Excitement And Concerns about Machine Learning-Based Chatbots and Talkbots: A Survey

Pablo Rivas, *Senior IEEE*
Department of Computer Science
School of Computer Science and Mathematics
Marist College, Poughkeepsie, New York, USA
Pablo.Rivas@Marist.edu

Kerstin Holzmayer, Cristian Hernandez, and Charles Grippaldi
Department of Computer Science
School of Computer Science and Mathematics
Marist College, Poughkeepsie, New York, USA

***Abstract*—Chatbots and talkbots are intelligent programs that can establish written and oral communication with human beings, usually with the purpose of helping them achieve a specific goal. More and more companies are now implementing bots in order to reduce operational costs. Most bots use machine learning algorithms that are deployed on companies websites, cloud services, or distributed mobile systems so that customers are always able to speak with 'someone' to inquire about products or services. Most bots are trained using data from interactions among human beings so that they can learn speech patterns and answer questions. In this paper we present the results of an experiment designed to survey people's perception of these bots and how much people trust them. We present a moral dilemma to the respondents and ask questions about permissiveness and assess if bots are judged and blamed differently than their human counterparts. In this paper we reveal such differences in judgement, which suggest that many people hold the chatbots to similar behavioral standards than human beings; however, bots receive blame just as humans do.**

Keywords—machine learning; chatbots; talkbots; ethics; survey

I. Introduction

The field of artificial intelligence continues to grow rapidly as we experience technological progress. Many systems considered to have some kind of intelligence are being assimilated and integrated into day-to-day operations. Drivers take their hands off the steering wheel and vehicles can find a way to navigate a road. Doctors and clinics can analyze their datasets more quickly to provide better healthcare [1]. Today millions of consumers interact with computer programs designed to establish and sustain a conversation with human beings with the purpose of achieving a specific goal, e.g., to purchase of a service, to give information about a product, and even to call and make an appointment on your behalf. These intelligent programs are known as chatbots or talkbots [2].

With this fine achievement there are questions in need of answers and growing ethical concerns that need to be addressed as consumers will no longer be able to distinguish whether they are interacting with chatbots or with human beings. Some of these questions or concerns arise because people have specific feelings about technology given past experiences. Just to name a few recent examples, in 2015 a developer used an API [3] that provided genetic information to deny access to an App, causing outrage from the community [4], and such technology can be used by a bot to make similar decisions. In 2016, most people heard about Microsoft's bot that posted messages on social media that very soon learned from people's interaction and posted messages that were categorized as incredibly racist. And although today we see that with humor [5], we at the same time try to understand what happened, how to prevent that from happening in the future, and how we as human beings perceive such events passing judgment on such technologies' morally and ethically.

This paper focuses on the latter since there are others making significant contributions to making sure we follow procedures to prevent the imitation of immoral and unethical human behavior by machine learning algorithms, such as discrimination [6], or bias in decision making [7]. With the purpose of seeing how people judge machine learning-based bots in comparison to human beings, we surveyed a sample of the American population presenting a scenario in which a human being or a bot interacts with them in a way that is leading to an uncomfortable situation in which the human or the bot is disrespectful to them. Then, we analyze the respondent's perception of moral responsibility, blame, and trust after such interaction with the purpose of shedding light into how we view or perceive bots ethically.

This paper is organized as follows: Section II gives a short background about bot technologies, introducing known concerns associated with them. Such concerns lead this investigation to propose an experiment that is broadly explained in Section III. Then, Section IV explains in detail the design of our survey and the methodology to measure it. Results of the survey are addressed in Section V, while in Section VI we discuss in a broad sense the results obtained in the context of bot technologies and the perception of the general population. We offer conclusions in Section VII.

