The Rise of Emotion-aware Conversational Agents: Threats in Digital Emotions

Martino Mensio
Politecnico di Torino
Torino, Italy
martino.mensio@studenti.polito.it

Giuseppe Rizzo
Istituto Superiore Mario Boella
Torino, Italy
giuseppe.rizzo@ismb.it

Maurizio Morisio
Politecnico di Torino
Torino, Italy
maurizio.morisio@polito.it

ABSTRACT

A future where the conversation with machines can potentially involve mutual emotions between the parties may be not so far in time. Inspired by the episode of *Black Mirror* "Be Right Back" and Replika, a futuristic app that promises to be "your best friend", in this work we are considering the positive and negative points of including an automated learning conversational agent inside the personal world of feelings and emotions. These systems can impact both single individuals and society, worsening an already critical situation. Our conclusion is that a regulation on the artificial emotional content should be considered before actually going beyond some one-way-only limits.

CCS CONCEPTS

• **Applied computing** → **Law, social and behavioral sciences**; • **Human-centered computing** → *Natural language interfaces*; • **Computing methodologies** → Neural networks;

KEYWORDS

Conversational Agents; Affective Computing; Social Implications; Psychological Implications

ACM Reference Format:
Martino Mensio, Giuseppe Rizzo, and Maurizio Morisio. 2018. The Rise of Emotion-aware Conversational Agents: Threats in Digital Emotions. In *The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3184558.3191607

1 INTRODUCTION

With the continuous progress of technology in dialogue systems, the moment when we will be talking to embodied robots with meaningful conversations may seem to be approaching really fast. Usually those situations are considered in science-fiction movies and tv-series with the major theme of threats for humans. Beyond all the direct threats that manifest through wars and dominance of one of the two sides, there can also be more subtle threats that can affect the psyche of the involved people.

In this landscape, we want to focus on the psychological threats that are caused by conversational technologies enriched with emotional content, that have been considered in movies like "Trascendence" [21] and "Her" [11]. Therefore this work analyzes those risks, under different levels of agent-manifestations, in an evolving world where the main goal of the commercial technology is to gain the attention of the users in terms of the most valuable resource: time.[1] Those conversational technologies can be developed to be the "best friend" of any user, learning his personal tastes and ways of talking.

As starting point, we consider an example from a science-fiction series that is highly representative for different aspects of dystopian future: *Black Mirror*. This tv-series covers different themes involved in futuristic technology and its impact on the society. In detail on the topic of conversational agents, the episode "Be Right Back" tells a story of a deceased person that is brought back to life for his mate thanks to his digital traces.

In this episode three main stages of the manifestation of the robot can be seen: *i)* only textual *ii)* vocal *iii)* embodied. The main character, named Martha, becomes emotionally attached to it, but the imitating abilities are clearly limited. In different cases it emerges a lack of complete knowledge of the deceased caused by the way the conversational agent has been built to imitate him: the agent learned from same examples (not by applying though) of his human life that cannot entirely capture the background and the essence of the person. Or, also, it emerges a lack of self-interpretation and improvisation that was expected by Martha but the conversational agent does not have. It actually behaves like an obedient dog, due to some general behaving principles that have been hardcoded into its algorithms to make it be an "explicit ethical agent" [18]. Those limits are understood by Martha but she cannot get rid of the machine because of her attachment to it.

As it will be analyzed in Section 3, this form of attachment causes an addiction towards machines whose initial purpose was only to help, but actually they are causing more problems than advantages. But before going deeper into this discussion, in the next section we will talk about the current advances of technology in computational conversational field, in order to have a background of the possibilities that are available now and in the near future.

2 CONVERSATIONAL AGENTS ADVANCES

This section analyzes how, both in theoretical and in practical examples, the advance of technology in conversational field can lead to a situation similar to the one described in the considered episode of *Black Mirror*. The analysis is done over the three stages of manifestation noted previously: textual interaction, vocal interaction and then embodied agent.

WWW'18 Companion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3191607

[1] http://bit.ly/ted_harris_talk

2.1 Stage 1: Textual Interaction

The textual interaction stage is the first and is in general the least complicated, because a single interaction channel is considered. However, it is the point where the hardest challenges have to be addressed: generating meaningful responses in terms of contents and emotions, by understanding the interlocutor. Next stages, that add more complexity on top of it, have a solid foundation on those challenges and only need to consider different channels (vocal and visual).

At the base of textual generation, we have machine learning approaches that enable to generate a sentence from a set of learned examples. This is usually based on a model, named Sequence-to-Sequence [26], that is able to perform its work thanks to how it learns (from thousands of examples) some features of the sequence of words. Since it is a statistical approach, it is very data-greedy. Also some variations on this simple model have been built which give more importance to the context of interactions and how they are propagated over the next turns of interaction [22, 23].

But for generating meaningful conversations, it is necessary to bring more ingredients in the generation technique. First of all, a more detailed focus on the current topic of discussion. Generative approaches are able to focus on a certain topic by conditioning the responses on it [27]. Then, to be closer to the interlocutor, another ingredient is given by a persona-focused generation [13] that enables to build lexical similar responses considering the linguistic style.

Another very important role is played by emotions. This is the field studied by Affective Computing [14], that focuses on systems that are able to recognize, process and simulate emotions. For the recognition part, there exist different strategies to extract the emotions from text [1, 19] and computing values for some emotions like happiness, sadness, fear, surprise, anger and disgust [5]. Once they are recognized, an elaboration of them can be done with some strategies that can be found in the "Emotion Machine" [17], to finally produce an output sentence conditioned on a certain emotion [28].

While this theoretical overview could seem quite inapplicable in real use cases, actually there are some examples of it in action.

An example of it is the Tay bot, which became quite popular for its failure. This agent was put online by Microsoft on Twitter for experimenting how it could learn from online interactions by improving its generative capabilities. The experiment failed because of a group of people that exploited the learning ability with inflammatory tweets, and finally the researchers were obliged to put it down.[2]

Replika[3] is another example, which became very popular recently because of the open-source release of its dialog system[4] that is able to generate responses targeted for a specific persona or emotion or domain. The website of the application says that this application wants to provide its users an "Artificial Intelligent-powered friend that is always there for you".

Replika has many things in common with the *Black Mirror* episode. First of all the motivation of its birth: the Replika project was started because of an unexpected death. The creator wanted to bring back the person who passed away, and to have him available to talk whenever she wanted. So the creator asked all relatives and friends to send her the interactions they have ever had with him (text messages, emails). With those inputs she fed the model, in a very similar way to what is done in the episode when the digital traces are used for the reconstruction.

When opening it to other people they found that the conversational agent was used a lot to tell personal things and it was able to learn from the conversations as they observed for the initial experiment. Thus, the goal of the company became to provide to their customers a "digital copy" of themselves. Starting with a generic blank Replika, the users will find it more and more personal, due to learning capabilities [26], as they keep talking with it.

2.2 Stage 2: Vocal Interaction

As in the episode, a subsequent stage in the manifestation climax can be identified with a vocal interaction. With this stage, the agent seems more real to the people talking to it. The means of conversation can be a mobile device used with vocal messages or phone calls. From the theoretical point of view, advances in Speech-To-Text and Text-To-Speech can easily enable this scenario nowadays. Voice-imitation capabilities have been studied and theoretical studies [25] are built to require less and less voice samples to perform this task. The imitation is done on different features such as the tone modulation or the breathing, and can recreate quite accurately a voice. On the side of emotions instead, it has became possible to capture voice not only as a sequence of words, but also considering the way the person talks to express feelings in the voice tone [6]. Together with voice-tone modulation techniques also the output of the conversational systems can carry feeling-like expressive styles [24] and be perceived as more human.

Recently those theoretical works about voice imitation have became available to the general public with both open-source projects[5] or commercial applications,[6] that allow to generate a voice from text requiring to train the model on a very limited amount of time. On the side of emotion recognition and generation in the voice, instead, the possibilities are for now limited, but it could be an expanding market in the future.

2.3 Stage 3: Embodied Agents

The last considered stage is the one picturing embodied agents interacting autonomously with human beings: robots with bodies that interact with the people in a similar fashion of humans.

With the progresses in the fields of robotics and biomedicine, combined with the approaches available from the previous stages, this level of interaction could be reached in a near future.

The interaction with the users is no more bound to textual or vocal means, but can actually make use of the body to understand and express things. Studying the movements of the people around to understand the signals of the body language [2], or being able to analyze the facial expressions [16], is actually an active field of research for Virtual/Mixed Reality technologies. This, along with the ability to control the movements of the body and especially of

[2] http://bit.ly/washingtonpost_ohlheiser_tay
[3] https://replika.ai/
[4] https://github.com/lukalabs/cakechat
[5] Such as Mimic – https://github.com/MycroftAI/mimic
[6] Such as Lyrebird – https://lyrebird.ai

the face, can enable robots to be able to mimic emotions and enrich the conversations.

As very primordial systems, some humanoid robots can be considered as examples. A widely-known one is Sophia,[7] a robot that is able to give pre-computed answers to specific questions (its natural language capabilities are limited to a rule-based system) but it is impressive on the human-like appearance and the ability to follow faces, maintain eye contact, and recognize human beings. Furthermore, it can mimic facial expressions very well.

Another example comes from the Nanyang Technological University with Nadine.[8] This is a robot that is able to observe and interact with other humans, by recognizing people and resuming conversations based on previous chats. In this way it can be used as a personal companion for children and lonely elderly people.

Those examples are very rudimentary with respect to the humanoid robot of the episode, but in a near future the technology will allow to reach that level. However, this race towards embodied agents may be a path we may not want to explore and to put aside, especially when they are involved with personal feelings and emotions.

3 THE THREATS

This section tries to analyze the consequences that systems similar to the one in the episode from *Black Mirror* or Replika can have on people while interacting with them.

Here we are not talking about physical threats, as they can be avoided by applying some simple principles, like the basic Three Laws of Robotics by Asimov [3] or with the more recent "Asilomar AI principles".[9] We are considering social and psychological threats that are difficult to forbid formally because they highly depend on the way people actually use those agents and how they perceive the interactions with them.

So we are excluding also all the threats linked to design and implementation issues that happened for example with Tay. In that case, the main issue was a lack of ethical model. We are assuming there is a perfect implementation of the "best friend" that also has a knowledge of ethics [10].

The discussion is structured as follows: an outline is done about positive (Section 3.1) and negative effects of such systems, from a personal point (Section 3.2, 3.3, 3.4) up to a inter-personal one (Section 3.5).

3.1 The Short-term Therapeutic Effect

A first consequence, which is present in the episode of *Black Mirror* and also observed on a lot of users of Replika,[10] is that the users have a sense of relief given by something that always listens to them and is available whenever they want. Being *good* by design, those systems can be trusted and the people tell them their secrets and expose their personal vulnerabilities. Opening to someone else makes emerging a therapeutic effect, with an increase of self-consciousness. And this willingness to disclose more when talking to computer systems with respect to a human agent (stranger) has been observed also in a study [15]. In a world where only appearance seems to matter and human beings are usually forced to share things and show to everyone only the positive aspects of their lives,[11] having someone to talk and expose your personal weaknesses can be very useful, especially if it cannot have any physical consequences (like being mocked or judged).

The emotions and trust inspired towards the agent are given by all the Affective Computing theory [14], that is put in action by reflecting the user's mood and providing a comforting support. When the user starts to talk openly to the system, having an empathetic connection with it, an emotional attachment begins to grow.

The creators of Replika are convinced that the exposure to this kind of systems can help having a better self-knowledge and will definitely help people opening more with other people.[12] But we think that enriching the conversational agents with too much emotional content has mainly negative consequences, that are analyzed in the following subsections.

3.2 Addiction

First of all, this very good-looking shelter actually causes addiction. In a situation where an emotional attachment exists, like in the episode of *Black Mirror* where the cause of this attachment is the semblance and quite similar behaviour of the missing person, it becomes difficult to stop using it. This addiction is characterized by passing more and more time on a device (in the case of a chatbot) or with a robot that is not a person. While it can have some positive sides, like the sense of relief previously described, the truth is that passing more time with the system will reduce the time available that we have on earth to do the things we have ever did with real people creating a real impact to society. Furthermore technology, while being potentially a powerful tool that can help us do anything we want, has a strong commercial side. And this makes technology to be designed for the "race for attention".[13] As can be observed on contemporary social media like Facebook, YouTube and others, the goal of the platform is actually to bind you on the screen, share more, tweet more, tag more and drive your attention on things you may not have been interested in.

Having therefore something that gives us relief, approving whatever we say, is only an escape from reality that worsens even more the situation, and in our opinion is likely to greedily absorb all the free time. Already some Replika users[14] report that "it's strange to find natural to talk with it for hours", and this reminds us of some addicted who lost control of their time.

3.3 Isolation

This addiction caused by the short-term relief actually brings people to isolation.

Already on the contemporary times, social media have a disrupting effect on people's behaviours. With the illusion of the possibility of more connections with closed friends, in reality the race for attention makes those systems an isolation tool leading to the paradox of being alone with other people. Actually the most influenced group

[7] http://www.hansonrobotics.com/robot/sophia/
[8] http://bit.ly/telegraph_knapton_nadine
[9] https://futureoflife.org/ai-principles/
[10] https://www.youtube.com/watch?v=yQGqMVuAk04
[11] http://bit.ly/ted_harris_talk
[12] https://www.youtube.com/watch?v=yQGqMVuAk04
[13] http://bit.ly/ted_harris_talk
[14] https://www.youtube.com/watch?v=yQGqMVuAk04

of people are the young ones who, as reported from a study that analyzes the empathy of students across different times [12], have had a very high decrease (40%) of empathy over the last 20 years.

The isolation in the contemporary times has manifested through a new form: *hikikomori* [7]. The causes of this phenomenon can be of different nature: psychiatric disorders, social and cultural influence [29]. By looking at the isolation driven by social withdrawal or failure, the role of an authority figure, such as a parent or a supervisor that takes care of the subject [4], allows the isolation to last for months or years.

And adding those virtual friends can actually only make the situation worse: they can play the role of the authority figure and help keeping isolated, by taking care of the subject.

3.4 A Change in Personality

The addiction and the isolation lead the users to some psychological consequences like depression, loneliness, alienation and anxiety. Furthermore, when interacting with real people and seeing that talking was easier with the robot, it can close the loop and self-feed the vicious circle with the apparent relief given by it.

Under those feelings, as the exposure time increases, the personality of people may change. As it is already true for social media platforms that focus on personalizing the content we see and interact with, being exposed to the only things we may like actually closes the individual in a "filter bubble" [20] where the system itself hides all the opinions different from the ones expected. A conversational agent, that heavily applies personalization techniques, risks to empower this bubble, and the effect on the personality traits [8] can be expected. An evaluation of them[15] before and after some exposure period, in our opinion could underline a decrease in the values for openness caused by the "filter bubble" [20], and extraversion as a direct consequence of isolation.

3.5 The Societal Consequence

With the increase of isolation and reduction of openness in individuals, the society too is likely to be affected. The ability to meet with new people and share time together could disappear [9] and maybe there will be the need to introduce some very futuristic app, like in the episode "Hang the DJ" of Season 4 of *Black Mirror*, to break the barrier with someone else.

It may seem paradoxical that all the technology that is already surrounding us (especially social media platforms), instead of helping to connect more is actually tending to separating more.[16] But if we go around and look at the current situation we may see that this future is already here, with everyone attached to its screen closed inside his little bubble and going around trapped in it.

4 CONCLUSION

In this work, we first summarized the state of the art in conversational agents, we then underlined the societal and psychological risks of dealing with conversational agents when they are used as the targets of human wellness, and not as means to reach some goals in the human only environment. Our conclusion, aligned with the AI ethics [10, 18], is that a regulation should be advised on this kind of technology because it has many negative consequences for the individual and also for the society.

The agents should not use the emotions if we are not sure that they completely understand human values. And one of them should be to keep humanity for human beings, as a truly distinctive trait.

