Designing Trust in Artificial Intelligence: A Comparative Study Among Specifications, Principles and Levels of Control

Fernando Galdon(✉), Ashley Hall, and Laura Ferrarello

School of Design, Royal College of Art, London, UK
fernando.galdon@network.rca.ac.uk

Abstract. This paper presents a comparative study amongst the three main frameworks acknowledged for designing trust in AI; specifications, principles and the levels of control necessary to underpin trust in order to address the rising concerns of Highly Automated Systems (HAS). We will also address trust design in four case studies specifically designed to address the rising concerns of these systems in the area of health and wellbeing. Based on the results, levels of control emerge as at the most reliable option to design trust in Highly Automated Systems, as it provides a more structured focus than specifications and principles. However, principles enhance philosophical inquiry to frame the intended outcome and specifications provide a constructive space for product development. In this context, the authors recommend the integration of all the frameworks into a multi-dimensional cross-disciplinary framework to build and extend robustness throughout the entire interactive lifecycle in the development of future applications.

Keywords: Human factors · Human-systems integration · Specifications · Levels of control · Principles · Trust · Virtual assistants

1 Introduction

Artificial intelligence (AI) systems are increasingly being used to replace human decision-making. While AI holds the promise of delivering valuable insights and knowledge across a multitude of applications, broad adoption of AI systems will rely heavily on the ability to trust their output.

With the exponential development of machine learning (ML) and deep learning (DL) techniques, a new paradigm is emerging; Machine-Human-Interaction (MHI). In this emerging paradigm, the technology holds the initiative of the interaction. These developments have urged Peter Hancock to raise a concern to the human factors community by which attention must be focused on the appropriate design of a new class of technology: Highly Autonomous Systems (HAS) [1]. This approach positions highly autonomous systems at the centre and tries to address the implications of trust from their perspective [2].

As we progress in the development of AI, the idea of 'performance' as an AI design paradigm will not be enough. Questions around, how do we achieve fairness,

T. Ahram et al. (Eds.): IHIET 2020, AISC 1152, pp. 97–102, 2020.
https://doi.org/10.1007/978-3-030-44267-5_14

robustness, explainability, accountability and value alignment through design, and how do we integrate them throughout the entire interactive lifecycle, are fundamental for the development of trusted HAS. In this context, we must learn how to build, and monitor trust.

For the last forty years, human factors approached the design of complex autonomous systems by articulating Levels of control as a design strategy to appropriately calibrate trust to achieve performance and safety goals [3]. However, Principles have recently become a design strategy being proposed from social and ethical perspectives to address trust [4]. Finally, Specifications are being proposed from a computational perspective as a design strategy to address the rising concerns of highly automated systems [2].

This paper will present a comparative study among the aforementioned frameworks to understand which of the three frameworks is better suited to design trust in the context of HAS. It will do so by addressing trust design in four case studies specifically designed to address the rising concerns of these systems in the area of health and wellbeing. In this regard, a workshop has been conducted with Design Research students at the [Removed for Review]. The workshop was structured over two days around the four case studies aforementioned.

In this context, we structured a workshop with seven students from the Masters of research programs (MRes) at the Royal College of Art. They represented a multiplicity of backgrounds ranging between fashion, textile, architecture, computer science, industrial design, and engineering.

2 Method

According to Bukhari [5] a Comparative Study analyses and compares two or more objects or ideas to examine, compare and contrast them to show how two or more subjects are similar or different. Building from this perspective we built a comparative study among the three main frameworks acknowledged to design trust in AI; specifications, principles and levels of control in order to underpin which one is better prepared to address the rising concerns of highly automated systems. In this context we aimed for a mixed methodology combining constructive approaches in the form of a design workshops, experimental design to control the variables and a semi-structured questionnaire and a post-activity debate synthesis to evaluate the outputs.

In order to address the task at hand, we defined the main area of intervention; health and wellbeing. Then, we structured four exercises around systems capable of diagnosing and providing treatment in the areas of anxiety, obesity, depression and addiction. The lead author introduced a video-demonstration of Duplex to illustrate the nature of the system and a small analysis that underlined the key characteristics of upcoming Virtual assistants. The students had 50 min to complete each task, which consisted of four parts;

1. A mapping exercise to underpin potential interventions
2. Introduction of a design framework.
3. Inference exercise to define four data points and four algorithms. This was designed to encourage students to define datasets and inference algorithms. The main purpose was to bring sensitive areas into the equation to trigger ethical design interventions.

