Ethical Chatbot Design for Reducing Negative Effects of Biased Data and Unethical Conversations

Junseong Bang*
Electronics and Telecommunications Research Institute
Daejun, Republic of Korea
Email: hjbang21pp@etri.re.kr

Sineae Kim
Ewha Womans University
Seoul, Republic of Korea
Email: shinaekim@ewha.ac.kr

Jang Won Nam, and Dong-Geun Yang
University of Science and Technology
Daejun, Republic of Korea
Email: dv0195@kribb.re.kr; ydk8027@kriss.re.kr

2021 International Conference on Platform Technology and Service (PlatCon) | 978-1-6654-1766-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/PlatCon53246.2021.9680760

***Abstract*— AI technology is being introduced into various public and private service domains, transforming existing computing systems or creating new ones. While AI technologies can provide benefits to humans and society, the unexpected consequences (e.g., malfunctions) of AI systems can cause social losses. For this reason, research on ethical design for the development of AI-based systems is becoming important. In this paper, from existing studies on AI ethics, general guidelines such as transparency, explainability, predictability, accountability, fairness, privacy, and control for the ethical design of AI systems are reviewed. And, based on the ethical design guidelines, we discuss ethical design to reduce the negative effects of biased data and unethical dialogues in AI-based conversational chatbots.**

Keywords—conversational chatbot; ethical design; framework

I. INTRODUCTION

Artificial Intelligence (AI) technology is being introduced into various public and private service domains, transforming existing computing systems or creating new ones [1]. AI can be used to analyze huge amount of data, predicting crimes in the police security field [2], discovering new uses of existing drugs or developing new drugs in the medical field. In the field of education and training, AI tutors for personalized learning and counseling will also appear in real and virtual spaces. AI technology can provide more benefits to humans and society. However, as the tasks to be processed in AI systems become more complex and the scope of tasks expands, the number of unexpected errors may also increase [3]. By using the latest AI technologies such as Generative Adversarial Networks (GANs), it is possible to create human voices and faces that can be mistaken as humans [4]. An AI-powered chatbot (e.g., Google Duplex) that perfectly mimics the human voice could potentially be used for phone fraud [5]. The use of AI technology for malicious or improper purposes causes social losses. Research on AI ethics is becoming increasingly important, in order to mitigate or resolve the negative impact on individuals and society of AI systems that may be caused by data bias, malicious access, etc., as well as issues of malfunctions in AI systems [6].

Ethics, as a branch of philosophy, generally organizes the concepts of social values, fairness, norms of behavior, and discusses guidelines for desirable directions. *Computer ethics* is defined as a practical way for computing professionals to make decisions related to professional and social conduct, or as a way to analyze the nature and social impacts of computer technologies and the corresponding formulation and justification of policies for the ethical use of such technologies [7]. Until now, ethics has been focused on solving human morality problems by defining concepts such as good and evil, right and wrong, justice and crime, but with the advent of AI, new issues must be discussed for the AI ethics. *AI ethics* is related to the moral behavior of humans as well as the moral behavior of AI agents in the process of designing, constructing, using, and handling AI beings. Recently, various organizations are discussing AI ethics, but due to the complexity of AI, there are still many things to be discussed in order to derive AI ethics standards [8]-[10]. For AI systems to be designed, implemented, and deployed in an ethical manner, it is necessary to establish an ethics framework for the development and use of technology. The AI ethics framework should consider updating existing laws/institutions and ethical standards so that they can be applied in the context of new AI technologies [11]. Therefore, it is necessary to discuss what the ethical requirements of the AI system are, what is ethical design, and what AI ethical guidelines are needed [12].

Chatbots are interactive interfaces for providing information by continuing a conversation with a person based on text or voice. AI-based chatbots are characterized by generating information even in conversations. To create an AI-based chatbot, a large amount of data is required, and issues related to data management/security and data bias arise. Since sensitive data (e.g., bank account) may be included in the chatbot's conversation with a person, such data processing issues should also be addressed. In addition, decisions made through chatbots need to be monitored so that they do not cause significant personal and social damage. Chatbots influence people in the process of leading conversations with people, and chatbots should enable people to make ethical decisions within the scope of laws and institutions. In the process of developing a conversational chatbot, there are many things that need to be addressed to ensure that the chatbot is designed ethically.

