Designing trust in virtual assistants: A taxonomy of levels of autonomy

Fernando Galdon, Ashley Hall and Stephen Jia Wang

Abstract This paper will present a guiding framework and a multi-level taxonomy of automation levels specially adapted to Virtual Assistants in the context of HHI. This trust-based framework incorporates interaction phases, trust-affecting design principles and design techniques. it also introduces a taxonomy of levels of autonomy explaining each level from a trust perspective. Based on the research insights, the author recommends designers to combine a holistic perspective on trust with contextual awareness, to be able to integrate the impact of contexts on interactions.

1 Introduction

Recently, Peter Hancock presented a warning to the field of human factors in which attention must be given to the appropriate design of a new class of technology: highly autonomous systems [1]. In the context of using such an autonomous system as Virtual Assistants (VA), this warning has become factual with a demo presented by Google called Duplex. This VA system demo presented an extraordinary level of fluidity, coherence, and autonomy never seen before.

With the development and combination of machine learning and deep learning techniques, a new paradigm is raising: Machine-Human interaction (MHI). In this paradigm the technology is holding the initiative of the communication. This approach positions highly autonomous systems at the centre and tries to address the implications of trust from their perspective [2]. However, the nascent nature of these systems and the unavailability of them to conduct research prevents an adequate development of strategies. Traditionally, complex autonomous systems required the

Fernando Galdon
School of Design, Royal College of Art, e-mail: fernando.galdon@network.rca.ac.uk,

Ashley Hall
School of Design, Royal College of Art e-mail: ashley.hall@rca.ac.uk

human operator to appropriately calibrate their trust in the automation to achieve performance and safety goals. In this context, literature has focused on the human-machine- interaction (HMI), and human-human-interaction (HHI) trust paradigms to precisely define and measure trust in automation. In this article, the author minds the warning and propose a human centred approach in the context of HHI directly aimed at ensuring that emerging highly autonomous systems interactions remain focused on the user's needs and preferences in the context of Virtual Assistants.

As we are placing this debate in the context of virtual assistants, We can observe a clear distinction by conducting a comparative study among Alexa and Duplex. The usability of Alexa is based on a one off query [3]. The system has the ability to stream audio over Bluetooth, request radio stations, play music, make lists, ask about the weather and news, and order products from Amazon.com [3]. On the other hand, Duplex is a system presenting an extraordinary level of fluidity, coherence, and autonomy never seen before. In a demo introduced by Google in May 2018, the system was able to make a hair appointment with a human without any supervision. This evolution represents a transition from automation to autonomy. This system, not only was able to deliver the task but did it without the human noticing that she was speaking to a robot. This system differs radically from Alexa in the sense that we are moving from one-off queries to conversations. And that the initiative in the interaction is not necessarily placed in the user but within the system. Finally, in the context of virtual assistants, Alexa recorded a private conversation and sent it to other users without the main user knowing [4]. The unpredictability on how these unsupervised agents may evolve and its goal-oriented nature present a fundamental problem for society and businesses. Is this regard a fundamental question arises; what strategies would enable us to establish and maintain trust in these systems?

1.1 From Machine-Human to Human-Human

Three main paradigms emerge on how to approach trust in Highly Automated Systems; human-machine-Interaction (HMI), human-human-Interaction (HHI) and an emerging machine-human-interaction perspective (MHI).

From an HMI perspective Interpersonal trust conceptualizations only provide a limited explanation. Distinct psychological constructs and mechanisms need to be considered to explain Human–Agent trust. In this paradigm, Anthropomorphised agents are related to lower initial trust. Computer systems are believed to be more capable, rational and objective than humans. (Automation bias: authority hypothesis) [5].

From an HHI perspective the same psychological constructs and mechanisms can be applied to explain interpersonal and Human–Agent trust. In this paradigm, Computers are social actors (media-equation hypothesis). Form a design perspective,

Anthropomorphised agents are related to higher initial trust because Anthropomorphism makes novel systems more familiar and controllable [6].

Finally, in a recent article posted in Medium by the DeepMind safety research team, they advocated focussing on a Machine-Human interaction (MHI) perspective. Their approach implies a superiority of the system to the agent. This strategy is focused on the autonomous system and design strategies are focused on how to control the system [2].

Although this emerging MHI perspective will be the dominant paradigm in the future, several reasons prevent us from adopting this paradigm nowadays; the problems of interpretability (the system does not know what is doing) and interruptibility (the system knowing when to stop) in addition to the nascent nature of this systems and the unavailability of them to conduct research prevent researchers to adopt this paradigm. In terms of HMI perspective, this was the paradigm used until 2017. The fundamental reason for the evolution to an HHI perspective is the nature of neural networks capable of learning form users interactions. In this context, de Visser proposes that we should approach the design of highly autonomous system from an HHI perspective by moving from focusing on interactions to focusing on relationships [7]. Consequently, In this paper, the authors positions the investigation from an HHI perspective.