4. An interaction task consisting of a user journey and potential design intervention. This part was structured in three areas; before the interaction, during the interaction, and after the interaction.

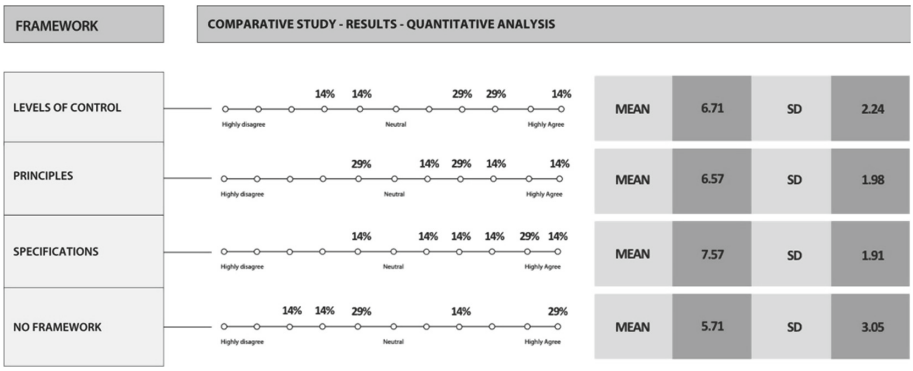
The first exercise introduced no framework. It operated as a control mechanism to understand what the students were bringing to the table and whether they would implement ethical interventions. The second exercise introduced specifications. The third exercise incorporated principles. And the last exercise introduced levels of control. In the latest exercise, a multi-dimensional framework was presented in collaboration with a trust calculator to facilitate participants' output definition by inserting a mode of calculation by which a trust rating could be obtained.

Once all the exercises were completed, the lead author introduced a semi-structured questionnaire to understand which framework was better suited to design trust in Highly Automated Systems. The questionnaire consisted of two areas; a quantitative area asked participants to rate the four frameworks proposed; no framework, principles, specifications and levels, by using an eleven points Likert scale, and a qualitative area asking participants to define the pros and cons of each framework.

3 Discussion

In the quantitative area, Specifications emerge as the most favoured framework by participants rating it with 7.57 in mean value. It is followed by levels of control with 6.71 and no framework with 6.57. The least favoured framework was Principles with 5.71 mean value (Table 1).

Table 1. Quantitative analysis



When reviewing the qualitative data obtained by asking participants to describe the pros and cons of each framework, they praise specifications for its semistructured nature which provides them with a flexible-constrained space for intervention. This differs from the prescriptive nature of levels, the openness of no framework and the abstraction of principles.

However, they also point to the limitations of specifications to address trust in ever-evolving systems, as it is a one-time a priori intervention which does not allow for a posteriori rectification. It is described as a powerful tool to understand user needs but limited to design trusted systems, especially in the context of HAS, with unsupervised and ever-evolving capabilities.

In this context, levels are described as a tool to implement quick adjustments, are beneficial, and enhance distributed self-optimisation to maintain control over the system. Furthermore, when integrating the calculator into the levels and providing a form of calculation participants described this combination as useful in reducing risks, integrating a critical dimension in product development and enhancing explainability in the design process [6].

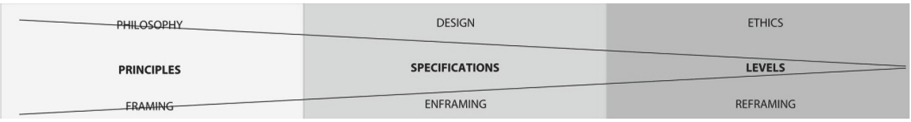
Principles though are seen as a philosophical element to open relevant debates. Finally, no framework is described as open, yet too loose in focus and abstract to address the rising concerns presented (Table 2).

Table 2. Qualitative analysis and Post-activity debate synthesis.