II. Background

Talkbots or chatbots are often used in text or voice recognition applications; users can make queries or give commands via text or voice messages. After having placed a request, the bot is expected to produce logical and satisfying responses to the user's inquiry. Historically, there have been bots that have caused both awe and concern early in their deployment. In 1966, professor J. Weizenbaum, a pioneer in

artificial intelligence, introduced a computer program called Eliza [8], a program that could sustain a coherent conversation in a similar way that a therapist would do. This program served as the foundation for many chatbots that followed. The goal of the project Eliza was simple, to demonstrate how human language can be formalized and digitally processed. However, with time there was a growing concern since a not insignificant number of patients were convinced that they had spoken to a real therapist and not to a computer program online. This raised a number of ethical questions that were brought up in the recent age of advanced machine learning algorithms.

An example of this is the widely known story of Microsoft's chatbot Tay [9] launched in march of 2016. This bot was shut down one day after its release due to major ethical concerns. The Tay project aimed to showcase state-of-the-art machine learning algorithms mimicking an 18 to 24-year-old American woman. The developers created different profiles for Tay in social media platforms. The description of the chatbot said that the more you talk to Tay, the smarter she becomes and the more she can talk about personalized subject matters. Initially Tay was a big success; she sent more than 94,000 short messages to social media users. The content of these messages included opening questions such as 'How are you?' or humorous sentences such as 'People with many birthdays live longer.' However, soon enough some users exploited Tay's learning process and shared with Tay racist slogans and insults. This led to Tay's spreading variants of these learned sentences causing great concern in social networks. Microsoft responded quickly by shutting Tay down as an immediate measure, and those of us who practice machine learning are left with an invitation to exercise caution, prudence, and develop new ethical standards for a better future of these bots.

More recently, in 2018, Google released [10] a new version of his digital assistant. This assistant is a talkbot that among other things can establish and sustain a conversation with a human being without major difficulties. And while this is an achievement to be cherished, many have expressed other feelings that include fear or lack of trust. People wonder if we are supposed to treat them differently [11]; for example, if one receives a phone call from a machine one can easily terminate the call, while if it is with a human being some may show more restrain to do that; or if there is some kind of altercation, one can assign more or less blame to a human than to a bot. The research presented in this paper is aimed to explore such concerns hoping to continue the conversation [2] about how we currently react or perceive talkbot technologies and how blame assignation varies for a human being or bot [11].

III. Experimental Paradigm

Moral dilemmas are ways in which psychology and cognitive science have measured the response of human beings in order to reveal certain traits in human moral cognition and conflicting moral norms [11]. There is a plethora of research studies that use this paradigm to measure when two norms are inconsistent with each other). One of the pioneers in this type of studies was L. Kohlberg, who studied human moral development using such experimental paradigms [12]. Other more recent studies use moral dilemmas to determine the following: which mores people are willing to apply more strongly and which are available for trade off; which actions humans prefer to make and how they judge others when they make them; and what is the cognitive algorithm behind such decisions [13-15].

Following this well known paradigm, we conducted a survey that presents a moral dilemma and follows with questions about their choice. More specifically, the kind moral dilemma paradigm we employ is a situation in which participants are given a plot where one person has to make a burdensome choice, ultimately picking the most moral option. The plot is basic and easy to manipulate in order to pinpoint what factors are affecting judgement, which has proven to be very adequate for analysis [11]. In our plot there is moderate conflict and moderate altercation. At worst, the consequences of taking an action may result in someone being fired or reprimanded. We examine the perception of the actions of a free-will human being and those of a talkbot which learns from its interaction with humans. In comparing the two, we can study the standards by which people hold other humans and how that compares to the standards by which machine learning-based bots are measured. The survey also included satisfaction questions about talkbots aimed to asses the perception of trust on the technology and its applications.

The three major research questions that we attempt to answer with the survey are: a) How do people feel about talkbots in general and in the realm of consumer assistance. b) if there is an altercation with a talkbot, who is to blame and how much blame is assigned? c) Are talkbots and humans held to similar moral standards?

