From Apology to Compensation: A Multi-level Taxonomy of Trust Reparation for Highly Automated 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. This paper presents a multi-level taxonomy of reparation levels specifically adapted to virtual assistants in the context of Human-Human-Interaction (HHI) with a specific focus on maintaining trust in the system. This taxonomy ranges from current models of apology to the newly integrated compensation area via a range of case studies specifically developed to address the rising concerns of unsupervised interactions in the context of Virtual Assistants (VA). Based on preliminary research, the author recommends the integration of reparation strategies as a fundamental variable in the ongoing development of VAs, as this element inserts a sense of balance in terms of vulnerability between users and developers to enhance trust in the interactive process. Present and future work is being dedicated to further understand how different contexts may affect integrity in highly automated virtual assistants.

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

1 Introduction

Recent developments of machine learning and deep learning are leading the rise of a new type of technology; highly automated systems (HAS). In this context, experts in the field of human factors [1] and data related human factors [2, 3] are calling for the development of tools and strategies to address the implications of autonomy in this type of systems. In this area, Virtual Assistants (VA) such as Google's Duplex are raising concerns due to an extraordinary level of autonomy, fluency and interactivity never seen before.

Literature in the area of automation are calling for the development of reparation strategies [4–6]. These strategies are becoming capital not only to address engagement but to maintain trust in these systems. According to research in the area, VAs need to generate less than 30% of errors, otherwise, the user would stop using them [7–9]. As these systems become more autonomous, ubiquitous and unsupervised, the

T. Ahram et al. (Eds.): IHIET 2019, AISC 1018, pp. 42–46, 2020.
https://doi.org/10.1007/978-3-030-25629-6_7

development of reparation techniques becomes fundamental for the adequate development and integration of these systems in society.

Traditionally, these papers focus on different types of reparations such as apologies or denials and the timing to deliver them. Bansal and Zahedi [10] investigated how trust may be rebuilt after it is violated by negative events in data privacy, including the efficacy of the three most used response types—apology, denial and no response. After conducting controlled experiments their results showed that apology emerged as a universally effective response, although its reparative power was far less effective in unauthorised sharing than in hacking. Denial emerged as a complex response and as a very negative approach. Finally, they also reported that is critical to investigate the typology of violation events.

Their research was groundbreaking, however, it was based on current models of automated VAs which are equipped with the capacity of responding to queries. However, with the emergence of highly autonomous systems such as *Duplex* (capable of getting the initiative of the interaction and establishing and maintaining conversations) and recent patents by Amazon to transform the VA into a medical adviser, it seems that reparation strategies around apology become limited in scope. In this paper the authors mind this evolution and propose a human-centred approach aimed at ensuring that these highly automated interactions remain focused on the user's interests and protection. In this scenario, the authors have structured a scale and integrated a gradation of compensation levels to test whether they could account for the type of interactions [3] emerging from highly automated virtual assistants.

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 have 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 [11]. Levels of automation (LoA) is acknowledged by Kaber as a fundamental design characteristic that determines the ability of operators to provide effective oversight and interaction with system autonomy [12]. According to Endsley, although Levels present a simplification of reality, they provide a system by which different stakeholders understand the full scope of the system at hand [13]. This method has proven successful in providing a solid foundation to understand automated systems 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 [13]. According to research in the area, automated systems which clarify their reasoning are more likely to be trusted [14–16].

In this context, a multi-level taxonomy of levels of reparation specifically designed to address the increasing autonomy of highly automated virtual assistants was designed by the lead author. It integrates a gradient spectrum ranging from no apology to high compensation. It is structured in seven distinctive levels organised in three main areas:

no apology (Level 1), a triple gradient around apology (Level 2, 3 and 4), and a triple gradient around compensation (Level 5, 6 and 7) (Table 1).

Table 1. Levels of reparation.

Levels	Subject	Explanation
Level 1	No apology	*The company who owns the platform*
Level 2	Basic apology	*The designer who designed the action*
Level 3	Generic apology	*The algorithm performing the action*
Level 4	Public apology	*The user performing the action*
Level 5	Low compensation	*Between 0$ and 100000$*
Level 6	Medium compensation	*Between 100000$ and 1 Million $*
Level 7	High compensation	*+ 1 Million $*

3 Discussion

Although the responsibility to design the scale remains in the designer, Kaber calls for empirical studies to address Level of Automation (LoA) [10]. Due to the highly contextual nature of VAs, a co-design workshop with students from the Royal College of Art underpinned. On one hand, four highly sensitives areas where high-level automated VAs may impact users significantly; health and wellbeing, identity, economically related activities and social interactions. On the other hand, four major unintended consequences; unhappy services, wrong predictions, unintended loses related to the service and an action unexpectedly ending violently.

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

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).