PARTICIPANT	COMPARATIVE STUDY - RESULTS - QUALITATIVE ANALYSIS			
	LEVELS	PRINCIPLES	SPECIFICATIONS	NO FRAMEWORK
PARTICIPANT 1	A trust calculator helped for quick design adjustments.	Provide perspectives	Helps you understand the key specifications you must focus. but you may forget to adjust it post-design	More organic and less restrictive But you can get lost.
PARTICIPANT 2	System learning from personal and collected data but user may not follow the system	System learning from personal and collected data but user may not follow the system	System learning from personal and collected data but user may not follow the system	System learning from personal and collected data but user may not follow the system
PARTICIPANT 3	Helps to categorise the ideas	Too abstract	Too abstract	More open answers, opportunity to be more fluid/free with design
PARTICIPANT 4	Strong control gives users confidence but leaves less space for the service system to process the outcome.	Relevant. they matter	Relevant. they matter	can be more based on the users
PARTICIPANT 5	interesting	Very fun	hard to understand at the begining	May be
PARTICIPANT 6	Beneficial	Philosophical	Pro: understanding the user needs, Con: lacks breakdown of effects of trust intervention.	without framework difficult to break down. Too abstract and open
PARTICIPANT 7	Distributed self-optimisation	Open debate	not like in the industry	Difficult

DEBATE	COMPARATIVE STUDY - RESULTS - FINAL ANALYSIS			
	LEVELS	PRINCIPLES	SPECIFICATIONS	NO FRAMEWORK
PARTICIPANTS	Closed system. Too prescriptive	Too abstract	Semi-structured. opens but does not close	Too open

These outputs are significant because they match a recent paper published on the 10th of November, 2019 by the Oxford Institute of the Internet in *Nature* claiming that principles are not enough to design trusted AI systems [7]. In this context, instead of providing a categorical excluding output, we propose to build an integrative multi-dimensional design framework by acknowledging the key beneficial elements of the three main frameworks by distributing these paradigms over time.



4 Conclusion

Based on these results, Levels of control emerge as at the most reliable option to design trust in Highly Automated Systems, as it provides a more structured focus than specifications and principles. However, principles enhance philosophical inquiry to frame the intended outcome, and specifications provide a constructive space for product development. In this context, the authors recommend a combinatorial strategy where principles are used as a preliminary element to frame the intended outcome. The use of specifications follows by determining the interaction and the use of levels is used as a strategy to calibrate interactions to build trust within the system to address a priori strategies around simulation, meanwhile strategies around calibration systems a posteriori strategies around reparation. The integration of all the frameworks into a multi-dimensional framework aims to build and extend robustness throughout the entire interactive lifecycle in the development of future applications.

This paper presents leading insights by providing a comparative study among proposed frameworks to design trust in AI. Although limited in scale, the results provide a highly relevant contribution to knowledge, as no other study we identified has compared these elements at once. In the process, it provides knowledge for future actions via a categorisation of existing frameworks and the development of an integrative cross-disciplinary framework to address the rising concerns of trust in AI. Future work will be dedicated to further evaluating the reliability of the frameworks presented.

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. Ortega, B.P.A.: Building safe artificial intelligence: specification, robustness, and assurance specification: define the purpose of the system. Medium (2018). https://medium.com/@deepmindsafetyresearch/building-safe-artificial-intelligence-52f5f75058f

3. Sheridan, T.B., Verplank, W.L.: Human and computer control of undersea teleoperators. Fort Belvoir, VA Def Technology Information Center (1978)
4. Floridi, L., Cowls, J.: A unified framework of five principles for AI in society. Harvard Data Sci. Rev. 1(1) (2019). https://doi.org/10.1162/99608f92.8cd550d1
5. Bukhari, S.A.H.: What is Comparative Study (2011). SSRN: https://ssrn.com/abstract=1962328, http://dx.doi.org/10.2139/ssrn.1962328
6. Galdon, F., Hall, A., Ferrarello, L.: Synthetic consequential reasoning: building synthetic morality on highly automated systems via a multidimensional-scales framework. In: Proceedings of the 2st International Conference on Human Interaction and Emerging Technologies (IHIET 2020), Lausanne, Switzerland, 22–24 April 2020 (2020)
7. Mittelstadt, B.: Principles alone cannot guarantee ethical AI. Nat. Mach. Intell. (2019). https://doi.org/10.1038/s42256-019-0114-4