2 Factors for designing trust in automation

Current models structure trust in automation from three different perspectives; the user, the environment and the system [8][9]. In this context, we can synthesise three main interactive stages; expectations; experimentation and reliability.

Expectations depend on preliminary knowledge, recommendation by relatives or endorsement by celebrities and appearance design attributes such as typological design or name. Experimentation is focused on design attributes such as communication style, ease-of-use or transparency-feedback. Elements such as pitch and porosity, intonation and wording or the system sounding comfortable and natural define its communication style. Fluidity and autonomy, the uselessness of recommendations and low error-rates define its ease-of-use and transparency and the communication of intend define its transparency and feedback level.

Reliability focuses on design strategies for reducing error-rates within an automation system fundamentally based on stages and levels of automation build around calibration systems. Research in the area of human factors present evidences that the more reliable the system, the more likely is to be trusted [10-11]. Therefore, positioning this area as the most relevant for building and establishing trust in HAS. In this context, as far as the error-rate is around 30 per cent or less, users will continue

using the automated system [12].

In terms of interaction, when users interact with automated systems misuse and disuse are the most common outcomes [13]. When the system fails less than 30 per cent it leads to user's misuse. Misuse refers to cases when the automation is used without skepticism, tending to result in overuse [10-11]. The main implications are; automation bias and complacency [14][10]. They fundamentally arise due to a lack of monitoring where lack of attention plays a central role [10].

On the other hand, when the system fails more than 30 per cent it leads to user's disuse. Disuse relates to an interaction that extends from the user barely using the automation to completely abandoning it in favour of a manual approach [13]. The main reasons are a high expectation of automation performance and unexpected errors [11][15].

Finally, Hoff, K. A., and Bashir, M. presented in a seminal paper the five fundamental principles to account when designing trust in automation; appearance (P1), Communication style (P2), ease of use (P3), transparency and/or feedback (P4) and Levels of control (P5) [9].

2.1 Two challenges: Reliability and Predictability

As autonomous systems become more and more complex, the ability of users to understand the system becomes more difficult. The higher the levels of automation, the more complex and less understandable the system, the lower the levels of trust [16]. In this context, reliability and predictability have been identified as a key factor influencing trust in automation [8]. Therefore, in order to address trust in highly automated systems, Trust must be appropriately calibrated to the actual system performance [17].

In the context of reliability, predictability has been identified as a fundamental quality for trust in automated systems. It is argued that prediction is necessary to mitigate potentially detrimental interaction behaviour to avoid unwanted results which may result in situations that cannot be changed [13].

In this context, for the system to enhance reliability, the calibration system must enhance predictability. In Predictability, prior knowledge about potential automation failures reduces the level of uncertainty and risk [16]. Once reliability has been judged, the most important factor of trust in automation is predictability of performance over time [18].

Predictability is enhanced by implementing levels of automation. The idea of gradient-base models of approximation with positive, negative and neutral spectrums

has been embodied through the concept of scales or Level of trust (LoT). The notion of different levels of automation has been persistent in the automation literature since its introduction by Sheridan [19]. Kaber [20] points out 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. 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 [21]. Automated systems which clarify their reasoning are more likely to be trusted [22][9].

In this context, prominent frameworks in the area of levels of automation (LoA) are for instance Parasuraman, Sheridan and Wickens [12]. These researches present a framework which differs radically from past approaches. Instead if structuring a scale, they propose a 4 levels structure outlining four classes or types of automatons functions to account in human-machine-interaction. Wickens et al. [23] degrees of automation proposes a similar approach to Parasuraman with a small addition of the notion of degrees (high and low). An approach more related to Sherindan [19] is presented by Westin, C., hilburn, B and Borst, C. [24]. They present a seven-point scale ranging from total human control to total autonomy in the context of air traffic management. Marinik et al. [25] Multi-variable framework integrates both approaches; stages and levels. This framework is widely used in current vehicle autonomy research.

As we have reviewed, a range of models of levels of automation (LoA) exits in the area, however, no scale has been designed specifically to address Virtual Assistants in its evolutive nature leading to an increasing level of autonomy. This paper will present the first level of automation designed ad-hoc to address present and near-future evolutions of these systems towards increased autonomy.