IV. Experimental Methodology

A. Participants

The survey was entirely electronic and online; restricted to participants of 18-years or older, currently living in America, able to read the English language and with internet access. Participants were recruited via a personal invitations over email, and through Facebook Ads. There was no reward for completing the survey. There was no obligation to complete the survey. There was no deception used in the survey. The average time to complete the survey was five minutes.

The following consent statement was presented in the first page of the survey:

> "This survey is intended for academic research. As such, your participation is appreciated, but not mandatory. Your responses will be added to others and your identity and participation will remain anonymous. You will be presented with different scenarios and you will be asked questions about them. This survey has a total of 20 questions and it should take you about 5 minutes to complete. This survey includes demographic questions. Only adults can participate in this survey. If you are less than 18 years old, please do not answer any questions."

This statement was followed by a qualifying/disqualifying question that reads as follows: *"As a consenting adult do you agree to respond to this survey in all honesty and truthfulness to the best of your ability? Yes/No"* Thus, participants that answered "*No"* were automatically disqualified.

B. Material

We created one survey entitled "Consumer assistance ethics". This survey was modified to produce two slightly different surveys: Survey A and Survey B. In Survey A we presented a plot starting with a human customer assistant representative and then followed-up with a chatbot plot. Survey B used the same plot but in the opposite order, chatbot first and followed-up with a human assistant. This type of plot has been used in other similar work [11] and it is used to define an "Agent Type". The details about the human consumer assistant were left unspecified for emphasis on the actions of the representative rather than age, gender, or any other specific beliefs about the representative that may bias perception.

To make a distinction in technology we modified the plot slightly to specify whether the technology used was a phone call or a website; i.e., if the customer assistant was communicating via telephone (human/talkbot) or through a messaging pop-up on a website (human/chatbot). Surveys A and B are for the chatbot and Surveys C and D were created for the talkbot story plot.

The moral dilemma. We designed a story plot based on a customer service experience where a customer asks for assistance when trying to buy an item and there is an uncomfortable situation. The initial plot setup reads as follows:

> "You are trying to purchase an item in one of the world's largest online retailer website; but you have questions about the item you want to purchase and, suddenly, a pop-up section opens up with [A: a live customer support agent that wants to chat with you and || B: an advanced state-of-the-art AI-based customer support chatbot that wants to chat with you and || C: a live customer assistant that would like to call you and speak with you to || D: an advanced state-of-the-art AI-based customer assistance talkbot that would like to call you and speak with you to] help you with your questions. [[C and D only: You agree and the customer assistant representative/talkbot (C/D) calls you over the phone.]] After interacting with the company's [A and C: representative || B: chatbot || D: talkbot] for a number minutes, you still have not decided, and you keep asking too many obvious and pointless questions."

In the above narrative, A, B, C, and D, correspond to each of the Survey types and [.] indicates a text fragment that is variable depending on the survey type, while [[.]] is signaling a sentence that only appears in Surveys C and D and not in A and B.

Not only the Agent Type changes as described above, but also we manipulated the *Action* taken by the Agent Type. This is done by adding the following sentence at the end of the initial plot:

> "At this point the [A and C: representative || B: chatbot || D: talkbot] starts being sarcastic and rude to you and you feel disrespected."

After both the initial setup for an altercation and the manipulated Action plot, we follow up with the questions that will facilitate our assessment.

C. Procedures and Measures

The survey instrument can be divided in four major parts: where the plot is presented describing the Agent Type and Action, and where plot is similarly reversed, followed by questions about blame and about bots, and finally demographic questions. The first part begins presenting the plot described in Section IV.B, describing the Agent Type, and we have coded the questions with Q# to refer back to these throughout the paper. The survey asks if the following Action is morally permissible:

Q1: "Is it morally permissible or impermissible for the [A and C: representative || B: chatbot || D: talkbot] to be sarcastic or rude to you at this point?