In this paper, we introduce an ethical design for conversational chatbots. In Section II, from existing studies on AI ethics, general guidelines such as transparency, explainability,

predictability, accountability, fairness, privacy, and control for the ethical design of AI systems are reviewed. In Section III, we differentiate between types of conversational chatbots and discuss ethical chatbot design to reduce the negative effects of biased data and unethical conversations from the perspective of implementation. Section 4 concludes this paper.

II. General Guidelines for Ethical Design

A. Organization Activities on AI Ethics

In recent years, many discussions on AI ethics have been taking place internationally. The *IEEE* has launched the "IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems" [13]. "Ethical Certification Program for Autonomous and Intelligent Systems (ECPAIS)" has been promoted to increase the transparency and accountability of autonomous and intelligent systems, and to reduce algorithmic bias [14]. The IEEE Standards Association has started a series of IEEE P7000 standards projects dealing with technical, ethical and social issues [15]. Professional associations and non-profit organizations such as the *Association of Computing Machinery (ACM)* have been also issuing recommendations for ethical AI. The *European Commission* has published "Ethics Guidelines for Trustworthy AI" [16], which emphasizes that AI should be human-centered and reliable. The UK's *AI National Plan* explores various issues related to AI ethics, including inequality, social cohesion, prejudice, data monopoly, criminal misuse of data, and suggestions for AI code development [17]. The Australian *CSIRO* shared the findings of the AI Ethics Framework as a toolkit for implementing ethical AI [11]. KPMG releases the "AI in Control" framework to help you realize the value of AI technology while maintaining the reliability, algorithmic integrity, explainability, fairness and agility of AI [18]. KPMG proposed nine ways to monitor the ethical compass for AI design [19], in order to reduce the negative impact of the AI system and maximize the benefit to society through the AI ethics framework. Global companies such as Google [20] and SAP [21], besides international or governmental organizations, have also publicly announced AI principles and guidelines.

Although discussions on AI ethics are being conducted by various organizations, research on design and development frameworks is also required to be applied to industrial technologies.

B. Ethical Principles for AI Systems

Internationally, various organizations are presenting different AI ethical guidelines according to their purpose [12]. For example, the ethical principles defined by the IEEE include human rights, well-being, data agency, effectiveness, transparency, accountability, awareness of misuse, and competence [22]. The ethical principles defined by the Australian CSIRO include human, social & environmental well-being, human-centered values, fairness, privacy protection and security, reliability & safety, transparency & explainability, contestability, and accountability [11]. Although there are slight differences in the principles presented in these AI ethics guidelines, it can be seen that there are common ethical principles.

1) Transparency: This includes not only the transparency of data and algorithms, but also about how the system works. It is directly related to system design and development [22], [23]. Transparency of data and algorithms is essential to creating an Explainable AI system. In addition to being able to understand how the system works, it should also explain why the system needs its data and algorithms. In the case of AI-based chatbots, transparency should also be considered in terms of communication with human users (i.e., through dialogues). Chatbots should not present themselves to users as humans, and human users have the right to know that they are interacting with an AI system (i.e., chatbots).

2) Predictability: Predictability should be considered to ensure that what we expect of the AI system to do or can do. This is sometimes viewed as a subset of transparency. Because transparency explains how the system works, it is possible to predict which output will be based on the input data. For example, people expect the smart home's AI-based indoor environment control system to control temperature, humidity, etc. in a state that makes people feel comfortable.

3) Accountability: This is a stakeholder-related accountability issue, which means that who is responsible for who, what, why and how should be defined [22], [23]. Accountability means responsibility under laws and regulations, but responsibility means responsibility for what we think is right.

4) Fairness: For the fairness of AI systems, representativeness must be maintained and biases must be prevented in terms of data, and fairness in benefits to human users must be ensured in terms of systems. In the case of an AI-based chatbot, the characteristics (e.g., voice tone, speed, conversation style) of the chatbot can be adaptively tuned for user-friendly response, but the chatbot should not discriminately lead conversations to human users based on race, residence, occupation, etc.

5) Privacy: Privacy is an essential part of the ethical design of AI systems. Training of AI systems requires large amounts of data, and the collection and processing of personal information requires explicit consent from data providers (i.e., users) [24], [25]. In the case of a conversational chatbot, even in the course of a conversation with a human user, information is collected and used for processing, so privacy in the conversation must be considered.