3 Method

Scales range from one to ten points. The most common types are odd or uneven scales which allow the participant to record a neutral trust level. The most used model is a seven-point scale traditionally articulated to measure global trust in automation.

Endsley [21] arguments that the most important benefit in LoA is its communicative value to key stakeholders (e.g., system operators, designers, and program managers) about the intrinsic notion that there are different ways and degrees to implement automation. The fact that there is a whole range of options between fully manual and fully automated enhances the understanding of these systems by non-experts. This method has proven successful in providing a solid foundation to understand HAI at a deeper level. This is highly relevant when confronting an invis-

ible entity making decisions while working in the background.

In this context, Kaber [20] points out, the LoA decision and design must be made by the system developers. Including adaptive automation, a granularity of control, and automation interface design, LoA is a fundamental design characteristic that determines the ability of operators to provide effective oversight and interaction with system autonomy. In this context, LoA remains a central design decision associated with the design of automated and autonomous systems that must be addressed in future system design.

3.1 Designing levels of automation for highly automated virtual assistants

Building from these insights, this study proposes the articulation of an odd scale. This type of scale proposes a neutral element and two extremes which allocate extreme perspectives. In this case no autonomy and full autonomy. As we have seen in the section before, taxonomies are structured between five to ten point scales. However, ten point scales overlap intentionalities in automation. Whereas five point scales seem too narrow to cover the range of possible interactions. They only provide a step in between the neutral and the end extreme measures. In this context, and following these insights, this study proposes a seven point scale which allocates two steps in-between (Figure. 2). It uses Sherindan [19] as its foundation to adapt the scale to Virtual Assistants and the increasing level of autonomy that is expected to evolve in future developments.

3.2 Proposed levels of automation

Please see below the proposed levels of automation (Figure 2).

LEVEL 1	NO AUTONOMY	The VA does not implement the action unless requested by the user
LEVEL 2	ASSISTANCE	The VA assist determining a range of options related to user's query.
LEVEL 3	PARTIAL AUTONOMY	The VA engage in conversation and suggests one option.
LEVEL 4	CONDITIONAL AUTONOMY	The VA selects action and implements it if human approves.
LEVEL 5	RELATIONAL AUTONOMY	The VA selects action, informs human with plenty of time to stop.
LEVEL 6	HIGH AUTONOMY	The VA can perform decisions solely on its own and necessarily tells human what it did
LEVEL 7	FULL AUTONOMY	The VA can perform decisions solely on its own without reporting to the user.

Fig. 1 Proposed Multi-level taxonomy of levels of autonomy for highly automated virtual assistants. Fernando Galdon

4 Discussion

Due to the highly contextual nature of virtual assistants, a preliminary investigation underpinned four highly sensitives areas where highly automated virtual assistants may impact significantly users; health and wellbeing, identity, economically related activities and social interactions.

Once the relevant contexts were identified, a workshop was conducted with design students at the Royal college of Art to map unintended consequences in these highly sensitive areas;

1. Health - death, harm, injury, dependency, addiction
2. Social - lack of diversity, privacy, dependency, accuracy, stalkers, hackers
3. Identity - segregation, insolation, manipulation, homogeneity, dependency
4. Economics - homogeneity, dependency, manipulation, lack of diversity, indoctrination, unethical investments

From this activity four main categories of unintended consequences emerged;

1. Unhappiness about the service - unethical investments, indoctrination, manipulation, addiction
2. Wrong predictions - accuracy
3. Losing something - dependency, privacy (stalkers, hackers), segregation, insolation, addiction, indoctrination, manipulation, homogeneity, lack of diversity.
4. Ends violently - death, harm, injury, addiction

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

The primary technique consisted on integrating an ‘other’ tab to test the scale. This space allowed participants to propose a missing new level (or area). The survey was structured around three sections addressing different aspects of trust. Each section contained the eight cases with their correspondent levels and the ‘other’ tab. One of the sections was designed to enhance an ‘other’ output by limiting the levels from seven to four. 44 per cent Participants engaged with the ‘other’ tab through the survey at different points. Whereas 66 per cent of the participants engaged with the proposed levels as they felt they were sufficient to address the proposed interactions.

Finally, a co-design activity was implemented with students of design at the Royal college of Art to test the framework from a different perspective. In this case participants assumed the role of an ethicist. Presented with the cases used for the survey, They were asked to push their imagination to the limit and came up with the most unexpected possible outcomes they could imagine. Then, they were presented

with the scale of levels of automation and asked whether this scale would be sufficient to cover the situation they had envisaged or a new level was needed. The outcomes would test the framework from a third perspective. The fundamental question was, can the proposed levels of automation address the most unexpected events imagined?