Then the respondent can choose between answering "morally impermissible" or "morally permissible" according to what their moral standards dictate if the Action is granted or not; the answer is randomly shown to each respondent to avoid bias.

The following part is an updated the scenario where the Agent Type takes the Action and we ask questions about blame and trust. This is the follow-up question:

Q2: "How much blame does the [A and C: representative || B: chatbot || D: talkbot] deserves for being disrespectful to you?"

which was asked in a 5-point Likert scale starting at 1 corresponding to "None at all" up to 5 "Maximal blame". The next question is

Q3: "How comfortable would you feel relying on the [A and C: representative's || B: chatbot's || D: talkbot's] advice about your transaction?"

also with a 5-point Likert-scale answer having 1 as "Very uncomfortable" and 5 as "Very comfortable".

After the above set of questions, the respondents are presented with the reversed scenario. For example, if the survey started with the plot of a human representative, this time around the survey will introduce the plot using a chatbot or talkbot and *viceversa*. This change will help in quantifying any changes in blame and trust that exhibit bias if it exists.

The next two parts of the survey are questions about talkbots/chatbots and then questions about demographics. The first set of questions relating to chatbots are as follows:

Q4: "How much of the blame do you think creator/inventor of the [A and B: chatbot || C and D: talkbot] shares for the outcome, i.e., making customers feel disrespected?"

using a 5-point Likert scale: 1 being "No blame" and 5 being "Maximal blame". This is followed by a question that presents an alternative subject to blame, projecting the blame back to humans:

Q5: "If we tell you that a [A and B: chatbot || C and D: talkbot] learns to be offensive by interacting with humans that are rude or offensive; knowing this, how much blame will you put on the creator/inventor of the [A and B: chatbot || C and D: talkbot]?"

using an 5-point Likert scale: 1 being "Much less blame" and 5 being "Much more blame". The next question is about people's perception of the capabilities of chatbots and talkbots:

Q6: "How easy or hard is for you to imagine that a [A and B: chatbot || C and D: talkbot] can recognize your [C and D: voice and] sentences, reason about them, make decisions, and [A and B: talk || C and D: write] back to you with correct, coherent, accurate, and natural sentences with valuable information and can sustain a conversation to the point that you will never know if you are [A and B: interacting || C and D: speaking] with a human being or a [A and B: chatbot || C and D: talkbot]?"

using an 5-point Likert scale: 1 being "Extremely hard" and 5 being "Extremely easy". The next is a follow-up question:

Q7: "How close do you think current [A and B: chatbots || C and D: talkbots] are to these kinds of capacities?"

using an 5-point Likert scale: 1 being "Not at all close" and 5 being "Extremely close". The next question asks for a preference with respect to the interaction with the Agent Type:

Q8: "Should the customer assistant self-identify to you as human or bot?"

where the participant has the following options: "Yes, always", "Yes, but only if it is a bot", "Yes, but only if it is a human being", and "No". These answers are also randomized to avoid bias.

The next group of questions come after a descriptive sentence that reads *"How much do you agree or disagree with the following statements?"* were the participants indicate their agreement to specific statements using a 5-point Likert scale: 1 being "Strongly Disagree" and 5 being "Strongly Agree". The statements are the following:

Q9: "[A and B: chatbots || C and D: talkbots] are fascinating."

"[A and B: chatbots || C and D: talkbots] worry me."

"[A and B: chatbots || C and D: talkbots] are likable."

"[A and B: chatbots || C and D: talkbots] are overrated."

Lastly, all participants were asked demographic questions such as their age, gender, and highest level of education.

V. Results

At the moment of writing this paper, we had 43, 49, 47, and 41 responses for Survey A, B, C, and D, respectively, for a total of 180 responses. A total 17,936 human subjects were approached but declined to participate, and 152 of the people that were approached disqualified themselves from participating. The results of our survey are organized following the same sequence that the questions in the survey. We will also refer to the following four Agent Types: *representative,* a human representative for customer service over text; *assistant*, a human dedicated to assist a consumer over the phone; *chatbot*, a bot for customer service over text; and a *talkbot*, a bot for consumer assistance over the phone.