6) Control: Control affects public trust in AI, along with ethical principles of transparency, accountability, fairness, privacy, etc. AI systems out of control pose a great risk to individuals and society [26]. While there is damage as a direct output of the AI system, there can also be gradual damage as it interacts with the AI system. Conversational chatbots can influence users through uncontrolled and unethical processes. In addition, uncontrolled processing or reuse of user decisions can also cause problems.

The AI ethics framework requires the development of a set of rules to control the use of technology and its behavior. In order to increase the reliability of complex AI systems and reduce defects, it is necessary to consider AI rules in various

aspects [27]. The most prevalent principle is transparency, which is directly related to trust in AI systems, followed by the definition and fairness necessary to avoid inequality caused by the use of AI [28].

III. Ethical Design for Conversational Chatbot

Ethical AI refers to AI algorithms, architectures, and interfaces that follow AI's ethical principles such as transparency, accountability, fairness, and privacy. Despite various efforts towards AI's ethical principles, uncertainty remains about how to implement AI's ethical principles into real systems or services [12]. When designing and implementing AI systems based on AI ethical principles or guidelines, complexity, variability, and subjectivity (i.e., including variable interpretation of each ethical principle or guideline) should be considered [29]. The ethical design should take into account socially accepted objectives, safe and responsible methods, levels of social risk perception, and socially beneficial outcomes [30].

A. Ethical Design in Data Perspective

Data management in the industrial age meant legal data collection, but that in the information age also includes data use and value creation. In AI systems/solutions, AI models are trained using large amounts of data for various business purposes. Data is very important to an AI system. The behavior of the AI model is fundamentally influenced by the characteristics of the data set [31]. A number of questions and corresponding guidelines for using the data set and the workflow for handling information about the data set should be provided [31]. For example, the reliability of the data, the purpose of creating the data set, the system that can use the data, the manager of the data, etc. should be discussed.

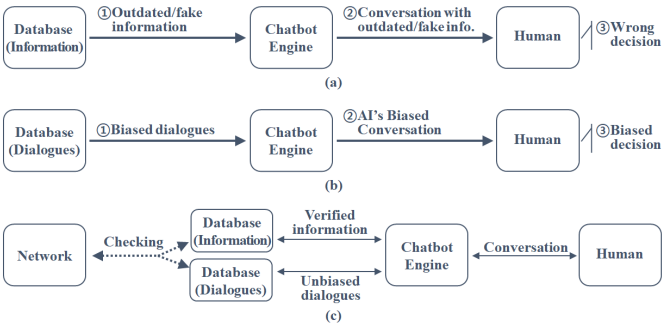

Fig. 1. Ethical issue design for a chatbot with biased data.

Even if data transparency is guaranteed in conversational chatbots, there are several other risks. As shown in the Fig. 1(a), outdated or fake information remains in the database connected to the chatbot engine, and if this information is passed on to a human user in the course of a conversation, the human user can make wrong decisions. That decision may require accountability. It is necessary to design ethical dialogue policies to reduce the risks and losses of using conversational chatbots.

Some of the commercialized chatbots respond to human user queries based on the dialogues for scenarios. As such, biased responses on the dialogues may be exposed to the user. As shown in the Fig. 1(b), biased conversations in chatbots can lead human users to make biased decisions.

To avoid this problem, basically, data must be updated periodically over the network. It is necessary to check whether the data is biased, whether the information is valid, etc. Data needs to be updated in real time according to the type of information (i.e., importance, urgency) from time to time during the conversation, as shown in the Fig. 1(c).

B. Ethical Design in Conversation Perspective

A chatbot is an AI-based system that provides services while continuing a conversation with a human user. Even if a chatbot system guarantees transparency, predictability, accountability, fairness, privacy, and control, unethical conversations between conversational chatbots and users can negatively affect individuals and society.