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 and neither of them proposed a new level of automation. Besides, the co-design activity generated 24 variations of unintended consequences and all of them could be covered by the framework proposed.

5 Conclusion

The survey (Table. 1) aimed to understand whether or not contexts affected the level of automation. In the central area of levels of autonomy, contexts play a role in determining which level of automation was needed, but it did not play any role in determining the spectrum. A generic scale covering from no automation to full automation is capable of addressing different contexts in highly automated virtual assistants. However, contexts determine which is the most appropriate level.

In terms of contexts affecting levels of trust in autonomy, the most susceptible context for users was Identity with a 54 per cent blended average; the system being capable of predicting sexual or political orientation led to 60 per cent of participants

AUTONOMY									
AREA	PREDICTING cough	PREDICTING depression	PREDICTING politics	PREDICTING sexuality	PREDICTING investment	PREDICTING best job	PREDICTING dating partner	PREDICTING domestic violence	TOTAL
LEVEL 1 No autonomy	24%	20%	**48%**	**60%**	**16%**	10%	**38%**	10%	28.25%
LEVEL 2 Assistance	24%	24%	16%	20%	12%	**32%**	16%	14%	19.75%
LEVEL 3 Partial autonomy	**26%**	**30%**	20%	8%	22%	18%	26%	**16%**	20.75%
LEVEL 4 Conditional autonomy	18%	10%	6%	4%	**34%**	28%	12%	14%	15.75%
LEVEL 5 Relational autonomy	6%	14%	8%	8%	8%	6%	2%	**16%**	8.50%
LEVEL 6 High autonomy	2%	2%	2%	0%	4%	2%	0%	**16%**	3.50%
LEVEL 7 Full autonomy	0%	0%	0%	0%	4%	2%	4%	12%	2.75%
OTHER ..	0%	0%	0%	0%	0%	2%	2%	2%	0.75%

Table 1 Table 1. Survey results. Fernando Galdon

in the first case and 48 per cent in the second to opt for Level 1 (No autonomy). Second most susceptible contexts were Health and well being and social interactions with 22 per cent and 24 per cent blended average respectively. However, the most distributed result was on the system predicting domestic violence. This is the only context where more than 10 per cent of the participants would use high or full autonomy. Participants preferred to have total control of the system. Level 1 is the preferred option on average. levels 2 and 3, account for 40.50 per cent of the participants preferring to be in control of the autonomous system. If we combine levels 1, 2, and 3; This presents an average of 68.75 per cent of participants demanding the initiative. The neutral level (level 4) is preferred by 15.75 per cent of the participants on average. On Levels where the initiative resides on the system (levels 5, 6, and 7), only 14.75 per cent of participants would decentralise their decision. Domestic violence is the only context were more than 10 per cent of participants would allow the initiative in the system on levels 5, 6, 7. Depression is the other area were participants would go beyond 10 per cent, but only in level 5.

6 Future work

Although several models address the nature and practice of automation systems, models in automation leading to autonomy specifically designed for Virtual assistants focusing on trust remained unsolved. The model presented is a first step in building a system capable of building and maintaining trust in Highly Automated Systems (HAS), however, other areas such as reparation and accountability must be further investigated. Future work will focus on these areas.

References

1. Hancock, P. A. (2017). Imposing limits on autonomous systems. Ergonomics, 60(2), 284–291. DOI: 10.1080/00140139.2016.1190035
2. Ortega, B. P. A. (2018). Building safe artificial intelligence; specification , robustness , and assurance. design the purpose of the system. Medium. Retrieved from https://medium.com/@deepmindsafetyresearch/building-safe-artificial-intelligence-52f5f75058f
3. Sciuto, A., Saini, A., Forlizzi, J., and Hong, J. I. (2018). “Hey Alexa, What’s Up?”: A Mixed-Methods Studies of In-Home Conversational Agent Usage. Proceedings of the 2018 on Designing Interactive Systems Conference 2018 - DIS ’18, 857–868. https://doi.org/10.1145/3196709.3196772
4. Chokshi, N. (2018). Is Alexa Listening? Amazon Echo Sent Out Recording of Couple’s Conversation. Hey York Times. https://www.nytimes.com/2018/05/25/business/amazon-alexa-conversation-shared-echo.html.
5. Dijkstra, J.J., Liebrand, W.B.G. and Timminga, E., (1998). Persuasiveness of expert systems. Behaviour and Information Technology, 17, pp. 155–163. https://doi.org/10.1080/014492998119526

