Addressing Accountability in Highly Autonomous Virtual Assistants

Fernando Galdon[1(✉)] and Stephen Jia Wang[2]

[1] Department of Global Innovation Design, School of Design, Royal College of Art, London, UK
fernando.galdon@network.rca.ac.uk
[2] Department of Innovation Design Engineering, School of Design, Royal College of Art, London, UK

Abstract. Building from a survey specifically developed to address the rising concerns of highly autonomous virtual assistants; this paper presents a multi-level taxonomy of accountability levels specifically adapted to virtual assistants in the context of Human-Human-Interaction (HHI). Based on research findings, the authors recommend the integration of the variable of accountability as capital in the development of future applications around highly automated systems. This element inserts a sense of balance in terms of integrity between users and developers enhancing trust in the interactive process. Ongoing work is being dedicated to further understand to which extent different contexts affect accountability in virtual assistants.

Keywords: Human factors · Human-systems integration · Systems engineering · Trust · Virtual assistant · Highly automated systems · Autonomy · Automation · Accountability

1 Introduction

With the rise of highly autonomous systems concerns in the field of human factors are focusing on designing appropriate tools to address this new class of technology [1]. Recent investigations, such as MIT's research paper on who should kill a self-driving car, are pointing to ethical decision making in the context of highly autonomous systems (HAS) [2]. Furthermore, due to its persuasive capabilities, concerns are also rising in the area of virtual assistants (VA) with the introduction of Duplex and Alexa by Google and Amazon. In this context, Amazon has recently filed a patent to transform its systems into a doctor diagnosing and providing treatment in the process [3]. Further innovations are transforming the VA into legal or financial advisers, set up romantic dates and provide jobs. They will engage with us, and by combining and inferring preliminary knowledge and in situ interaction they will have the potential and capability to change our preliminary decisions and take actions on our behalf on highly sensitive areas such as health and wellbeing, identity, social interactions or economically related activities. However, one fundamental question remains, if something goes wrong, who should be accountable for the action?

As we are moving into a machine-human paradigm questions of accountability remain unsolved. This fact positions highly autonomous systems at the centre and

T. Ahram et al. (Eds.): IHIET 2019, AISC 1018, pp. 10–14, 2020.
https://doi.org/10.1007/978-3-030-25629-6_2

re- search must try to address the implications of trust from their perspective [4]. Traditionally, accountability in complex autonomous virtual assistants has been out of research due to the automated nature of the interactions. They were based on one-off query focused on non-dangerous outcomes such as playing songs. Nowadays, as systems become more autonomous and unsupervised, the potential outcomes of these interventions are probing capital for the successful development and implementation of these systems in society. Humans, therefore, must be considered as a key component in even a fully automated system [5], and implement a human-centred approach to tackle this challenge [6].

Recent research in the area of robustness in HAS shows 0% adversarial accuracy when evaluating a deep network against stronger adversaries [7, 8]. In order to address this problem they are using interval bound propagation [9, 10, 11] to great success. However, as the researcher acknowledge "no amount of testing can formally guarantee that a system will behave as we want. In large-scale models, enumerating all possible outputs for a given set of inputs…is intractable due to the astronomical number of choices for the input perturbation" [12].

In the context of unsupervised HAS constantly evolving we must change our approach. Instead of talking about transparency we have to start talking about trust and accountability as an a priori and a posteriori elements to address. Although I agree that preventive strategies must be seen as the preferred area of intervention, systems of accountability must be put in place to address errors and failures in the system.

In this paper the authors minds the warning and propose a human-centred approach aimed at ensuring that these highly automated interactions remain focused on the user's interests and protection.

2 Method

Research into the area of automation present levels as a tool to address trust in automated systems. In this context gradient-base models of approximation has been embodied through the concept of scales or Level of trust (LoT). This approach of different levels of automation has been persistent in the automation literature since its introduction by Sheridan and Verplanck [13]. Kaber [14] emphasises that levels of automation (LoA) is a fundamental design characteristic that determines the ability of operators to provide effective oversight and interaction with system autonomy. According to Endsley [15], although they represent a simplification of reality, they provide a tool. This method has proven successful in providing a solid foundation to understand HAI at a deeper level. This is highly relevant when confronting an invisible entity making decisions while working in the background.