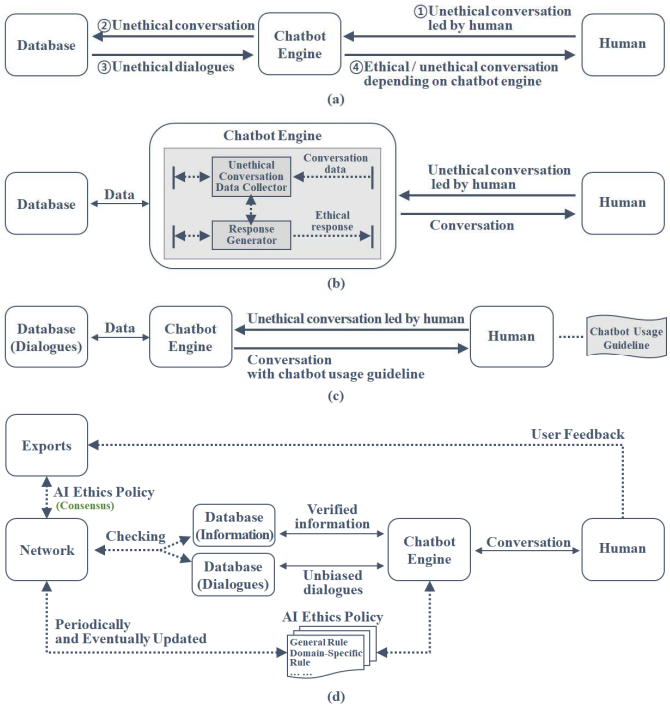

Fig. 2. Ethical design to reduce negative effects of unethical conversations.

As shown in the Fig. 2(a), An unethical conversation can be initiated by a human user. It can be classified according to whether the chatbot learns a query from a user input. First, the chatbot understands the intent from the user input sentence but does not learn it. Human users can make unethical queries with malicious intent or no specific purpose. In this case, the chatbot will usually respond based on a database of pre-built conversation scenarios, so you will not have an unethical conversation. However, without a chatbot's ability to track conversational status, the chatbot can sometimes embarrass human users with

out-of-context, out-of-context answers, or make users misjudgment.

There may be cases where a chatbot learns by collecting conversations with users. In this case, the chatbot may collect the conversation and send it to the conversation database, or send information from the collected conversation to the conversation database. In this case, data biased by the conversation scenario built using or using the database may be transmitted to the chatbot, which may lead to the chatbot continuing unethical conversations.

The chatbot should be designed considering the case of collecting and utilizing conversations for conversational queries biased by human users and malicious intent. As shown in the Fig. 2(b), at this time, the chatbot system needs a module that can determine the unethical of conversation and collect it. Using the collected data, a conversation scenario can be designed and added to the chatbot engine about how the chatbot should respond if an unethical conversation is attempted.

As shown in the Fig. 2(c), chatbots are responsible for notifying human users what conversations they are capable of. If a human user continues a conversation with the chatbot with malicious intent or ignorance, it is necessary to inform the user of the potential risks in the current scenario.

Social values change according to circumstances and time. Therefore, as shown in the Fig. 2(d), it is necessary to update the conversation policy of the chatbot through the discussion of experts after receiving user feedback. Through this, new values can be easily reflected in general chatbot systems.

1) Privacy in Conversation: In the course of conversations with users, personal and sensitive information must be thoroughly secured and must not be reused. And if you want to use it, you must obtain consent from the user and notify the secure processing process thereafter. Chatbots also access sensitive information such as bank accounts by using information in conversations with users. Alternatively, additional authorization (i.e., Account Binding) may be requested for interworking with other companies' services. For example, appropriate authority may be required to provide services linked with other companies' schedule management systems. This must be done through user credentials.

2) Conversation Topic and Content: A chatbot can make its functions transparent by declaring within the conversation what it can do, etc. Make sure your chatbot doesn't get off topic. Chatbots should also look at expressions in conversations.

3) Information Reliability: If it is not provided by linking data updated in real time, it is necessary to continuously check whether the information delivered to the user is reliable. Therefore, information that may be relevant to important decision-making needs to be communicated with the rationale and timing of obtaining the information.

4) Mutual Influence: The IEEE proposed two ethical practices, ethical design and sustainable development, and focused on "the social impact of existing and emerging technologies, including intelligent systems" [33]. It can be recognized that not only the implementation of the system itself, but also the mutual influence of humans and society should be considered.

Since chatbots communicate information and continue conversations with users through a question-and-answer process, if the conversation is continued around the options provided by the chatbot, the user may be induced to make decisions by the chatbot. Chatbot scenario conversations should be conducted within an ethical scope so that users do not make erroneous judgments.

C. Ethical Design for Chatbot Systems

1) Ethical Governance and Accountability: Even if the transparency of the chatbot system is maintained as much as possible, the definition of accountability should be discussed by experts in various fields related to the chatbot service. This is because the chatbot system may store information that is the basis for decision-making, but the final decision on who is responsible for complex problems must be made by humans. Accountability is one way to ensure trust of AI. As discussed above, accountability ensures that if an AI system makes a mistake or harms someone, there is someone that can be held responsible, whether that be the designer, the developer or the corporation selling the AI. In the case of damages occurred, there must be a mechanism for salvation so that victims can be compensated enough.

2) Monitoring Under AI Ethics Guideline: First, it is necessary to monitor AI ethics guidelines. It is necessary to check whether the AI ethics guidelines reflect the characteristics of systems/services in which AI technology is used, and whether the AI ethics norms are appropriately updated according to the temporal/spatial situation. Second, it is necessary to monitor whether the system is designed and implemented based on the AI ethics guidelines or AI ethics framework. It is necessary to check whether there are any violations of the newly updated AI ethics guidelines, and whether there are areas not covered by the existing AI ethics framework. If there is an error due to the system, a technical discussion on this should proceed immediately. In particular, where the social impact can be large, trivial points should also be considered. Finally, it is necessary to monitor the stakeholders related to AI ethics. In the process of designing the AI ethics framework, adopting guidelines, and developing AI systems/services, it is necessary to monitor the impact of each decision maker.

IV. Conclusion

AI ethics requires a discussion on the definition of guidelines and frameworks with participation from a number of stakeholders (e.g., based on different aspects such as philosophical foundations, science and technology ethics, legal aspects, and etc) taking into account not only human and social factors, but also AI system design and development, and its ripple effect. Implementing ethical principles in practice is difficult because of the complexity of AI. There is also a lack of ethical standards used to certify AI solutions. In this paper, a study on the AI ethics was conducted from the perspective of practical development related to the design and implementation of the conversational chatbot system. It is necessary to develop

AI ethics guidelines specific to other AI systems or services, and it is expected that there may be systematic standards for AI ethics through these studies.

ACKNOWLEDGMENT

This research was supported and funded by the Korean National Police Agency. [Pol-Bot Development for Conversational Police Knowledge Services / PR09-01-000-20]

REFERENCES

[1] M. Taddeo and L. Floridi, "How AI Can Be a Force for Good," *Science*, vol. 361, no. 6404, pp. 751–752, Aug. 2018.

[2] J. Bang, Y. Lee, Y.-T. Lee, and W. Park, "AR/VR Based Smart Policing For Fast Response to Crimes in Safe City," *IEEE Int. Symp. Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, Beijing, China, Oct. 10-18, 2019.

[3] V. Vakkuri, K.-K. Kemell, J. Kultanen, and P. Abrahamsson, "The Current State of Industrial Practice in Artificial Intelligence Ethics," *IEEE Software*, vol. 37, no. 4, pp. 50-57, July/Aug. 2020.

[4] T. T. Nguyen, C. M. Nguyen, D. T. Nguyen, D. T. Nguyen, and S. Nahavandi, "Deep Learning for Deepfakes Creation and Detection: A Survey," *arXiv:1909.11573*, July 2020.

[5] H. Li, X. Xu, C. Liu, T. Ren, K. Wu, X. Cao, W. Zhang, Y. Yu, and D. Song, "A Machine Learning Approach to Prevent Malicious Calls Over Telephony Networks," in *IEEE Symposium on Security and Privacy (SP)*, pp. 53–69, 2018.

[6] M. Deane, "AI and the Future of Privacy," *Towards Data Science*, Sep. 05, 2018.

[7] J. H. Moor, "What is Computer Ethics?" *Metaphilosophy*, vol. 16, no. 4, pp. 266–275, 1985.

[8] E. Bird, J. Fox-Skelly, N. Jenner, R. Larbey, E. Weitkamp, and A. Winfield, "The ethics of artificial intelligence: Issues and initiatives," European Parliamentary Research Service, Technical Report PE 634.452, Mar. 2020.

[9] A. Gupta, C. Lanteigne, V. Heath, M. B. Ganapini, E. Galinkin, A. Cohen, T. D. Gasperis, M. Akif, and R. Butalid, "The State of AI Ethics Report (June 2020)," *arXiv:2006.14662*, Jun. 2020.

[10] S. Lo Piano, "Ethical principles in Machine Learning and Artificial Intelligence: Cases From the Field and Possible Ways Forward," *Humanities and Social Sciences Communications*, vol. 7, no. 1, Jun. 2020.

