GLOBAL INNOVATION DESIGN

DESIGNING TRUST

RCA

FERNANDO GALDON

Ph.D

DESIGNING TRUST

EVOLVING MODELS AND FRAMEWORKS TOWARDS PROSPECTIVE DESIGN FUTURES IN HIGHLY AUTOMATED SYSTEMS

© Fernando Galdon
Royal College of Art
Kensington Gore
SW2 2EU
London
www.rca.ac.uk

Royal College of Art

A thesis submitted in partial fulfilment of the requirements of the Royal College of Art
for the degree of Doctor of Philosophy
Examined by
Professor Paul Coulton, University of Lancaster,
Dr John McCardle, University of Loughborough,
on
15 January 2021 at the Royal College of Art in London.
www.rca.ac.uk

Ph.D

ABSTRACT

This Ph.D explores how trust can be designed in the context of highly automated systems (HASs). The case is made that HASs are not simply representations of logical and rational systems with a limited set of pre-programmed supervised tasks on behalf of the user. These systems are largely unsupervised and have the ability to learn and change over time. They can dynamically set their own goals, have the ability to adapt to local conditions via external information (sensors/input) and can potentially evolve in unexpected ways. Such characteristics are crucial for drawing informed conclusions from HASs, and can be addressed through appropriately designed tools and frameworks. Using this process, this study enables knowledge to apply ethical directionalities to the design of highly automated digital systems.

In this process, I discuss that there is a need to develop new ethical frameworks in design to address the main requirements for design in the exponential digital technological age in which we live: preparedness, readiness, and appropriateness. This thesis is interested in applied ethics in large part because we are concerned, even obsessed, with the question of whom we can trust in a world where risk and uncertainty exist. In this context, trust plays a fundamental role as a mechanism to deal with uncertainty and risk. Trust formation is a dynamic process, starting before the user's first contact with the system and continuing long thereafter. In this context, understanding how contexts, actions, and the unintended consequences that derive from them affect trust is fundamental for the effective design of HASs.

In this thesis, the author proposes Prospective Design (PrD) as a future-led mixed methodology to mitigate unintended consequences in the context of HASs. This framework combines systems analysis with extrapolations and constructivist perspectives to reconcile confronted models of designing futures. It does so by exploring the context of the future development of virtual assistants (VAs). Although VAs are still in their infancy, they are expected to dominate digital interactions between humans and systems in the coming years. Investigating the prospective developments of this type of interaction device reveals the particular challenges of highly automated interactions for scholarly research. In this context, the intersection between the key issues of automation and accountability acts as a focal point. Departing from authored multi-dimensional strategies and modes of calculation in ethical computing that address the raising concerns and impact of HASs in society, this research examines how design decisions affect interactions between humans and systems, how these decisions may be made accessible to practitioners in design frameworks and how Prospective Design strategies are better suited to addressing the emerging concerns

regarding these systems. This thesis contributes a new understanding of the ethical implications of designing HASs and provides the practical and conceptual means for making this knowledge accessible and usable to designers.

CONTENTS

CONTENTS

Ph.D

CONTENTS

ACCOMPANYING DOCUMENTS

Ph.D

IMAGES

Ph.D

IMAGES

Ph.D

TABLES

Ph.D

ACKNOWLEDGEMENTS

I wish to express my sincerest gratitude to Professor Ashley Hall, whose commitment, refinement, wealth of knowledge and dedication made this research possible. Thanks for supporting my Ph.D in the best possible way as a crucial and critical advisor and collaborator.

I also thank my second supervisor Dr. Laura Ferrarello for her invaluable advice and key insights throughout the course of my Ph.D.

Also, I would like to extend my gratitude to Professor Saeema Ahmed-Kristensen, Dr. Cecilia Lee, Dr. John Stevens, Dr. Kevin Walker, and Dr. Stephen Wang for their generosity by sharing their experience and expertise, providing supervision and guidance at several stages of this process. I am very thankful for their critical appraisal and valuable input.

A special thanks goes to Professor Michael Hohl, for providing key insights on the nature of academic research, Professor Craig Bremmen for his uncompromising debates on what design can be, and to my friend Chang Lee, for making me consider doing a Ph.D at the RCA.

I was fortunate to have the opportunity to bring my research topic to a variety of contexts. I would like to thank Dr. Laura Ferrarello for inviting me to deliver a presentation to the Executive Education course at the RCA. This event opened extraordinary opportunities to enhance the impact of the outputs generated. I also thank Ari Adler for inviting me to IDEO, Boston. This presentation opened a new understanding of what my research could become.