Levels aim to improve transparency by simplifying interactions. In this context, transparency refers to the extent to which the actions of the automation are understandable and predictable [15]. According to research in the area, automated systems which clarify their reasoning are more likely to be trusted [16, 17, 18].

In this context, a multi-level taxonomy of levels of accountability specifically designed to address the increasing autonomy of highly automated virtual assistants was designed by the authors. It integrated a gradient spectrum of levels ranging from the

system to the user. It was structures in four distinctive levels; the platform hosting all the interactions (Level 1), the developer/designer designing the actions/skills algorithms (Level 2), due to its unsupervised and evolutive nature I decided to include the algorithm performing the action (Level 3), and finally, the user (Level 4) (Table 1).

Table 1. Proposed levels of accountability.

Levels	Subject	Explanation
Level 1	Platform	*The company who owns the platform*
Level 2	Developer	*The designer who designed the action*
Level 3	Algorithm	*The algorithm performing the action*
Level 4	User	*The user performing the action*

3 Discussion

Due to the highly contextual nature of virtual assistants, a co-design workshop with students from the Royal College of Art underpinned four highly sensitives areas where highly automated VAs may impact significantly users; health and wellbeing, identity, economically related activities and social interactions.

From the areas aforementioned and based on demos, patents and prototypes, eight case studies were built to address different outcomes. Two cases addressed each sensitive area ranging from low to high impact. Then, a survey was designed to establish whether the proposed levels of accountability in highly automated virtual assistants were sufficient to address all the cases.

To test the scale, the main technique consisted on integrating an *other* tab in each case. This space allowed the participant to propose a new level or area missing, questioning the existing scale in the process. Participants engaged with the *other* tab though the survey at different points.

50 participant, 21 men, 27 women and 2 who didn't want to identify themselves, from 14 different countries with an age range between 18–67 years old from different professions have undertaken the survey (Table 2).

Table 2. Survey results.

	Unhappy service medicine	Unhappy service newspaper	Ends in violence addiction	Ends in violence raped	Wrong prediction sexuality	Wrong prediction jailed	Loses money	Loses job	Total
Level 1 Platform	20%	26%	12%	6%	14%	22%	18%	8%	15.75%
Level 2 Designer	18%	**28%**	18%	14%	22%	24%	32%	8%	20.50%
Level 3 Algorithm	16%	20%	8%	6%	**38%**	**30%**	16%	8%	17.75%
Level 4 User	**38%**	24%	**56%**	**38%**	24%	22%	**34%**	**74%**	38.75%
Other	8%	2%	6%	36%	2%	2%	0%	2%	7.25%

In average, 38, 75% of participants found that the user is the main element accountable for unexpected consequences. This result was highly unexpected as in all cases the VA was initiating and doing actions on behalf of the user. The designer/developer is placed second with 20.50% of participants pointing them as accountable. Third position is for the algorithm with 17.75%. Finally, the platform would be the least accountable level with 15.75% of the participants. Other elements (third-parties) add the other 7.25%.

4 Conclusion

The survey aimed to understand whether or not contexts and actions affected the level of accountability. In the main area of levels of accountability, contexts and actions play a role in determining which level of accountability is needed. In addition, they did play a role in determining the spectrum. These elements demanded the integration of a third-party level in the scale. However, at the same time, a generic granular scale of 5 levels covering from platform to user and integrating third-parties is capable of addressing different contexts and actions in highly automated virtual assistants.

Finally, if we combine Level 1 (platform), 2 (designer), and 3 (algorithm) 54% of participants place accountability in the system's side. Therefore, inserting the accountability variable in the design process is capital for the correct integration of Highly Automated Systems in society, as this element inserts a sense of balance in terms of integrity between users and developers enhancing trust in the interactive process (Table 3).

Table 3. Final levels of accountability.

