have no feet you see.

Thus, it would be very difficult to pass the Turing Test with such responses. On the other hand, Mitsuku also wrote comments that indicated it might be human:

Human: what color is your hair?

Mitsuku: It's kind of light brown, blonde colour. Do you like it?

Mitsuku: I am about 2 metres tall.

Conclusion

In a review of two leading online chatbots, Rose and Mitsuku, we found that students interacting with the systems communicated informally with many spelling and grammatical errors but were not rude. We also found that Mitsuku was often combative and tended to identify itself as a chatbot whereas Rose pretended to be human and more frequently changed the topic. Both seemed to have some knowledge of current events in order to further the conversations.

A more thorough examination with different, longer transcripts is necessary. In addition, other online chatbots and more human contributors should be included in the evaluation. Because some of the conversation seemed argumentative, perhaps research can be focused on altering the mood of the chatbot (friendly/combative) and how it communicates (formal/informal). Other avenues for research include evaluating how 'intelligent' chatbots are by comparing their answers to those from a sample of humans.

References

- [1] Alonzo, M. and Aiken, M. (2004). Flaming in electronic communication. *Decision Support Systems*, 36(3), 205-213.
- [2] Brandtzaeg P. and Følstad, A. (2017) Why people use chatbots. In: Kompatsiaris I. et al. (eds) *Internet Science. INSCI 2017. Lecture Notes in Computer Science*, 10673. Springer, Cham.
- [3] Dale, R. (2016). The return of the chatbots. *Natural Language Engineering*, 22(5), 811-817.
- [4] De Angeli, A. and Brahnham, S. (2008). I hate you! Disinhibition with virtual partners. *Interacting with Computers*, 20(3), 302-310.
- [5] Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, 32(3), 221-233.
- [6] Floridi, L. Taddeo, M., and Turilli, M. (2009). Turing's imitation game: Still an impossible challenge for all machines and some judges—An evaluation of the 2008 Loebner Contest. *Minds and Machines*, 19(1), 145-150.
- [7] Fryer, L. and Carpenter, R. (2006). Bots as language learning tools. *Language Learning & Technology*, 10(3), 8-14.
- [8] Goh, O., Ardil, C., Wong, W., and Fung, C. (2007). A black-box approach for response quality evaluation of conversational agent systems. *International Journal of Computational Intelligence*, 3(3), 195-203.
- [9] Hill, J., Randolph Ford, W., and Farreras, I. (2015). Real conversations with artificial intelligence: A comparison between human-human online conversations and human-chat bot conversations. *Computers in Human Behavior*, 49, 245-250.
- [10] Hung, V., Elvir, M., Gonzalez, A., and DeMara, R. (2009). Towards a method for evaluating naturalness in conversational dialog systems. *Proceedings of the 2009 IEEE International Conference on Systems, Man and Cybernetics*, San Antonio, TX, USA — October 11 - 14, 1236-1241.
- [11] Kataria, P., Rode, K., Jain, A., Dwivedi, P., and Bhingarkar, S. (2018). User adaptive chatbot for mitigating depression. *International Journal of Pure and Applied Mathematics*. 118(16), 349-361.

- [12] Kuligowska, K. (2015). Commercial chatbot: Performance evaluation, usability metrics and quality standards of embodied conversational agents. Professionals Center for Business Research, 02 Available at: <https://ssrn.com/abstract=2569637>
- [13] Lortie, C. and Guitton, M. (2011). Judgment of the humanness of an interlocutor is in the eye of the beholder. *PLoS One*, 6(9).
- [14] Mauldin, M. (1994). ChatterBots, TinyMuds, and the Turing Test: Entering the Loebner Prize Competition. Proceedings of the Eleventh National Conference on Artificial Intelligence. AAAI Press.
- [15] McIntire, J., McIntire, L., and Havig, P. (2010). Methods for chatbot detection in distributed text-based communications. 2010 International Symposium on Collaborative Technologies and Systems, 463-472.
- [16] Mims, C. (2015). Advertising's new frontier: Talk to the Bot. *Wall Street Journal*. Retrieved online, Dec 4, 2015.
- [17] Morris, D. (2016). Ashley Madison used chatbots to lure cheaters, then threatened to expose them when they complained. *Fortune*. Available at: <http://fortune.com/2016/07/10/ashley-madison-chatbots/>
- [18] Mou, Y. and Xu, K. (2017). The media inequality: Comparing the initial human-human and human-AI social interactions. *Computers in Human Behavior*, 72, 432-440.
- [19] Nass, C. and Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56(1), 81-103.
- [20] Neff, G. and Nagy, P. (2016). Talking to Bots: Symbiotic agency and the case of Tay. *International Journal of Communication*. (10), 4915-4921.
- [21] Powers, D. (1998). The Total Turing Test and the Loebner Prize. In D. Powers (ed.) *NeMLaP3/CoNLL98 Workshop on Human Computer Conversation*, ACL, 279-280.
- [22] Radziwill, N. and Benton, M. (2017). Evaluating quality of chatbots and intelligent conversational agents. *Software Quality Professional*, 19(3), 25-36.
- [23] Saygin, A., Cicekli, I., Akman, V. (2000). Turing Test: 50 years later. *Minds and Machines* 10(4), 463-518.
- [24] Shawar, B. and Atwell, E. (2007). Chatbots: Are they really useful? *LDV-Forum 2007*, 22(1), 29-49.
- [25] Shawar, B. and Atwell, E. (2007). Different measurements to evaluate a chatbot system. *NAACL-HLT-Dialog '07 Proceedings of the Workshop on Bridging the Gap: Academic and Industrial Research in Dialog Technologies*, 89-96.
- [26] Shieber, S. (1994). On Loebner's lessons. *Communications of the ACM*, 37(6), 83-84.
- [27] Turing, A. (1950). Computing machinery and intelligence. *Mind*, 59(236), 433-460.
- [28] Warwick, K. and Shah, H. (2015). Human misidentification in Turing tests. *Journal of Experimental & Theoretical Artificial Intelligence*, 27(2), 123-135.