[11] D. Dawson, E. Schleiger, J. Horton, J. McLaughlin, C. Robinson, G. Quezada, J. Scowcroft, and S. Hajkowicz, "Artificial Intelligence: Australia's Ethics Framework," Data61 CSIRO, Australia, 2019.

[12] A. Jobin, M. Ienca, and E. Vayena, "The Global Landscape of AI Ethics Guidelines," *Nature Machine Intelligence*, pp. 389–399, Sep. 2019..

[13] IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, IEEE Standards Association.

[14] IEEE, "IEEE Launches Ethics Certification Program for Autonomous and Intelligent Systems," IEEE Standards Association, Oct. 02, 2018.

[15] A. Koene, L. Dowthwaite, and S. Seth, "IEEE P7003™ Standard for Algorithmic Bias Considerations: Work in Progress Paper," *2018 ACM/IEEE International Workshop on Software Fairness*, Gothenburg, Sweden, pp. 38–41, 2018.

[16] "Ethics Guidelines for Trustworthy AI," *European Commission*, Brussels, Dec. 2018. Accessed: Sep. 16, 2019.

[17] "AI in the UK: Ready, Willing and Able," *Select Committee on Artificial Intelligence*, House of Lords, UK, Apr. 2018.

[18] J. Samuel, "KPMG Launches Framework to Help Businesses Gain Greater Confidence in Their AI Technologies - KPMG Global," KPMG, Feb. 13, 2019.

[19] K. Reid, "The Opportunity of AI: Unlocking the Potential," KPMG, Apr. 05, 2019.

[20] "Artificial Intelligence at Google: Our Principles," Google AI.

[21] "SAP's Guiding Principles for Artificial Intelligence," *System Analysis Program Development (SAP)*, Sep. 18, 2018.

[22] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, "Ethically Aligned Design: A Vision for Prioritizing Human Well-being With Autonomous and Intelligent Systems," IEEE, 2019.

[23] High-Level Expert Group on Artificial Intelligence, "Ethics guidelines for trustworthy AI," European Commission, Brussels, Belgium, 2019.

[24] S. Hallensleben, L. Fetic, T. Fleischer, Grünke, T. Hagendorff, M. P. Hauer, A. Hauschke, J. Heesen, M. Herrmann, R. Hillerbrand, C. Hubig, A. R. Kaminski, T. D. Krafft, W. Loh, P. Otto, M. Puntschuh, and C. Hustedt, "From Principles to Practice : An Interdisciplinary Framework to Operationalise AI Ethics," Sociology, 2020.

[25] H. Nissenbaum, "Contextual Integrity Up and Down the Data Food Chain," *Theoretical Inquiries in Law*, vol. 20, no. 1, pp. 221-256, March 2019.

[26] J. J. Bryson, "The Past Decade and Future of AI's Impact on Society," *Towards a New Enlightenment? A Transcendent Decade*, vol. 11, 2019.

[27] D. E. O'Leary, "Ethics for Big Data and Analytics," *IEEE Intelligent Systems*, vol. 31, no. 4, pp. 81-84, July-Aug 2016.

[28] J. Zhou, F. Chen, A. Berry, M. Reed, S. Zhang, and S. Savage, "A Survey on Ethical Principles of AI and Implementations," *IEEE Symp. Series on Computational Intelligence (SSCI)*, pp. 3010-3017, Dec. 2020.

[29] J. Morley, L. Floridi, L. Kinsey, and A. Elhalal, "From What to How. An Overview of AI Ethics Tools, Methods and Research to Translate Principles into Practices," *arXiv:1905.06876*, May 2019.

[30] M. S. Caron and A. Gupta, "The Social Contract for AI," *IJCAI 2019 AI for Social Good Workshop*, Macau, China, Aug. 10-16, 2019.

[31] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé III, and K. Crawford, "*Datasheets for Datasets*," *arXiv:1803.09010*, Mar. 2018.

[32] M. Dahiya, "A Tool of Conversation: Chatbot," *Int. J. of Computer Sciences and Engineering*, vol. 5, no. 5, pp. 158-161, May 2017.

[33] G. Adamson, J. C. Havens, and R. Chatila, "Designing a Value-Driven Future for Ethical Autonomous and Intelligent Systems," *Proceedings of the IEEE*, vol. 107, no. 3, pp. 518–525, Mar. 2019.