Levels	Subject	Explanation
Level 1	Platform	*The company who owns the platform*
Level 2	Developer	*The designer who designed the action*
Level 3	Algorithm	*The algorithm performing the action*
Level 4	Third party	*A third party affecting the service*
Level 5	User	*The user performing the action*

References

1. Hancock, P.A.: Imposing limits on autonomous systems. Ergonomics **60**(2), 284–291 (2017)
2. Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.F., Rahwan, I.: The moral machine experiment. Nature (2018). https://doi.org/10.1038/s41586-018-0637-6
3. Jin, H., Wang, S.: Voice-based determination of physical and emotional characteristics of user (2018). http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO2&Sect2=HITOFF&u=%2Fnetahtml%2FPTO%2Fsearch-adv.htm&r=1&p=1&f=G&l=50&d=PTXT&S1=10,096,319&OS=10,096,319&RS=10,096,319

4. Ortega, B.P.A., Maini, V.: Building safe artificial intelligence: specification, robustness, and assurance specification: de ne the purpose of the system. Medium. Retrieved from https://medium.com/@deepmindsafetyresearch/building-safe-artificial-intelligence-52f5f75058f (2018)
5. Wang, S.J.: Fields interaction design (FID): the answer to ubiquitous computing supported environments in the post-information age. Homa & Sekey Books (2013)
6. Wang, S.J., Moriarty, P.: Big Data for Urban Sustainability. Springer (2018)
7. Athalye, A., Carlini, N., Wagner, D.: Obfuscated gradients give a false sense of security: circumventing defenses to adversarial examples. Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden, PMLR 80 (2018)
8. Uesato, J., O'Donoghue, B., van den Oord, A., Kohli, P.: Adversarial risk and the dangers of evaluating against weak attacks. Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden, PMLR 80 (2018)
9. Ehlers, R.: Formal verification of piece-wise linear feed-forward neural networks. BT - Automated Technology for Verification and Analysis - 15th International Symposium, ATVA 2017, Pune, India, October 3–6, 2017, Proceedings (2017)
10. Katz, G., Barrett, C.W., Dill, D.L., Julian, K., Kochenderfer, M.J.: Reluplex: an efficient SMT solver for verifying deep neural networks. BT - Computer Aided Verification - 29th International Conference, CAV 2017, Heidelberg, Germany, July 24–28, 2017, Proceedings, Part I (2017)
11. Mirman, M., Gehr, T., Vechev, M.: Differentiable abstract interpretation for provably robust neural networks. Proceedings of the 35th International Conference on Machine Learning, in PMLR 80, pp. 3578–3586 (2018)
12. Kohli, P., Gowal, S., Dvijotham, K., Uesato, J.: Towards robust and verified AI: specification testing, robust training, and formal verification. Deepmind. Medium 28 March 2019. https://deepmind.com/blog/robust-and-verified-ai/ (2019). Accessed 29 Mar 2019
13. Sheridan, T.B., Verplank, W.L.: Human and computer control of Undersea teleoperators: Fort Belvoir, VA: Defense Technical Information Center (1978). https://doi.org/10.21236/ADA057655
14. Kaber, D.B.: Issues in human-automation interaction modeling: presumptive aspects of frameworks of types and levels of automation. J. Cogn. Eng. Decis. Making **12**(1), 7–24 (2018). https://doi.org/10.1177/1555343417737203
15. Endsley, M.R.: From here to autonomy: lessons learned from human–automation research. Hum. Factors: The J. Hum. Factors Ergon. Soc. **59**, 5–27 (2017). https://doi.org/10.1177/0018720816681350
16. Simpson, A., Brander, G.N., Portsdown, D.R.A.: Seaworthy trust: confidence in automated data fusion. In: Taylor, R.M., Reising, J. (eds.) The Human-Electronic Crew: Can We Trust the Team, pp, 77–81. Hampshire, UK: Defence Research Academy. Retrieved from http://www.dtic.mil/dtic/tr/fulltext/u2/a308589.pdf (1995)
17. Hoff, K.A., Bashir, M.: Trust in automation: integrating empirical evidence on factors that influence trust. Hum. Factors: The J. Hum. Factors Ergon. Soc. **57**, 407–434 (2015)
18. Galdon, F., Wang, S.J.: Designing trust in highly automated virtual assistants: a taxonomy of levels of autonomy. International Conference on Industry 4.0 and Artificial Intelligence Technologies. Cambridge, UK. ISBN: 978-1-912532-07-0 (2019)