VICTORIA UNIVERSITY OF WELLINGTON TE HERENGA WAKA | Wellington Faculty of Engineering Te Wāhanga Ahunui Pūkaha | Ethics and Sustainability Symposium

Virtual Assistants: A Code of Ethics

Tian Welgemoed

Abstract

An intelligent virtual assistant (IVA) like Amazon's Alexa supports clients remotely by performing administrative, technological, or creative tasks. IVAs have become a part of our daily lives and collect large sums of data about their users, meaning we need a proper code of ethics to protect people's rights to privacy as there are data leakage concerns. In this paper, I outline a code of ethics for IVAs based on research and a case study on Amazon Alexa to recognize the key principles that a code of ethics for IVAs requires to protect its users. This code of ethics can be used for development of new IVAs or updating of existing ones in order for them to follow improved ethics principles.

Keywords: Intelligent virtual assistant; Artificial intelligence; Internet of Things.

1. Introduction

IVAs are becoming increasingly popular for both personal and business use cases as they are easily accessible by users and can perform a wide variety of tasks. IVAs can save people time by doing mundane daily tasks so their users do not have to, and there is a new trend for automation where IVAs can be used to automatically perform tasks when user-specified conditions are met [1]. This can be done by using a time-based system or through an Internet of Things (IoT) approach wherein a dedicated division of the system is used to detect when the conditions are met for a set of tasks to be completed by the IVA [1]. There are many approaches that have been used to create adequate and useful IVAs and although they had an independent upbringing from augmented reality (AR), it is becoming increasingly popular for IVA researchers to use IoT and AR approaches for IVA commercialisation and development [2]. One example of this is using sensors and actuators in the system's environment to trigger tasks to be done when the sensors send certain signals to the actuators. The conditions for the signals can be set up via the IVAs.

Users can communicate with IVA systems through software applications, or, most commonly, voice activation devices. The user communication side of IVAs are developed by using Artificial Intelligence (AI) that must include Speech Recognition and Natural Language Processing techniques for the system to understand the input (human language) and execute the relevant procedures for the given commands [3, 4, 1]. For my case study, I am going to investigate Amazon Alexa, a popular IVA that can communicate with other IoT products and thus perform tasks for its users. Alexa has a mobile application for iOS and Android [5] and is primarily accessed through Amazon Echo products (smart speakers, TVs, etc.), however, Alexa's accessibility is not limited to Amazon products. It can also access third-party systems like Uber or Domino's. A user's activity history is stored on Amazon's cloud servers [5], which is potentially dangerous because there have been a significant number of cases where cloud servers were hacked [6].

1.1. Objective

The objective of this paper is to create a code of ethics for IVAs that will protect user's data from the aforementioned threats. The set of principles involved will address threats that have been proven to violate people 's data privacy rights either directly or indirectly.

This code of ethics could be used for the design of new IVAs or the upgrade of existing IVAs. By following this code, the developers or designers will prevent any illegal and immoral physical or mental actions done towards its users and will thus result in an ethical IVA.

2. Ethical Issues for IVAs

In this section, I will be performing a literature review to discuss ethical issues that relate to IVAs and how they can be a threat to their users. These rights might be country-specific, consequently I will be using the "The EU Approach to Ethics Guidelines for Trustworthy Artificial Intelligence" by the High-Level Expert Group on Artificial Intelligence (HLEG AI) [7] as a guideline to decide what aspects of present-day IVAs might be threatening to the general society of IVA users. The selected issues will take into account the concerns of eight respected, already-existing ethical frameworks [8] that altogether consider 19 different factors. I will provide four principles for the four most common concerns between these frameworks.

2.1. IVA Data Protection and Privacy Issues

As mentioned in the introduction, there have been issues relating to the storage of users' data. Fortinet is a multinational, cybersecurity company that posted cyber attacking statistics about household IoT threats in 2015 and 2016 [9]. They found that there were about 800,000 attacks on home routers alone in 2015 and around 25,000,000,000 in 2016, indicating that homes are being targeted by cyber attackers more and more, especially now that smart homes trending. A researcher from the University of Oxford found that there is a lack of control from a user side over what data is collected and used by smart assistants [10]. A separate group of researchers found that by using Ultrasonic waves, they could for example, hijack a phone's SMS passcode [11]. IVAs operate on a cloud-storage environment, but there are still privacy and security concerns for such environments [12]. It is clear that the security systems of IVAs are not strong enough to be considered safe.

2.2. Accountability Issues

The concerns for accountability in terms of IVAs mainly arise from the IVAs carrying out actions for users when they should not. The main reason for IVAs mistakenly carrying out actions is malicious input by unauthorised users. If a user leaves their IVA "unlocked", other users may be able to use it and get the IVA to perform harmful tasks, which is why Apple's Siri should be turned off when an Apple device is locked [13]. Another reason could be because of "misheard" input where a user said one phrase, but the IVA heard something different due to bad microphone quality or bad Speech Recognition software. IVAs are currently not transparent enough with their data collection and usage [14], which implies that users do not know the risks involved with using IVAs.

2.3. Inequality

Inequality could occur in IVAs when certain groups are not accounted for during development. For example, when an IVA is released to the USA but the Boston accent was left out or their main uses for the IVA were not included in the IVA system – this would be an example of discrimination by design. Discrimination by design is on purpose as it would be part of a plan to leave out representation for or (negatively) target of a group [15]. Discrimination could happen by accident too if a group was left out of training data without the AI developers noticing it. The IVA would thus learn how to interact with everyone except the group that had been left out, thus discriminating against them [15]. The feature selection from the training data could also prove to be discriminatory if only features from the data selected would prove to benefit certain groups or detriment a specific group [15, 16]. An example of this could be when the IVA is developed to be able

to translate words to all languages as features, except for one group's language, thus making it harder for that group to use the IVA.

2.4. Humanity Issues and Purpose

Being reasonably socially conscious [17] is important for an AI that interacts with people through verbal and text conversation. The reason for this is that people have emotions, and they might listen to what an IVA has to say due to it being developed by facts and statistics. According to a survey [18], about 70% of the Chinese respondents trust AI, which puts it in a position of power. People have learned to trust technology, one example being AI that detects and treats mental health issues [19]. This can be dangerous as someone might ask about their self-worth which objectively or economically might not be much, but when this is repeated back to them every day, it could be detrimental to their mental health. This would also be an example of over-stepping the scope. An IVA might not have been trained to take care of people's mental health, but people might not be aware of this and thus let it affect theirs.

3. Code of Ethics

In this section I will be providing the principles that follow the aforementioned guideline provided by the HLEG AI, meaning all principles will be lawful, ethical, and robust [7]. Each of the principles will address the issues from the previous section either via elimination or mitigation.

3.1. IVAs must use modern security approaches for data management

Data should be encrypted whenever it is stored, sent, or used wherever possible with up-to-date security measures. Where data is left unencrypted, it leaves the users vulnerable, and when there are millions of users of an IVA, it is especially important. When attempting to gather sensitive information that is external to the IVA system, there should be a) an option to opt-out/in and b) reasoning for why the system would need it. If a user opts out, they should be made aware of the features they are missing out on. This gives users full control over the data flow in their environments [17, 7]. Requesting permissions for features first would include giving users the option to not have their conversations recorded (even for legal reasons) to protect their privacy rights. Following the release, maintenance will be a high priority due to cyber attackers and researchers finding flaws in the system.

3.2. IVAs must be transparent by allowing users to control which of their data is used

It is important for IVAs to be as transparent and secure as possible [7, 8], they need to include secure procedures for features that deal with external systems, users' privacy or financial matters. This will include a relatively secure way of confirming actions such that features that have little to no impact on these matters don't need confirmation, but features that have an impact on these matters would need a biometric, two-factor or similarly levelled authentication procedure to be followed, thus improving the system's transparency through traceability [15]. Using these procedures would result in a more secure system that would never leave the users blaming the IVA for performing malicious tasks regarding sensitive matters. The system is allowed to not require confirmation of any kind if the user puts the system in a mode wherein authentication is provided to enable the confirmation-less state of the IVA. The user then needs to be made aware that they are responsible for all of the proceeding actions.

3.3. IVAs must represent and consider all user groups of their target audience

As per the issue section regarding equality, all groups from the target audience should be accounted for at all stages of the IVA development plan and in all data to avoid discrimination [15]. This includes consideration of all groups in the target audience

during every step of development for the AI [16] and taking into account all groups' main usages for the IVA and their main interaction methods. Following this principle will result in an equality-safe IVA system.

3.4. IVAs must only fulfil their purpose

The IVA should serve its purpose and not interfere with out-of-scope decision making. When an IVA makes or influences decisions regarding topics that it's unfamiliar with or hasn't been trained on, it will not account for all of the factors that it needs to, thus providing fraudulent solutions. It should be transparent [7] by clarifying what it knows and does not know, and when users make out-of-scope queries, it is especially important to communicate this. It should thus be socially trained enough to affect a user's environment positively and honestly [17] by being respectful, unintrusive and, as mentioned, transparent.

4. Case study: Amazon Alexa

Amazon Alexa is a very popular IVA with 40 billion users just in the USA [20], and 100,000 skills [21] including the capabilities to record voices and shop online. It is thus a well-supported and researched case for evaluating its applied ethics compared to the one I have outlined.

4.1. IVAs must use modern security approaches for data management

Researchers [22] found that it was recently possible to gain control of Amazon Echo devices that are connected to Alexa and perform malicious tasks like listening to private conversations and buying unwanted items by reproducing audio files to create commands. Although this has been patched, there are still flaws in Alexa's security. An analysis of Alexa's environment was done by researchers from the Korea University [5] to create a more efficient and forensically-inclusive environment for Alexa so that it may be used to provide evidence to a court case if necessary – which is becoming an increasingly important topic [23]. They found that by using unlisted Alexa APIs, they could find unencrypted user accounts, Wi-fi settings and passwords, and ways to invoke other cloud services. Although it is good that data is transferred to the cloud storage with an encrypted connection [5], the data that is stored locally on Alexa-enabled devices should be encrypted, and currently, the users are not aware of the stored sensitive data. As per my principle, the users should be asked what data is acceptable to be used in the system, and what features they would miss out on if they declined.

4.2. IVAs must be transparent by allowing users to control which of their data is used

Alexa mainly operates through voice activity, which can be muted. The audio logs for the time that it is muted will not be recorded and old audio logs can be deleted, but companion systems can also store audio logs and often do [24]. A study [14] has shown that cyber attackers can realise a user's behaviours by studying their IVA data that Alexa is not keeping secure enough and suggested that the system should communicate to their users the implications of data breaches. They further explained that there are more data security (and thus privacy) concerns –users should have more control over the data access/usage. It should also be more convenient for users to do so, which would be prioritising the users' needs as per the outlined code of ethics.

4.3. IVAs must represent and consider all user groups of their target audience

It is likely that the African American Vernaculars were underrepresented during the training of Alexa's speech recognition AI as there is a significant performance difference in Alexa's acoustic models for African Americans versus others [25]. It is not proven that underrepresentation is the cause, implying that this could be a design flaw of the AI model, but in either scenario, the current model is not acceptable.

4.4. IVAs must only fulfil their purpose

Due to Alexa having over 100,000 skills, it will try to answer all user queries. It has a purpose to be generally useful in a household environment, but it cannot be used to replace a social environment [26]. Alexa does however clarify that she works best with single specific queries, rather than maintaining a whole complex conversation [27], which indicates good transparency for this principle.

5. Conclusion and recommendations

It can be concluded that IVA systems were in definite need of a proper code of ethics, which I have now provided. As stated, it can be used for the development of a new IVA or for updating the applied ethics to a currently existing system. The resulting system will account protect all of its users' data and consider their needs in the IVA.

I have analysed Amazon's most popular IVA, Alexa, and have found that although it fulfils its purpose by executing innovative and helpful tasks, the system still lacks proper security measures, user control over the data, equality, and social awareness.

In terms of future works, I believe that there should be more research done into the IVA's effects on its users' mental health. More specifically, how IVAs should be communicating with their users [10] and how it can improve the users' mental health. I also believe that security measures are best when they are transparent, which IVAs can improve on, considering recent Alexa events relating to security flaws [22].

References

[1] V. Këpuska and G. Bohouta, "Next-generation of virtual personal assistants (Microsoft Cortana, Apple Siri, Amazon Alexa and Google Home)," in *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)*, 2018.

[2] N. Norouzi, G. Bruder, B. Belna, S. Mutter, D. Turgut and G. Welch, "A Systematic Review of the Convergence of Augmented Reality, Intelligent Virtual Agents, and the Internet of Things," Springer, Cham, 19 February 2019. [Online]. Available: https://par.nsf.gov/servlets/purl/10105846. [Accessed 17 April 2021].

[3] Z. Eberhart, A. Bansal and C. Mcmillan, "A Wizard of Oz Study Simulating API Usage Dialogues with a Virtual Assistant," *IEEE Transactions on Software Engineering,* doi: 10.1109/TSE.2020.3040935, 27 November 2020.

[4] C. Delgrange, J. Dussoux and P. F. Dominey, "Usage-Based Learning in Human Interaction With an Adaptive Virtual Assistant," *IEEE Transactions on Cognitive and Developmental Systems,* vol. 12, no. 1, pp. 109-123, March 2020.

[5] H. Chung, J. Park and S. Lee, "Digital forensic approaches for Amazon Alexa ecosystem," 5 August 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1742287617301974. [Accessed 30 April 2022].

[6] J. D. Groot, "The History of Data Breaches," 1 December 2020. [Online]. Available: https://digitalguardian.com/blog/history-data-breaches. [Accessed 30 April 2022].

[7] N. A. Smuha, "The EU Approach to Ethics Guidelines for Trustworthy Artificial Intelligence," vol. 20, no. 4, pp. 97-106, 2019.

[8] K. Siau and W. Wang, "Artificial Intelligence (AI) Ethics: Ethics of AI and Ethical AI," *Journal of Database Management,* vol. 31, no. 2, pp. 77-87, 1 April 2020.

[9] G. Chow, "FortiGuard Labs Telemetry – Roundup and Comparison of 2015 and 2016 IoT Threats," 6 March 2017. [Online]. Available:

https://www.fortinet.com/blog/threat-research/fortiguard-labs-telemetry-roundup-and-comparison-of-2015-and-2016-iot-threats. [Accessed 1 May 2022].

[10] W. Seymour, "How loyal is your Alexa? Imagining a Respectful Smart Assistant," 18 April 2018. [Online]. Available: https://dl.acm.org/doi/abs/10.1145/3170427.3180289?casa_token=SqLKB06cnQ0AAAAA:jiGRM8NYkCYgk04F9U1vlGxqMnBfaq4Z4lBTAGnkaNERDNYaDPavtSqNk-cTvjIIco4Kf95IF8p3. [Accessed 2 May 2022].

[11] Q. Yan, K. Liu, Q. Zhou, H. Guo and N. Zhang, "SurfingAttack: Interactive Hidden Attack on Voice Assistants Using Ultrasonic Guided Waves," February 2020. [Online]. Available: https://par.nsf.gov/servlets/purl/10186166. [Accessed 2 May 2022].

[12] M. Adelmeyer, P. Meier and F. Teuteberg, "Security and Privacy of Personal Health Records in Cloud Computing Environments – An Experimental Exploration of the Impact of Storage Solutions and Data Breaches," 2019. [Online]. Available: https://www.researchgate.net/profile/Michael-Adelmeyer/publication/331398404_Security_and_Privacy_of_Personal_Health_Records_in_Cloud_Computing_Environments_-_An_Experimental_Exploration_of_the_Impact_of_Storage_Solutions_and_Data_Breaches/links/5c778ebb9. [Accessed 2 May 2022].

[13] L. H. Newman, "Turn Off Siri on Your Lock Screen for Better iOS Security," 23 November 2018. [Online]. Available: https://www.wired.com/story/turn-off-siri-lock-screen-attacks/. [Accessed 1 May 2022].

[14] H. Chun and S. Lee, "Intelligent Virtual Assistant knows Your Life," Korea University, 28 February 2018. [Online]. Available: https://arxiv.org/pdf/1803.00466.pdf. [Accessed 30 April 2022].

[15] F. Zuiderveen Borgesius, "Discrimination, artificial intelligence, and algorithmic decision-making," Council of Europe, Directorate General of Democracy, 2018. [Online]. Available: https://pure.uva.nl/ws/files/42473478/32226549.pdf.

[16] F. J. Z. Borgesius, "Strengthening legal protection against discrimination by algorithms and artificial intelligence," 25 March 2020. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/13642987.2020.1743976. [Accessed 1 May 2022].

[17] L. Floridi, "Establishing the rules for building trustworthy AI," June 2019. [Online]. Available: https://www.nature.com/articles/s42256-019-0055-y.pdf. [Accessed 2 May 2022].

[18] M. Szmigiera, "Share of people who agree they trust artificial intelligence in 2018, by country," 19 October 2021. [Online]. Available: https://www.statista.com/statistics/948531/trust-artificial-intelligence-country/. [Accessed 1 May 2022].

[19] S. D'Alfonso, "AI in mental health," December 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352250X2030049X?casa_token=bEdEnM6UTPIAAAAA:jmOZT27mwUGJmrJW1FC2yM8cD4yO1RJ7BmfjaU5ov-bvzikdfZGzWnN5qWoHP2rBn5HxH76-fg. [Accessed 2 May 2022].

[20] "Intriguing Amazon Alexa Statistics You Need to Know in 2022," 14 February 2022. [Online]. Available: https://safeatlast.co/blog/amazon-alexa-statistics/#gref. [Accessed 1 May 2022].

[21] F. Laricchia, "Total number of Amazon Alexa skills from January 2016 to September 2019," 14 Februrary 2022. [Online]. Available: https://www.statista.com/statistics/912856/amazon-alexa-skills-growth/#:~:text=In%20a%20little%20over%20three,services%20such%20as%20Amazon%20Echo.. [Accessed 1 May 2022].

[22] S. Esposito, D. Sgandurra and G. Bella, “Alexa versus Alexa: Controlling Smart Speakers by Self-Issuing Voice Commands,” in *Proceedings of the 2022 ACM Asia Conference on Computer and Communications Security (ASIA CCS '22)*, 2022.

[23] S. Li, K.-K. R. Choo, Q. Sun, W. J. Buchanan and J. Cao, “IoT Forensics: Amazon Echo as a Use Case,” *IEEE Internet of Things Journal,* vol. 6, no. 4, pp. 6487-6497, doi: 10.1109/JIOT.2019.2906946, August 2019.

[24] J. Lau, B. Zimmerman and F. Schaub, “Alexa, Are You Listening? Privacy Perceptions, Concerns and Privacy-seeking Behaviors with Smart Speakers,” November 2018. [Online]. Available: https://dl.acm.org/doi/pdf/10.1145/3274371?casa_token=V136c35bTuAAAAAA:uzZT908UQxcghcgsuZ3eJ3kNOKnibgedbYJlkZm9Fwu91h26Z2zIqYxtMYva5tOW63oa9PPpMbTh. [Accessed 2 May 2022].

[25] A. Koeneckea, A. Namb, E. Lakec, J. Nudelld, M. Quarteye, Z. Mengeshac, C. Toupsc, J. R. Rickfordc, D. Jurafskyc and S. Goeld, “Racial disparities in automated speech recognition,” vol. 117, no. 14, pp. 7684-7689, 2020.

[26] A. Reis, D. Paulino, H. Paredes, I. Barroso, M. J. Monteiro, V. Rodrigues and J. Barroso, “Using intelligent personal assistants to assist the elderlies,” 2018. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8559503&casa_token=A27wFl9cZJ4AAAAA:MeJlwVcHzFCkyAbpdhuDhoa5xQLOzMxUSN7JLTjpWsz2Ng6SSvClWylpVySEFvIPgb6mLyNwug. [Accessed 2 May 2022].

[27] M. McTear, “Conversational Modelling for Chatbots: Current Approaches and Future Directions,” 2018. [Online]. Available: http://www.spokenlanguagetechnology.com/docs/McTear_ESSV_2018.pdf. [Accessed 2 May 2022].

[28] V. Këpuska and G. Bohouta, “Next-Generation of Virtual Personal Assistants (Microsoft Cortana, Apple Siri, Amazon Alexa and Google Home),” 2018. [Online]. Available: https://d1wqtxts1xzle7.cloudfront.net/54374216/IRJET-V4I8120-with-cover-page-v2.pdf?Expires=1651298444&Signature=G3UqOOgf~I1cJJVxzZVeti9OXzonQXKkLj73ZrYk~6D1kV0xnbd5r4RL5UozsiRrdajvvl6UVFDGl1KkZIvSTDQ7BTL0LDxsHBH6e0lo8AxxML9oIC6-ZRoLo7TvVqGE-e5ZaAKTQzLWU9. [Accessed 30 April 2022].

[29] J. Beskow, C. Peters, G. Castellano, C. O’Sullivan, I. Leite and S. K. (Eds.), “Intelligent Virtual Agents,” in *17th International Conference, IVA 2017*, Stockholm, 2017.