In average, 48,75% of participants found that no reparation is the best option when the system generate an unintended consequence in highly sensitive areas. Thus, placing the responsibility in the user. This result was unexpected as the initiative of the action was placed in the system. 23.25% of participants would accept some sort of apology. Finally, 23.00% in average would demand some sort of compensation to repair trust in

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 No reparation	36%	48%	56%	64%	46%	56%	40%	44%	48.75%
Level 2 Basic apology	10%	8%	10%	6%	10%	4%	4%	6%	7.25%
Level 3 pers. Personal apology	22%	12%	2%	0%	24%	2%	2%	6%	8.75%
Level 4 Public apology	6%	20%	6%	4%	4%	4%	6%	8%	7.25%
Level 5 low compensation	6%	2%	4%	4%	2%	0%	10%	4%	4.00%
Level 6 medium compensation	6%	0%	10%	2%	10%	8%	14%	18%	8.50%
Level 7 full compensation	4%	6%	10%	12%	2%	20%	20%	10%	10.50%
Other	10%	4%	2%	8%	2%	6%	4%	4%	5.00%

the system. If we combine reparation strategies (apology and compensation) a total of 46.25% of participants would request them to keep using the system. Other elements such as third-parties were not present in the other tab. Responses in this area demanded combinations of apologies and compensation to repair trust in the system. None of the participants demanded a new level.

4 Conclusion

The survey aimed to understand whether or not contexts and actions affected the level of reparation. They partially determine the level of reparation, however, they did not play a role in determining the spectrum. A generic granular scale of 7 levels covering from no reparation to high compensation would be capable of addressing different contexts and actions in highly automated virtual assistants.

From the survey conducted, contexts (highly sensitive areas) and actions (unintended consequences) play a role in determining user engagement. The 50/50 split presented by this research presents an empirical need for approaching the design of these system equally from a preventive *a priori* strategies to reparative *a posteriori* strategies. Therefore, inserting the reparation variable in the design process is as important as preventing strategies for the correct integration of Highly Automated Virtual Assistants in society.

Based on the results, the author recommends the integration of reparation strategies as a fundamental variable in the ongoing development of virtual assistants, as this element inserts a sense of balance in terms of vulnerability between users and developers enhancing trust in the interactive process.

References

1. Hancock, P.A.: Imposing limits on autonomous systems. Ergonomics **60**(2), 284–291 (2017). https://doi.org/10.1080/00140139.2016.1190035
2. Wang, S.J., Moriarty, P.: Big Data for Urban Sustainability. Springer (2018)
3. Wang, S.J.: Fields Interaction Design (FID): The Answer to Ubiquitous Computing-Supported Environments in the Post-Information Age. Homa & Sekey Books, Paramus (2013)
4. Bottom, W.P., Gibson, K., Daniels, S.E., Murnighan, J.K.: When talk is not cheap: substantive penance and expressions of intent in rebuilding cooperation. Organ. Sci. **13**(5), 497–513 (2002)
5. Kim, P.H., Ferrin, D.L., Cooper, C.D., Dirks, K.T.: Removing the shadow of suspicion: the effects of apology versus denial for repairing competence-versus integrity-based trust violations. J. Appl. Psychol. **89**(1), 104 (2004)
6. Kohn, S.C., Quinn, D., Pak, R., de Visser, E.J., Shaw, T.H.: Trust repair strategies with self-driving vehicles: an exploratory study. Proc. Hum. Factors Ergon. Soc. Annu. Meet. **62**(1), 1108–1112 (2018). https://doi.org/10.1177/1541931218621254
7. Parasuraman, R., Sheridan, T.B., Wickens, C.D.: A model for types and levels of human interaction with automation. IEEE Trans. Syst. Man Cybern. A Syst. Hum. **30,** 286–297 (2000)
8. Wickens, C.D., Dixon, S.R.: The benefits of imperfect diagnostic automation: a synthesis of the literature. Theor. Issues Ergon. Sci. **8**, 201–212 (2007)
9. Wang, L., Jamieson, G.A., Hollands, J.G.: Trust and reliance on an automated combat identification system. Hum. Factors **51**, 281–291 (2009)
10. Bansal, G., Zahedi, F.M.: Trust violation and repair: The information privacy perspective. Decis. Support. Syst. **71**(2015), 62–77 (2015)
11. Sheridan, T.B., Verplank, W.L.: Human and Computer Control of Undersea Teleoperators. Defense Technical Information Center, Fort Belvoir, VA (1978). https://doi.org/10.21236/ADA057655
12. Kaber, D.B.: Issues in human-automation interaction modeling: presumptive aspects of frameworks of types and levels of automation. J. Cogn. Eng. Decis. Mak. **12**(1), 7–24 (2018). https://doi.org/10.1177/1555343417737203
13. Endsley, M.R.: From here to autonomy: lessons learned from human–automation research. Hum. Factors: J. Hum. Factors Ergon. Soc. **59**, 5–27 (2017). https://doi.org/10.1177/0018720816681350
14. 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. Defence Research Academy, Hampshire, UK (1995). http://www.dtic.mil/dtic/tr/fulltext/u2/a308589.pdf
15. Hoff, K.A., Bashir, M.: Trust in automation: integrating empirical evidence on factors that influence trust. Hum. Factors: J. Hum. Factors Ergon. Soc. **57**, 407–434 (2015)
16. Galdon, F., Wang, S.J. (2019). Designing trust in highly automated virtual assistants: a taxonomy of levels of autonomy. In: International Conference on Industry 4.0 and Artificial Intelligence Technologies. Cambridge, UK. ISBN: 978-1-912532-07-0