6. Nass, C., Steuer, J., Tauber, E.R. (1994). Computers are social actors. In: Proceedings of the SIGCHI conference on human factors in computing systems, pp. 72–78, New York. https://doi.org/10.1145/191666.191703
7. de Visser, E. J., Pak, R., and Shaw, T. H. (2018). From 'automation' to 'autonomy': the importance of trust repair in human–machine interaction.Ergonomics, 61(10);1409-1427. doi: https://doi.org/10.1080/00140139.2018.1457725
8. Lee, J. D., and See, K. A. (2004). Trust in automation; Designing for appropriate reliance. Human Factors, 46, 50–80.
9. Hoff, K. A., and Bashir, M. (2015). Trust in automation; Integrating empirical evidence on factors that influence trust. Human Factors: The Journal of the Human Factors and Ergonomics Society, 57(3), 407–434. Doi: https://doi.org/10.1177/0018720814547570
10. Parasuraman, R., and Manzey, D. (2010). Complacency and bias in human use of automation: an attentional integration. Hum. Factors 52, 381–410. https://doi.org/10.1177/0018720810376055
11. Parasuraman, R., and Riley, V. (1997). Humans and automation; use misuse, disuse, abuse. Hum. Factors 39, 230–253. https://doi.org/10.1518/001872097778543886
12. Parasuraman, R., Sheridan, T. B., and Wickens, C. D. (2000). A model for types and levels of human interaction with automation. IEEE Trans. Syst. Man Cybern. A Syst. Hum. 30, 286–297. https://doi.org/10.1109/3468.844354
13. Drnec, K., Marathe, A. R., Lukos, J. R., and Metcalfe, J. S. (2016). From trust in automation to decision neuroscience: Applying cognitive neuroscience methods to understand and improve interaction decisions involved in human automation interaction. Frontiers in Human Neuroscience, 10. Doi: 10.3389/fnhum.2016.00290
14. Manzey, D., Bahner, J. E., and Hueper, A.-D. (2006). Misuse of automated aids in process control: complacency, automation bias and possible training interventions. Proc. Hum. Factors Ergon. Soc. Annu. Meeting, 50, 220–224. https://doi.org/10.1177
15. Lee, J. D., and Moray, N. (1994). Trust, self-confidence, and operator's adaptation to automation. International Journal of Human-Computer Studies, 40, 153–184. https://doi.org/10.1006/ijhc.1994.1007
16. Lewis, M., Sycara, K., and Walker, P. (2018). The role of trust in human-robot interaction. In H. A. Abbass, J. Scholz, and D. J. Reid (Eds.), Foundations of Trusted Autonomy (pp. 135–159). Australia: Springer Open.
17. Muir, B. M. (1994). Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. Ergonomics, 37, 1905–1922. https://doi.org/10.1080/00140139408964957
18. Lee, J., and Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. Ergonomics, 35, 1243–1270. https://doi.org/10.1080/00140139208967392
19. Sheridan, T. B., and Verplank, W. L. (1978). Human and Computer Control of Undersea Teleoperators: Fort Belvoir, VA: Defense Technical Information Center. https://doi.org/10.21236/ADA057655
20. Kaber, D. B. (2018). Issues in Human–Automation Interaction Modeling: Presumptive Aspects of Frameworks of Types and Levels of Automation. Journal of Cognitive Engineering and Decision Making, 12(1), 7–24. https://doi.org/10.1177
21. Endsley, M. R. (2017). From here to autonomy: Lessons learned from human–automation research. Human Factors: The Journal of the Human Factors and Ergonomics Society, 59, 5–27. https://doi.org/10.1177/0018720816681350
22. Simpson, A., Brander, G. N., and Portsdown, D. R. A. (1995). Seaworthy trust: Confidence in automated data fusion. In R. M. Taylor and J. Reising (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
23. Wickens, C. D., Li, H., Santamaria, A., Sebok, A., and Sarter, N. B. (2010). Stages and levels of automation: An integrated meta-analysis. In Proceedings of the Human Factors and Ergonomics Society 54th Annual Meeting (pp. 389–393). https://doi.org/10.1177

24. Westin, C. A., C. Borst, and B. Hilburn. (2013). Mismatches between Automation and Human Strategies: An Investigation into Future Air Traffic Management Decision Aiding. In Proceedings of the 17th International Symposium on Aviation Psychology, Dayton, OH.
25. Marinik, A., Bishop, R., Fitchett, V., Morgan, J. F., Trimble, T. E., and Blanco, M. (2014). Human factors evaluation of level 2 and level 3 automated diving concepts: Concepts of operation. (Report No. DOT HS 812 044). Washington, DC: National Highway Traffic Safety Administration. http://hdl.handle.net/10919/55082