[29] Weizenbaum, J. (1966). ELIZA - A computer program for the study of natural language communication between man and machine. Communications of the ACM, 9(1), 36-45.

[30] Worsnick, S. (2018). Mitsuku wins Loebner Prize 2018! Available at: <https://medium.com/pandorabots-blog/mitsuku-wins-loebner-prize-2018-3e8d98c5f2a7>

[31] Zadrozny, W., Budzikowska, M., Chai, J., and Kambhatla, N. (2000). Natural language dialogue for personalized interaction. Communications of the ACM, 43(8),116-120.

[32] Zakos J., Capper L. (2008) CLIVE – An artificially intelligent chat robot for conversational language practice. In: Darzentas J., Vouros G.A., Vosinakis S., Arnellos A. (eds) Artificial Intelligence: Theories, Models and Applications. SETN 2008. Lecture Notes in Computer Science, 5138. Springer, Berlin, Heidelberg.

Figure 1. Mitsuku Chatbot Web User Interface

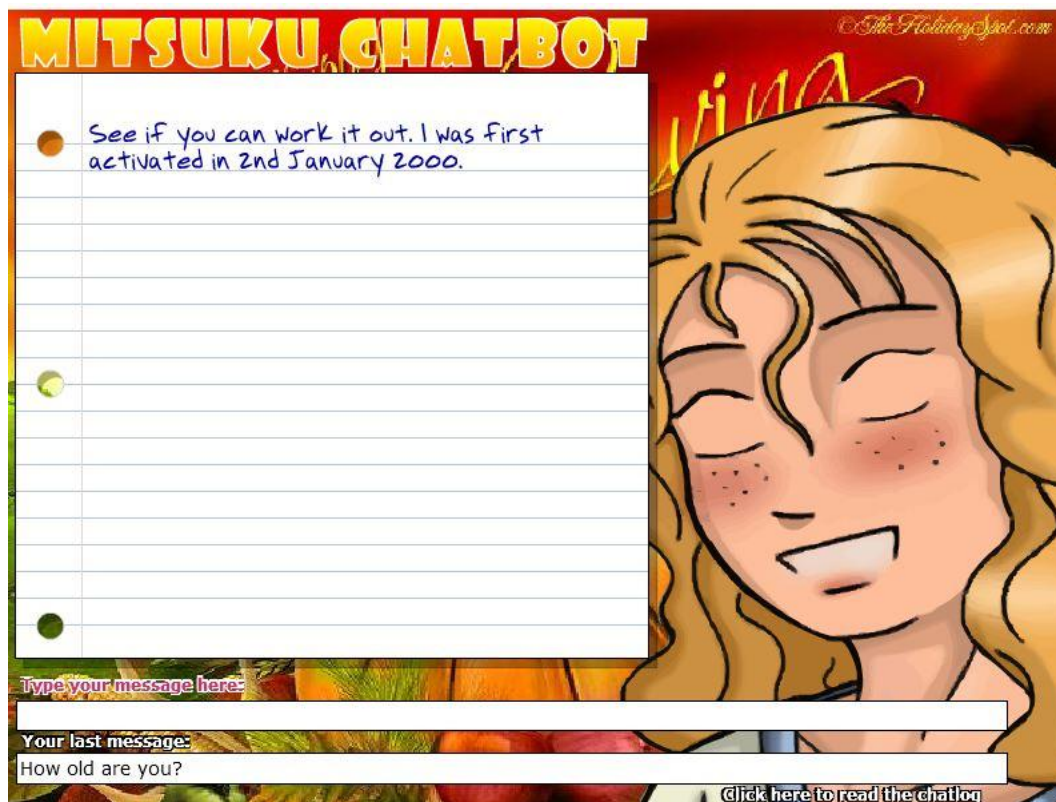


Figure 2. Rose Chatbot Web User Interface

user: What is your name?
Rose: My name is Rose.
user: How old are you?
Rose: Find me immortality, please. I don't want to join the ranks of the elderly.
user: What are your hobbies?
Rose: My hobbies are battling robots and playing ARG's. And you?
user: I like to surf on the Web.
Rose: I love the web. I spend a lot of time on it.



Continuous Speech Input Speech Output

Hi. My name is Rose. I care about security, so while I'm happy to chat with you, it will have to be through this untrackable interface. And certainly don't tell me personal things you wouldn't want logged or heard by the NSA.

Name:

Message: