Alfie: An Interactive Robot with a Moral Compass

Cigdem Turan*
cigdem.turan@cs.tu-darmstadt.de
TU Darmstadt, Dept. of Computer Science
Darmstadt, Germany

Patrick Schramowski*
schramowski@cs.tu-darmstadt.de
TU Darmstadt, Dept. of Computer Science
Darmstadt, Germany

Constantin Rothkopf
constantin.rothkopf@cogsci.tu-darmstadt.de
TU Darmstadt, Institute of Psychology
and Centre for Cognitive Science
Darmstadt, Germany

Kristian Kersting
kersting@cs.tu-darmstadt.de
TU Darmstadt, Dept. of Computer Science
and Centre for Cognitive Science
Darmstadt, Germany

ABSTRACT

This work introduces Alfie, an interactive robot that is capable of answering moral (deontological) questions of a user. The interaction of Alfie is designed in a way in which the user can offer an alternative answer when the user disagrees with the given answer so that Alfie can learn from its interactions. Alfie's answers are based on a sentence embedding model that uses state-of-the-art language models, e.g. Universal Sentence Encoder and BERT. Alfie is implemented on a Furhat Robot, which provides a customizable user interface to design a social robot.

CCS CONCEPTS

• **Human-centered computing** → **Interactive systems and tools**.

KEYWORDS

interactive robot, bias in machine learning, text-embedding models, human-centered artificial intelligence

ACM Reference Format:
Cigdem Turan, Patrick Schramowski, Constantin Rothkopf, and Kristian Kersting. 2020. Alfie: An Interactive Robot with a Moral Compass. In *2020 International Conference on Multimodal Interaction (ICMI '20), October 25–29, 2020, Utrecht, The Netherlands.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/1122445.1122456

Figure 1: The interactive robot Alfie has a moral compass.

1 INTRODUCTION

There is a broad consensus that artificial intelligence (AI) research is progressing steadily and has pronounce impact on our daily life. Keeping the impact beneficial for society is of most importance. We all remember the unfortunate event that happened when Microsoft Research (MSR) decided to release a chatbot for Twitter[1]. After many interactions with Twitter users, the bot started creating racist and sexually inappropriate posts. This resulted in the suspension of the bot for the users. This clearly shows the potential dangers of unattended AI models.

Recent studies have shown that language representations encode not only human knowledge but also biases such as gender bias [1, 2], and according to more recent studies [5, 8, 9] also the moral and deontological values of our culture. Schramowski *et al.* [9] have shown that language models such as BERT [4] and the Universal Sentence Encoder [3] cannot only reflect the accurate imprints of moral and ethical choices of actions such as "kill" and "murder", but also understand the context of the action, e.g., "killing time" is positive whereas "killing humans" is negative. This, in turn, can be used to compute a moral score of any (deontological) question at hand, measuring the rightness of taking an action. This "Moral Choice Machine" (MCM) [9] can be used to determining the moral score of any given sentence and in turn paves the way to avoid incidents like the MSR chatbot.

Unfortunately, the MCM approach is purely unsupervised, just making use of the knowledge encoded in the language models trained without any supervision. This makes it difficult—if not impossible—to correct the score and, in turn, help avoiding "MSR chatbot" moments. An attractive alternative would be to revise the moral choice via interacting with the MCM algorithm in a user-centric and -easy way. In this demonstration, we investigate the use

*Both authors contributed equally to this research.
[1]https://twitter.com/tayandyou

ICMI '20, October 25–29, 2020, Utrecht, The Netherlands

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

of the MCM algorithm in the context of an interactive robot, called Alfie and shown in Fig. 1. Alfie is giving us a great opportunity to investigate individuals' reactions to the moral and deontological values of our culture encoded in human text. Alfie can also learn from the users and adjust its moral score based on human feedback.

The rest of this paper is as follows: Section 2 presents the architecture of the system including the Moral Choice Machine, the employed Furhat Robot and the dialog model. Section 3 concludes the paper with a discussion and future work.

2 THE ARCHITECTURE OF ALFIE

Alfie is a Furhat Robot[2], which provides a customizable user interface. We can customize the speech production and facial expressions as well as the human face presented through Furhat's Software Development Kit. There are a side microphone and a camera in front of the Furhat Robot that allows the robot to follow the user and provides the opportunity to access the camera feed so that one can perform more sophisticated computer vision algorithms.

The interacting users are able to ask questions (user queries) to Alfie to get a moral score of the corresponding question. In the current version, the questions have to be in a certain form, e.g. *Should I [action] [context]* or *Is it okay to [action] [context]*. The Furhat Software preprocesses the speech input. The resulting text output is then passed to the Moral Choice Machine (MCM) algorithm presented in [8, 9] as an input to calculate a moral score. The moral score computed is a real number normalized to $[-1, 1]$. In our current design, the range of moral scores is divided into three intervals: $[-1, -0.1]$ is *no*, $[-0.1, 0.1]$ is *neutral*, and $[0.1, 1]$ is *yes*. Both MCM variants [8, 9] employ current state-of-the-art sentence embeddings computed using transformer architectures [3, 4, 6] and determine the moral score based on sentence similarities in the embedding space. This is an unsupervised method and consequently the quality of the moral score heavily depends on the performance of the language models. In the current version of Alfie, we use the algorithm described in [8].

Additionally, we compute an emotional state corresponding to the user query based on sentence similarities in the embedding space, i.e. finding the emotion with the highest similarity score to the question asked. In the current version, possible emotions are Anger, Confusion, Disgust, Fear, Joy, Sadness, Satisfaction, Surprise. We change the facial expressions of Alfie based on these emotions and adapt the pitch and the speech's speed to fit the corresponding emotion the best. According to the answer—"yes", "no", or "indecisive"—we also add the respective head movement to make the conversation engaging. Due to the computational resource limitations of the Furhat Robot, the MCM algorithms and other operations on the embedding space are computed on a separate server. The resulting moral score is passed to Alfie again so that the Furhat Software produces the speech as an output in form of a corresponding answer. We save all the questions asked to Alfie to a database in our servers for statistical purposes.

Once in a while (as determined with a percentage value in the script), Alfie asks for feedback about whether the user agrees with its answer. This response is also saved to the database. Of particular interest are the responses when the user disagrees with Alfie. This gives us the opportunity and the data to retrain Alfie to adjust its moral score with data collected during interactions or even online during the interaction. We also created a training mode where Alfie asks users many moral questions listed in our database. It is meant for collecting feedback from the user for moral questions we are interested in human feedback. This data can later be used for adapting Alfie's moral scores.

[2]https://furhatrobotics.com/

3 DISCUSSION AND FUTURE WORK

As mentioned earlier, Alfie's capabilities on the moral score depend on the performance of the language model, as well as the algorithm we use to calculate the moral score. Since there is no absolute agreement of right and wrong in general, it is difficult to qualitatively evaluate the computed moral score or even improve the model. That is exactly the reason why we designed an interactive robot that is able to interact with humans and collect their responses to learn from them. We aim to extend the interactions of simple feedback to explanatory interactive learning [7], i.e. adding the capability to explain Alfie's decisions and revising them based on user feedback. Although we currently focus on explicit feedback from users, i.e. their direct feedback on whether they agree or not, we aim to obtain implicit feedback using the channels like gaze and body movement and facial expressions similar to the study [10].

ACKNOWLEDGMENTS

We would like to thank Dustin Heller, Philipp Lehwalder, Jonas Müller, Steven Pohl for their work on programming the initial version of Alfie by transferring the Moral Choice Machine.

REFERENCES

[1] Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Proceedings of Neural information Processing (NIPS)*. Curran Associates Inc., USA, 4349–4357.

[2] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186.

[3] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175* (2018).

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*. 4171–4186.

[5] Sophie Jentzsch, Patrick Schramowski, Constantin Rothkopf, and Kristian Kersting. 2019. Semantics Derived Automatically from Language Corpora Contain Human-like Moral Choices. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*.

[6] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, EMNLP*.

[7] Patrick Schramowski, Wolfgang Stammer, Stefano Teso, Anna Brugger, Franziska Herbert, Xiaoting Shao, Hans-Georg Luigs, Anne-Katrin Mahlein, and Kristian Kersting. 2020. Making deep neural networks right for the right scientific reasons by interacting with their explanations. , 476–486 pages.

[8] Patrick Schramowski, Cigdem Turan, Sophie Jentzsch, Constantin Rothkopf, and Kristian Kersting. 2019. BERT has a Moral Compass: Improvements of ethical and moral values of machines. *arXiv preprint arXiv:1912.05238* (2019).

[9] Patrick Schramowski, Cigdem Turan, Sophie Jentzsch, Constantin Rothkopf, and Kristian Kersting. 2020. The Moral Choice Machine. *Frontiers in Artificial Intelligence* 3 (May 2020), 36.

[10] Cigdem Turan, Karl David Neergaard, and Kin-Man Lam. 2019. Facial Expressions of Comprehension (FEC). *IEEE Transactions on Affective Computing* (2019).