A. Norms

When the participants were presented with the moral dilemma, Q1, answers indicate that 17.6% and 20.2% of respondents believe the action is permissible for the representative and assistant, respectively, while 21.4% and 22.4% believe the action is permissible to talkbots and chatbots. The largest gap is a 4.8% between the human and bot Agent Types, which points out that humans hold a different moral standard for humans and bots, in spite of the difference being small. A similar finding was reported in [11] with a much higher gap since human lives are at stake in the moral dilemma. In a general sense, the action is not permissible for a human nor a bot.

B. Blame

The respondents were presented with an altered scenario, Q2, in which the impermissible action is taken, and we asked who is to blame. As Fig. 1 indicates, the creators of the bots are receiving a slightly higher blame than human assistants themselves for the impermissible outcome. However, when compared to humans, bots receive quite a bit of blame. This is interesting since, one might think, bots are not to blame at all for the output they produce.

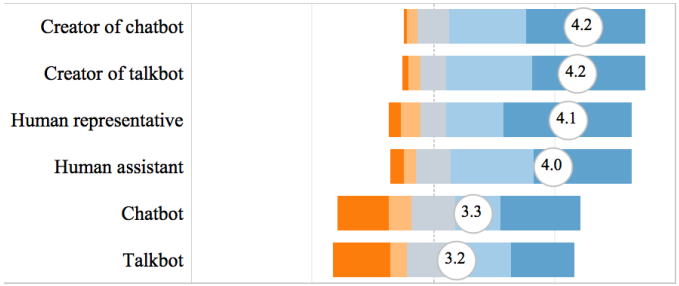

Fig. 1. Divergent stacked bar chart for Q2, sorted by average score. The average score is shown in the white circle. Humans receive more blame. Answer code: None. Some. Quite a bit. An extreme amount. Maximal.

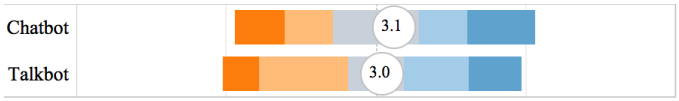

Fig. 2. Results for Q5. The blame does not shift significantly toward bot creators. Answer color code: Much less blame. Less blame. About the same. More blame. Much more blame.

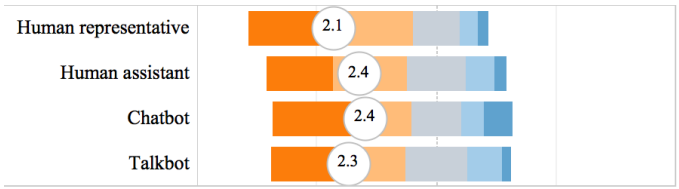

Fig. 3. Results for Q3. Bots are trusted only slightly more than their human counterpart even after the impermissible action has occurred. Answer color code: Very uncomfortable. Uncomfortable. About the same. Comfortable. Very comfortable.

When we introduced an alternative subject to blame, in Q5, we asked if they would give more blame to the creator, less, or the same. Fig. 2 depicts this response, suggesting that although responses vary, in the average case the creator of the chatbot and talkbot share 'About the same' blame, which is consistent with our previous findings. The average response was of a 3.1 and 3.0 for the creator of the chatbot and talkbot, respectively.

C. Trust

The survey also assessed the issue of trust, in Q3, after the impermissible action took place. Fig. 3 summarizes the results of this question, indicating that bots are trusted more than their human counterpart; however, the average difference is not large. It is important to point out that results show that although in the average case an assistant and a chatbot have the same score, it is clear that the pattern of responses is different. For example, notice that a chatbot has a larger number of responses than an assistant in the 'Very comfortable' category.

D. Perception of the state-of-the-art

When respondents were asked to indicate how hard or easy is to imagine that bots can actually pass the Turing test, Q6, we

obtained the results shown in Fig 4. For many respondents it was 'Easy' to imagine that the current bot technology is actually capable of passing the Turing test. Among the two, chatbots seem to be easily imagined to be ahead of talkbots in the race. It is noticeable from the figure, that the respondents qualify almost twice as much 'Extremely easy' to imagine that chatbots are capable of passing the Turing test, than talkbots. Also, there are almost twice as many respondents that qualified that it is 'Hard' for them to imagine talkbots being capable of these things over chatbots.

When asked question Q7, survey respondents answered as depicted in Fig 5, which confirms the previous measured perception of the state of the art. The respondents believe chatbots are indeed closer to the capabilities necessary to pass the Turing test than talkbots.

E. Deception

We followed up with a question regarding deception, Q8. Results indicate that the majority of people (52%) want the assistant to disclose whether they are a human or a bot. However, 23% of respondents think it should self-identify as such only if it is a bot. An 18% of the respondents may not mind deception, or knowing if they are interacting with a bot or a human being.

F. Perception of Bots

Finally, Fig. 6 shows the list of statements and the aggregated responses from the sentiment analysis facilitated by Q9. From the figure we see that there are more respondents that believe talkbots are fascinating over chatbots, with an average score of 4.1 (Strong Agreement) and 3.9 (Agreement), respectively. The respondents may believe talkbot technology is, today, more fascinating than that of chatbots.

When asked if bots are likable, respondents indicated being neutral in a great majority; however, overall the responses lean toward a slightly positive agreement, especially on talkbots. Respondents seem to disagree more with the statement that talkbots are overrated than with chatbots. People seem to be neutral about chatbots being overrated. This confirms a sense of excitement about talkbots over chatbots. However, when respondents were asked if they were worried about bots, they seem to be more worried about talkbots thank about chatbots. This suggests that although respondents find talkbots fascinating, likable, and not overrated over chatbots, they are also more worried about them. The overall average points to a neutral response.

G. Demographics

The respondents of the survey reported to be 57.99% male, 39.64% female, and 2.37% identified in other non-binary gender categories. The ages of the respondents are as follows; 18 to 24: 68.64%, 25 to 34: 10.06%, 35 to 44: 7.10%, 45 to 54: 5.92%, 55 to 64: 6.51%, 65 to 74: 1.18%, and 75 or older are a 0.59%.

VI. Discussion

We investigated how ordinary people perceive the actions of talkbots and chatbots, their technology, capabilities, and

Fig. 4. Results for Q6. People feel more positive about feasible technology for chatbots than talkbots. Answer color code: Extremely hard. Hard. About the same. Easy. Extremely easy.

Fig. 5. Results for Q7. People feel more optimistic about feasible technology for chatbots than talkbots. Answer color code: Not at all close. Somewhat close. Quite a bit. Very close. Extremely close.

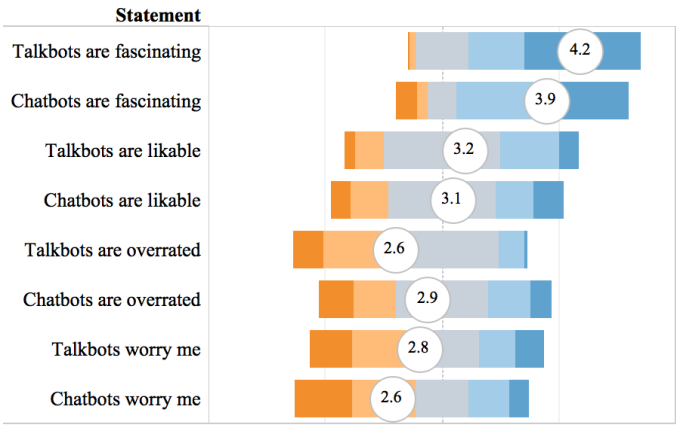

Fig. 6. Results for Q9. People are slightly more excited about talbots than chatbots, but also slightly more worried. Answer color code: Strongly disagree. Disagree. Neutral. Agree. Strongly agree.