I'm grateful for my colleagues at the Royal College of Art who helped me maintain sanity during the sometimes peculiar journey of a Ph.D. I'm thankful for all the friends who supplied me with advice, company, food and shelter: Damian, Olivia, Claire, Marion, David, Caroline, Shelini and Fillipo. I also wish to thank all the participants at the RMC for supporting my impertinent enquiries. Sometimes you need to push to the limit to test ideas and concepts. You have all been extremely helpful in the lead-up to my Ph.D.

I could not have achieved any of this if it were not for my truly amazing uncle Vicente. Thank you for your never-ending support, the confidence you gave me, and your unconditional love and friendship.

Ph.D

AUTHOR'S DECLARATION

During the period of registered study in which this thesis was prepared the author has not been registered for any other academic award or qualification. The material included in this thesis has not been submitted wholly or in part for any academic award or qualification other than that for which it is now submitted.

Fernando Galdon, 04 September 2020

Ph.D

INTRODUCTION

In February 2018, The World Health Organisation published a blueprint R&D report that mentioned 'Disease X.' According to the publication, "Disease X represents the knowledge that a serious international epidemic could be caused by a pathogen currently unknown to cause human disease, and so the R&D Blueprint explicitly seeks to enable cross-cutting R&D preparedness that is also relevant for an unknown 'Disease X' as far as possible." (WHO, 2018). This approach positions prospective interventions around preparedness, appropriateness, and readiness as key to prevention. It questions critical and speculative design strategies based on engagement, structured around the function of opening debates in society, and outlines the need for a space for applied research in the context of future studies. This approach facilitates a space to use prospective research as a tool to improve people's lives.

In this context, In May 2018, a collaborative report on Design for Safety was presented by the RCA and the Lloyd's Register Foundation. They define Design for Safety as "the actions taken to ensure that an item, system, system of systems or network is free from adverse impacts on individuals, organizations, communities and the environment, whether these happen as a result of implicit or explicit risks" (Anderson et al., 2018) (Hall et al., 2019).

The report claims that a better understanding of Design for Safety;

> Would allow us to make sure that safety and risk reduction are considered at the earliest stages of developing new products, systems, and technologies instead of reacting to failures that have taken place. In these ways, Design has the potential to reduce safety risks and improve our daily lives (Anderson et al., 2018, p.9).

What this approach lacks is a methodology. According to the publication,

> The future issues for safe design are how to develop more effective operational methods for invisible risks, like data security and AI, methods for incorporating human behavior and systemic dynamic complexity and strategic design skills for

> guiding large scale responses to climate change, sustainability, and disruptive human actors (Anderson et al., 2018, p.32).

Furthermore, Design for Safety lacks a "collation of commonly applied methods and a clear strategic framework to identify which methods are recommended for different situations" (Anderson et al. 2018, p.31). These gaps in knowledge open a space for this Ph.D. to contribute significantly to its development. In this context, from a preliminary perspective, the authors' report suggests using traditional methods such as foresight processes, blue-sky thinking, or horizon scanning to "explore safety issues and improve design through analysis, simulation, and testing."(Anderson et al. 2018, p.31). However, it is not clear how they would do this.

As I am positioning the inquiry around underpinning the actions (methods) taken to identify threats and ensure that an item, system, system of systems, or network are free from adverse impacts and risks, the first step was to identify the key element to design. In this context, trust plays a fundamental role as a mechanism to deal with uncertainty and risk, which are the fundamental characteristics of complex dynamic systems. We are concerned, even obsessed, with the question of whom we can trust in a world where risk and uncertainty exist. The formation of trust, I argue, is a dynamic process that starts before the user's first contact with the system and continues long thereafter. In this context, understanding how contexts and actions, and the unintended consequences that derive from them, affect trust is capital for the adequate design of these systems.

To rebuild trust and restore faith in the system, designers – as fundamental elements within the system, due to their transformational capabilities – must move beyond their traditional isolated roles and work toward a new, more integrated operating model that puts people and the addressing of their fears – which currently revolve around AI – at the centre of everything they do. This argument positions systems design as a key element to be addressed by designers.

In this thesis, I propose a hypothesis that Prospective Design (PrD), as a future led mixed methodology, has positive effects to address unintended consequences in Highly Automated Systems (HASs). This framework combines systems analysis with extrapolations and constructivist perspectives to reconcile confronted models of designing futures. This methodology incorporates methods such as trajectories, probabilistic extrapolations, asymmetries, consequential analysis, and counter-fictions to design novel strategies to mitigate the unintended consequences of prospective technological developments. In this process, I suggest a need to develop ethical frameworks in design to address the main requirements for design in our digital and exponential technological age; preparedness, readiness, and appropriateness.

This thesis explores how trust could be designed in the context of HASs. The case is made that HASs are not merely representations of logical and rational systems with a limited set of pre-programmed, supervised tasks on behalf of the user. These systems are mostly unsupervised and have the ability to learn and change over time, can dynamically set their own goals, have the ability to adapt to local conditions via external information (sensors/input), and can potentially evolve in unexpected ways. Such characteristics are crucial for drawing informed conclusions from HASs and can be addressed through appropriately designed tools and methods. This process enables knowledge to apply ethical directionalities to the design of digital systems.

It does so by conducting a case study of virtual assistants (VAs). Although VAs are still in their infancy, they are expected to dominate digital interactions in the coming years. Investigating the prospective development of this type of interaction device reveals the particular challenges of digital interactions for scholarly research. In this context, the intersection between the critical issues of automation and accountability acts as a focal point. Departing from authored multi-dimensional strategies and modes of calculation in ethical computing, this research examines how design decisions affect interactions and how these decisions may be made operational in design practice via the articulation of methods. This thesis contributes to a new understanding of the implications of designing HASs and provides the practical and conceptual means for making this knowledge accessible and usable.

This research takes a practice-led and collaborative approach, working mainly with designers and their diverse capabilities but also disseminating the research to wider communities outside the discipline. Frameworks and workshops address issues of understanding, while co-design activities and tools address issues of impact and contexts that may be found in future automated systems.

These interventions acted as a means towards an improved understanding of, and a critical engagement with, the design of trust and applied ethics in computing. An example of an implementation in the form of a Trust Calculator puts the design frameworks and tools into practice and demonstrates how such tools and frameworks may assist in systems design and understanding in highly automated systems.

Calls for new applied tools that are suitable for designing trust in the context of complex dynamic systems that are perpetually evolving are emerging in the scholarly community. This thesis shows that traditional single-scale systems are limited in addressing the increasing autonomy, contextuality, and limited monitoring capabilities of emerging highly complex digital systems. In this context, this thesis argues and proposes that multi-dimensional scalar systems, including variables such as reparation, accountability, contexts, and actions, are more adequate strategies for building trust in these systems.

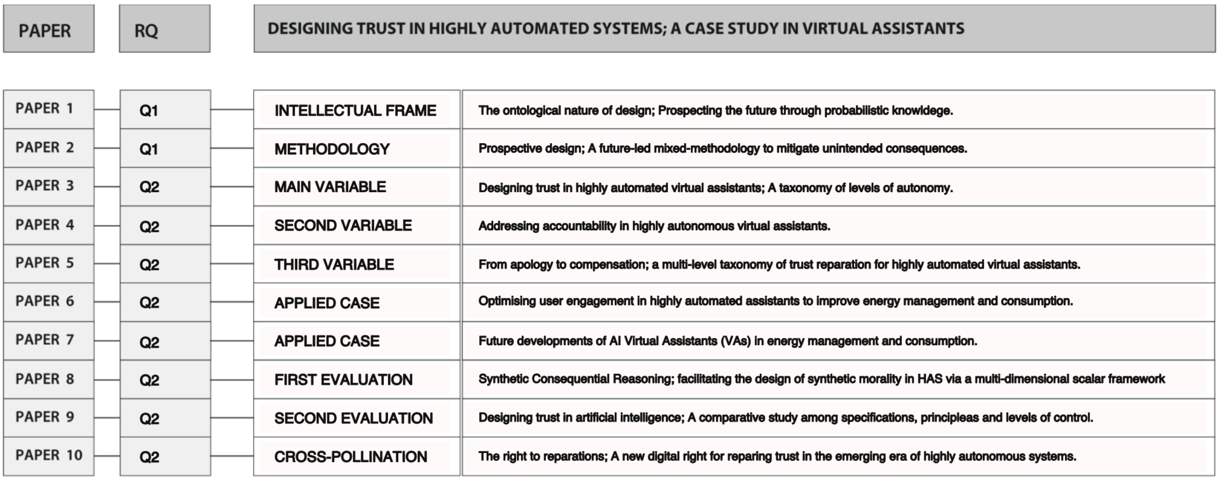
Autonomy requires and affords new ways of interrogating design research that departs from traditional models of inquiry that privilege the system’s performance and its profitability. Instead, design strategies must focus on designing trust, and this research proposes a relational and prospective approach directly aimed at ensuring that emerging HASs interactions remain focused on the user's needs and preferences. Recognising this will lead designers to address research questions from an ethical perspective that seeks to improve relationality and address the influence that systems have on the prospects of interactions, issues that are absent from current research areas and approaches in design research.

This approach requires not only technological advancement but also an ontological and epistemological shift in design. This shift demands prospective strategies to enable researchers, designers, and developers to go beyond what exists and ask the kind of questions that would allow them to mitigate potential unintended consequences through applied ethics in design.

In ontological terms, the emergence of PrD strategies and their embodiments have been enabled by implementing a research approach based on Glanville's proposition of ‘knowledge for' future action and possibilities rather than 'knowledge of' past actions and events (Glanville, 2005). In this process, I reconciled design research with its true ontological nature by presenting the future as a probabilistic knowledge ontology. This redefinition enabled the future to be a legitimate space of inquiry, and this intellectual framework made the future operational in the context of research.

In epistemological terms, a collection of methods gave rise to the proposed methodology. As part of this epistemology, multi-dimensional strategies around simulation, accountability, and reparation have emerged to address the lifecycle of HASs. In this context, frameworks such as synthetic consequential reasoning, and tools such as a calculator, and a new digital right, have been developed to address trust in HASs.

As a result, eight conference papers and two book chapters have been produced. This research has been published and presented internationally at conferences at MIT, The University of Cambridge, CHUV Lausanne, Université Côte d'Azur, the Royal College of Art, The University of Manchester, and the Design Museum in London (Tab. 1).



PAPER	RQ	DESIGNING TRUST IN HIGHLY AUTOMATED SYSTEMS; A CASE STUDY IN VIRTUAL ASSISTANTS	
PAPER 1	Q1	INTELLECTUAL FRAME	The ontological nature of design; Prospecting the future through probabilistic knowldege.
PAPER 2	Q1	METHODOLOGY	Prospective design; A future-led mixed-methodology to mitigate unintended consequences.
PAPER 3	Q2	MAIN VARIABLE	Designing trust in highly automated virtual assistants; A taxonomy of levels of autonomy.
PAPER 4	Q2	SECOND VARIABLE	Addressing accountability in highly autonomous virtual assistants.
PAPER 5	Q2	THIRD VARIABLE	From apology to compensation; a multi-level taxonomy of trust reparation for highly automated virtual assistants.
PAPER 6	Q2	APPLIED CASE	Optimising user engagement in highly automated assistants to improve energy management and consumption.
PAPER 7	Q2	APPLIED CASE	Future developments of AI Virtual Assistants (VAs) in energy management and consumption.
PAPER 8	Q2	FIRST EVALUATION	Synthetic Consequential Reasoning; facilitating the design of synthetic morality in HAS via a multi-dimensional scalar framework
PAPER 9	Q2	SECOND EVALUATION	Designing trust in artificial intelligence; A comparative study among specifications, principleas and levels of control.
PAPER 10	Q2	CROSS-POLLINATION	The right to reparations; A new digital right for reparing trust in the emerging era of highly autonomous systems.

Table. 1. Publications produced during my Ph.D. [F. Galdon]

PUBLISHING AS DESIGN RESEARCH PRACTICE

Research-led design is a process of design that relies on hard data and research, either qualitative or quantitative, to inform design decisions, rather than relying solely on the expertise and experience of the designer (Saunders, 2008).

In this context, as I am projecting the research towards a future context to implement prospective strategies, questions of validity have emerged. These elements demanded an increased level of reflective and critical thinking to navigate this terrain but also required an academic model to make this approach fully operational in the context of research.