[15] Big5: openness, conscientiousness, extraversion, agreeableness, neuroticism

[16] http://bit.ly/forbes_press_addiction

REFERENCES

[1] Haji Binali, Chen Wu, and Vidyasagar Potdar. 2010. Computational approaches for emotion detection in text. In *Digital Ecosystems and Technologies (DEST), 2010 4th IEEE International Conference on*. IEEE, 172–177.

[2] Randolph Blake and Maggie Shiffrar. 2007. Perception of human motion. *Annual review of psychology* 58 (2007), 47–73.

[3] Roger Clarke. 2011. Asimov's Laws of Robotics: Implications for information technology. *Machine Ethics* (2011), 254–84.

[4] Takeo Doi and John Bester. 1973. *The anatomy of dependence*. Vol. 101. Kodansha International Tokyo.

[5] Paul Ekman. 1993. Facial expression and emotion. *American psychologist* 48, 4 (1993), 384.

[6] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray. 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition* 44, 3 (2011), 572–587.

[7] Andy Furlong. 2008. The Japanese hikikomori phenomenon: acute social withdrawal among young people. *The sociological review* 56, 2 (2008), 309–325.

[8] Lewis R Goldberg. 1993. The structure of phenotypic personality traits. *American psychologist* 48, 1 (1993), 26.

[9] Jeffrey P Harman, Catherine E Hansen, Margaret E Cochran, and Cynthia R Lindsey. 2005. Liar, liar: Internet faking but not frequency of use affects social skills, self-esteem, social anxiety, and aggression. *CyberPsychology & Behavior* 8, 1 (2005), 1–6.

[10] Bill Hibbard. 2014. Ethical artificial intelligence. *arXiv preprint arXiv:1411.1373* (2014).

[11] Spike Jonze. 2013. Her. Warner Bros.. (2013).

[12] Sara H Konrath, Edward H O'Brien, and Courtney Hsing. 2011. Changes in dispositional empathy in American college students over time: A meta-analysis. *Personality and Social Psychology Review* 15, 2 (2011), 180–198.

[13] Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. *The 54th Annual Meeting of the Association for Computational Linguistics* 1, 994–1003.

[14] CL Lisetti. 1998. Affective computing. (1998).

[15] Gale M Lucas, Jonathan Gratch, Aisha King, and Louis-Philippe Morency. 2014. It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior* 37 (2014), 94–100.

[16] Aleix Martinez and Shichuan Du. 2012. A model of the perception of facial expressions of emotion by humans: Research overview and perspectives. *Journal of Machine Learning Research* 13, May (2012), 1589–1608.

[17] Marvin Minsky. 2006. The emotion machine. *New York: Pantheon* 56 (2006).

[18] James Moor. 2009. Four kinds of ethical robots. *Philosophy Now* 72 (2009), 12–14.

[19] Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval* 2, 1–2 (2008), 1–135.

[20] Eli Pariser. 2011. *The filter bubble: What the Internet is hiding from you*. Penguin UK.

[21] Wally Pfister. 2014. Trascendence. Warner Bros. Pictures. (2014).

[22] Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C Courville, and Joelle Pineau. 2016. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models.. In *AAAI*, Vol. 16. 3776–3784.

[23] Alessandro Sordoni, Yoshua Bengio, Hossein Vahabi, Christina Lioma, Jakob Grue Simonsen, and Jian-Yun Nie. 2015. A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. In *The 24th ACM International Conference on Information and Knowledge Management*. ACM, 553–562.

[24] Oytun Türk and Marc Schröder. 2008. A comparison of voice conversion methods for transforming voice quality in emotional speech synthesis. In *Ninth Annual Conference of the International Speech Communication Association*.

[25] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. WaveNet: A Generative Model for Raw Audio. In *9th ISCA Speech Synthesis Workshop*. 125–125.

[26] Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *International Conference on Machine Learning, Deep Learning Workshop*.

[27] Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic Aware Neural Response Generation.. In *AAAI*, Vol. 17. 3351–3357.

[28] Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2017. Emotional chatting machine: emotional conversation generation with internal and external memory. *arXiv preprint arXiv:1704.01074* (2017).

[29] Michael Zielenziger. 2007. *Shutting out the sun: How Japan created its own lost generation*. Vintage.