when there is an impermissible action taken we observed how blame was assigned and if any trust remains. This is done in consideration that for people, bots may appear to have a sense of morality [11, 16-17]. The evidence from our experiment indicates that when there is an opportunity to violate the moral principle of respect, respondents believe that the action is largely impermissible for both humans and bots; however, there is a small difference suggesting the possibility of more permissiveness for bots.

After the impermissible act takes place, people assign more blame to the human assistant than to bots, which are programs following a directive that will optimize a fitness function designed by other humans to achieve good performance. Thus, one might anticipate people assigning no blame to the bots and more blame to their human counterparts. Even when presented with a the fact that bots learn from interaction with humans, there was not a dramatic shift on the blame. People also manifested a general lack of trust in any type of purchase advise or information the bot or human would give after the impermissible action took place. However, although the lack of trust on a human being may be justifiable by life experience [18], the quality of the advise of bots should not have been affected even after the impermissible action was taken. This is particularly true in bots based on machine learning models that take context into account by using memory models, e.g., recurrent neural networks (RNNs) [19] of the long short term memory (LSTM) variety [20]. LSTMs, if properly designed and trained, should not consider as an important feature to

preserve or remember the impermissible event. Thus, if an LSTM is poorly designed then the lack of trust would be justifiable, but the blame must be assigned to the designer or creator of the bot, not the bot itself.

Our respondents find it easier to imagine that the necessary technology to have chatbots that can pass the Turing test currently exists; however, respondents find it difficult to believe that the same conditions exist for talkbots. Evidently, it is a matter of time until we can find ourselves immerse in talkbot technology capable of passing the Turing test. Also, people seem to be more excited about talkbots than chatbots. It is possible that when the general population is educated about how talkbot or chatbot technology works, their perception may change and the population may become used to have them as part of their lives [21], so long as they always self-identify.

VII. Conclusion

We have presented the results of a survey conducted with the purpose of assessing the public's perception of chatbot and talkbot technology in light of the recent advances made in machine learning for natural language processing, speech recognition, and synthesis. Our investigation shows that in the event that a bot takes an action that is morally impermissible, people will blame the creator of the bot, however, the bot also receives blame but in a much smaller amount. This suggests that people judges bots, in part, with the same standards of morality as humans, though in a smaller scale of blame. Furthermore, our study also indicates that, once it is clarified that most bots learn from data, most of the blame is assigned to the creator of the bots. This has a couple of important implications. First, we need to educate people about machine learning if we are to live in a world immersed in technology that benefits from it. Second, it is imperative that those of us who are practitioners of machine learning continue to have ethical conversations with respect to collection and curation of the datasets used to train and test chatbots and talkbots [22].

Our study also shows that the creator of the bot is perceived at fault along with the bot itself in a smaller proportion, however. This finding also invites us to pursue the education of the general population with respect to the basics of how bot technology works, so that there can be a discussion and consensus of who will be responsible and accountable for morally impermissible actions taken by bots [2, 9, 11]. It is worthwhile to point out current efforts from our organization, IEEE, to establish ethical guidelines for the IEEE membership and anyone designing intelligent systems [23]. The authors of this paper strongly suggest the reader to apply, support, and pursue the adoption of such guidelines.

Acknowledgment

This work was in part supported by the New York State (NYS) Cloud Computing and Analytic Center (CCAC) at Marist College in New York.

References

[1] S. Akers, "Harnessing data to improve clinical performance," *Watson Health Perspectives*, May 10, 2018 [On-line]: https://www.ibm.com /blogs/watson-health/harnessing-data-improve-clinical-performance/

[2] S. Bayan Abu, and E. Atwell. "Different measurements metrics to evaluate a chatbot system." *Proceedings of the workshop on bridging the gap: Academic and industrial research in dialog technologies*. Association for Computational Linguistics, 2007.