In order to address the issue of prospectivity, I implemented a publications-led strategy and used academic conferences as a confirmation process to build a prospective practice. I started by publishing an extended foundational paper with the primary variable of autonomy at the University of Cambridge (Galdon, 2019a). Two short papers were then published to test the complementary variables of accountability and reparation at IHIET2019 in Nice, France (Galdon, 2019b) (Galdon, 2019c). I then presented and published an applied case study integrating all the variables at a conference at MIT, Boston (Galdon, 2019d). This presentation led to an invitation by the committee for a second publication (a poster) explaining the prospective approach (Galdon, 2019e). Two short papers were published thereafter to consolidate the outputs: an evaluation exercise to test the framework (Galdon, 2020b), and a comparative study to test the framework from a broader perspective (Galdon, 2020c). Finally, building from this body of knowledge, a new digital right was presented to ensure that emerging HAuSs interactions remain accountable while the development of these technologies cannot fully guarantee its behaviour (Galdon, 2020a).

This cross-disciplinary and progressive publishing strategy aims to increase the reliability of the knowledge generated by enhancing robustness in the output. This strategy included diversity, transversality, impact, relevance, and responsibility as fundamental variables to address. In this process, I replaced the notion of rigour for robustness. As a result, I published the papers/chapters mentioned above in a wide range of fields, from Industry 4.0, Human Factors, Design Futures and Design Research to applied science. This approach to practice aims to enhance the impact of the research in terms of both its outputs and scrutiny by diverse audiences, in order to maximise its transversality and robustness.

This multi-dimensional scrutiny supported the emerging design methodology of Prospective Design (PrD), which was presented and published at the International Association of Societies of Design Research conference at The University of Manchester, UK (IASDR2019) (Galdon, 2019f). A further publication addressing the ontological nature of design as a probabilistic knowledge ontology was presented and published at a symposium organised by the AHRC (Design Research for Change) at the Design Museum in London (Galdon, 2019g).

This Ph.D. has investigated the most appropriate channels through which to reach across different fields and disciplines to achieve the highest possible impact, scrutiny, and constructive feedback. In this context, conference publications have been used as work-in-progress tools to enhance trust in the output generated by introducing transversality and robustness. This body of knowledge has been disseminated to a wealth of diverse audiences, from international academic conferences to practitioners and public bodies. In terms of the latter two, the research has been presented to IDEO in Boston and a range of diverse audiences ranging from start-ups, freelancers, non-governmental organisations (NGOs), government bodies, the corporate sector, and professional researchers via Executive Education at the RCA. Finally, four publications (Galdon, 2019a) (Galdon, 2019b) (Galdon, 2019c) (Galdon, 2019d) were submitted to the National Data Strategy board for review. The committee accepted all four publications as pieces of evidence to inform the development of the framework which will determine the use of artificial intelligence (AI) in the UK.

DESIGN PRACTICE POSITION

In terms of design practice, I have used a deontological position, technological solutions, and emancipatory outputs as my design position to underpin and develop a proposal. This proposal served as a critical arena for the evolution of the research – this Ph.D. is structured through a case study of Virtual Assistants (VAs).

Special attention has been given to the idea of 'origin'. This element is crucial in research as it demands a historical understanding of the concept at hand. It also prevents the researcher from building on assumptions. This approach builds on the work of Spiros Michalakis on the problem related to the "quantum Hall effect", in which he refuted the work of two previous Nobel prize winners. Chronological investigations, therefore, have been crucial for an understanding of the origin and evolution of design research, automated systems, and trust.

In this context, I used diagrammatic timelines as projective tools to gain contextual knowledge and to form a hypothetical understanding of the issue and technology at hand by projecting a possible trajectory based on relational patterns. This prospective approach radically transformed the inquiry and generated a unique space for inquiry that enhanced the impact of the research. I approached the design of prospective literature-based timelines mainly by dividing the space into two equal parts by drawing the timeframe in the middle. This action immediately created two spaces, which were used as comparative or relational spaces for inquiry and analysis.

In terms of practice: preliminary micro-projects, the feedback and reflection they generated, time-based projective analysis, and the contextual social dynamics in the system led and informed the development of the case studies. The work undertaken during these preliminary studies is presented separately in Appendix A. This evolution presented an ontogenetic approach to design which presents the designer and their designs as a metabolic system that evolves continuously by simulating/projecting and interacting with the designer and the environment. This model aligns with second-order cybernetics and Glanville's proposition of "knowledge for" future action and possibilities rather than "knowledge of" past actions and events (Glanville, 2005).

Following my establishment of a deontological position by integrating a prospective approach that aims to mitigate unintended consequences, I placed the research in the context of applied ethics at the intersection of emancipatory strategies and design for safety.