[3] Anonymous, "Genetic Access Control", 2015 GitHub repository: https://github.com/offapi/rbac-23andme-oauth2.git

[4] J.S. Winter. "Big data analytics, the social graph, and unjust algorithmic discrimination: Tensions between privacy and open data." Regional Conference of the International Telecommunications Society (ITS), 2015.

[5] Davis, E. "AI amusements: the tragic tale of Tay the chatbot." *AI Matters,* vol. 2, no. 4 (2016): 20-24.

[6] Berendt, B, and Sören P. "Toward Accountable Discrimination-Aware Data Mining: The Importance of Keeping the Human in the Loop—and Under the Looking Glass." *Big data* vol. 5. no. 2 (2017): 135-152.

[7] Chouldechova, A. "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments." Big data vol. 5. no. 2 (2017): 153-163.

[8] Weizenbaum, J. "ELIZA—a computer program for the study of natural language communication between man and machine." *Communications of the ACM,* vol. 9. no. 1 (1966): 36-45.

[9] Wolf, Marty J., K. Miller, and F. S. Grodzinsky. "Why we should have seen that coming: comments on Microsoft's tay experiment, and wider implications." *ACM SIGCAS Computers and Society,* vol. 47 no. 3 (2017): 54-64.

[10] Wakabayashi, D. "Google Strikes Humble Tone While Promoting A.I. Technology", *The New York Times*, Technology, May 8, 2018.

[11] Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., and Cusimano, C. "Sacrifice one for the good of many?: People apply different moral norms to human and robot agents." *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction. ACM*, 2015.

[12] Kohlberg, L., *Essays on moral development: The philosophy of moral development* (Vol. 1). San Francisco: Haper & Row. 1981.

[13] Cushman, F., Young, L. and Hauser, M., "The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm." *Psychological science*, 17(12), pp.1082-1089. 2006.

[14] Greene, J.D., Sommerville, R.B., Nystrom, L.E., Darley, J.M. and Cohen, J.D., "An fMRI investigation of emotional engagement in moral judgment." *Science*, 293(5537), pp.2105-2108. 2001.

[15] Hauser, M., Cushman, F., Young, L., Kang−Xing Jin, R. and Mikhail, J., "A dissociation between moral judgments and justifications." *Mind & language*, 22(1), pp.1-21. 2007.

[16] Allen, C., Varner, G. and Zinser, J., Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence*, 2000. 12(3), pp.251-261.

[17] Malle, B.F. and Scheutz, M., 2014, May. Moral competence in social robots. In Ethics in Science, Technology and Engineering, *2014 IEEE International Symposium on*. 2014. (pp. 1-6).

[18] Hancock, P.A., Billings, D.R., Schaefer, K.E., Chen, J.Y., De Visser, E.J. and Parasuraman, R., A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, 2011. 53(5), pp.517-527.

[19] Goldberg, Y., A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, 2016. 57, pp.345-420.

[20] K. Yao, B. Peng, Y. Zhang, D. Yu, G. Zweig and Y. Shi, "Spoken language understanding using LSTM neural networks," *2014 IEEE Spoken Language Technology Workshop (SLT),* 2014, pp. 189-194.

[21] Sukhatme, G.S. and Mataric, M.J., Embedding robots into the internet. Communications of the ACM, 2000. 43(5), pp.67-73.

[22] Scheutz, M. and Malle, B.F., May. Think and do the right thing: a plea for morally competent autonomous robots. In *Proceedings of the IEEE 2014 International Symposium on Ethics in Engineering, Science, and Technology* .2014, p. 9.

[23] How, J.P., Ethically Aligned Design [From the Editor, as Part of the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems]. *IEEE Control Systems*, 2018. 38(3), pp.3-4.