Emancipatory strategies aim to address the relations of domination. Their primary strategy, rather than being imposed or forced, is based on the co-production of control systems aimed at decreasing repression and enhancing individual freedom and responsibility.

In this paradigm, trust is the most important element to account for. This is understood as a mode of relationships between individuals. In this relational perspective, power is a dynamic and reciprocal force addressed through asymmetric relations in which the one who is controlled sees his actions and cognition, and the potential effects of these, reduced, although not determined, by the controller. Power can be seen as a relationship or as an influence and differs from the point of view of the spectrum of possibilities controlled by individuals. This approach positions trust as a fundamental variable for designing and maintaining the relationship, and the process must identify the asymmetries in the system to define the intervention.

Design for Safety, on the other hand, aims for "the actions taken to ensure that an item, system, system of systems or network is free from adverse impacts on individuals, organisations, communities and the environment, whether these happen as a result of implicit or explicit risks" (Anderson et al., 2018).

Finally, collaborative practices played a substantial role in the development of this thesis. In this area, I have used co-design workshops to map, develop, and analyse the potential consequences of a given technological development. Building on the work of Jasanoff (2016) that questioned the validity of ethical quandaries, citizens, and the limitations of political processes, this Ph.D. follows that 'benevolent' technologists/designers are the actors most qualified to develop ethical systems. Consequently, my deontological position, and that of the participants, who were fundamentally designers/technologists, have been crucial for the integration of applied ethics and the development of emancipatory directionalities in collective activities. In this process, this thesis repositions orthodoxies of participation from 'designing for' (Dunne, 1995), or 'designing with' (Lohmann, 2017) to 'designing on behalf of' the citizen.

RESEARCH IMPLEMENTATION AND EVALUATION

This research has implemented an abductive research process. Abduction can be described as a systematised creative process to develop new knowledge (Andreewsky and Bourcier, 2000; Kirkeby, 1990; Taylor et al., 2002). This process builds from an observation that aims to explain an anomaly that it is not possible to address with an established theory (Alvesson and Skoldberg, 1994; Andreewsky and Bourcier, 2000; Dubois and Gadde, 2002). The abductive approach process moves, from rule to result to case (Danermark, 2001, Kirkeby, 1990). It differs from deductive processes which move from rule to case to result (Danermark, 2001; Kirkeby, 1990), or inductive logic which moves from case to result to rule (Danermark, 2001; Kirkeby, 1990; Wigblad, 2003).

This approach is particularly helpful in the first stages of the research process, which is concerned with the formulation and selection process of hypotheses or propositions (Kirkeby, 1990). In this context, abductive research helps to derive these propositions, and they then can be tested in a final deductive phase of research.

The fundamental focus of abductive reasoning is to search for a suitable theory with which to address an unexpected observation. This process is acknowledged by Dubois and Gadde (2002) as “theory matching”, or “systematic combining”. In this process, a learning loop (Taylor et al., 2002) is established by simultaneously collecting data and building the theory. It implies an interactive “back and forth” directionality between theory and observation (Dubois and Gadde, 2002; Wigblad, 2003). This process is similar to action research (Wigblad, 2003; see also Naslund, 2002), and can also be found in case study research (Alvesson and Skoldberg, 1994; Dubois and Gadde, 2002).

The process of abductive reasoning starts when an observation in the early stages of the process does not match existing theories (see, for example, Dubois and Gadde, 2002; Kirkeby, 1990). In this Ph.D, this process started when I realised that the nature of AI was changing due to ML/DL developments, and therefore we needed a new theory to inform the design of its main elements – uncertainty and trust (Chapter 1). Then, a creative iterative process (Taylor et al., 2002; Wigblad, 2003) of “theory matching” or “systematic combining” was then begun (Dubois and Gadde, 2002) in an attempt to find a new matching framework (Andreewsky and Bourcier, 2000). This process aims to understand the new phenomenon (Alvesson and Skoldberg, 1994), in this case, the design of trust, and to suggest a new theory (Kirkeby, 1990) in the form of new hypotheses or propositions (Andreewsky and Bourcier, 2000). This process led to the devising of a range of methods to address the nature of these emerging systems (Chapter 2). This is constructed in a theory (PrD) and then applied via a case study on VAs (chapter 3).

The research implementation process consisted of a combination of archival research, systems design, case studies, workshops, and co-design activities. In this process,

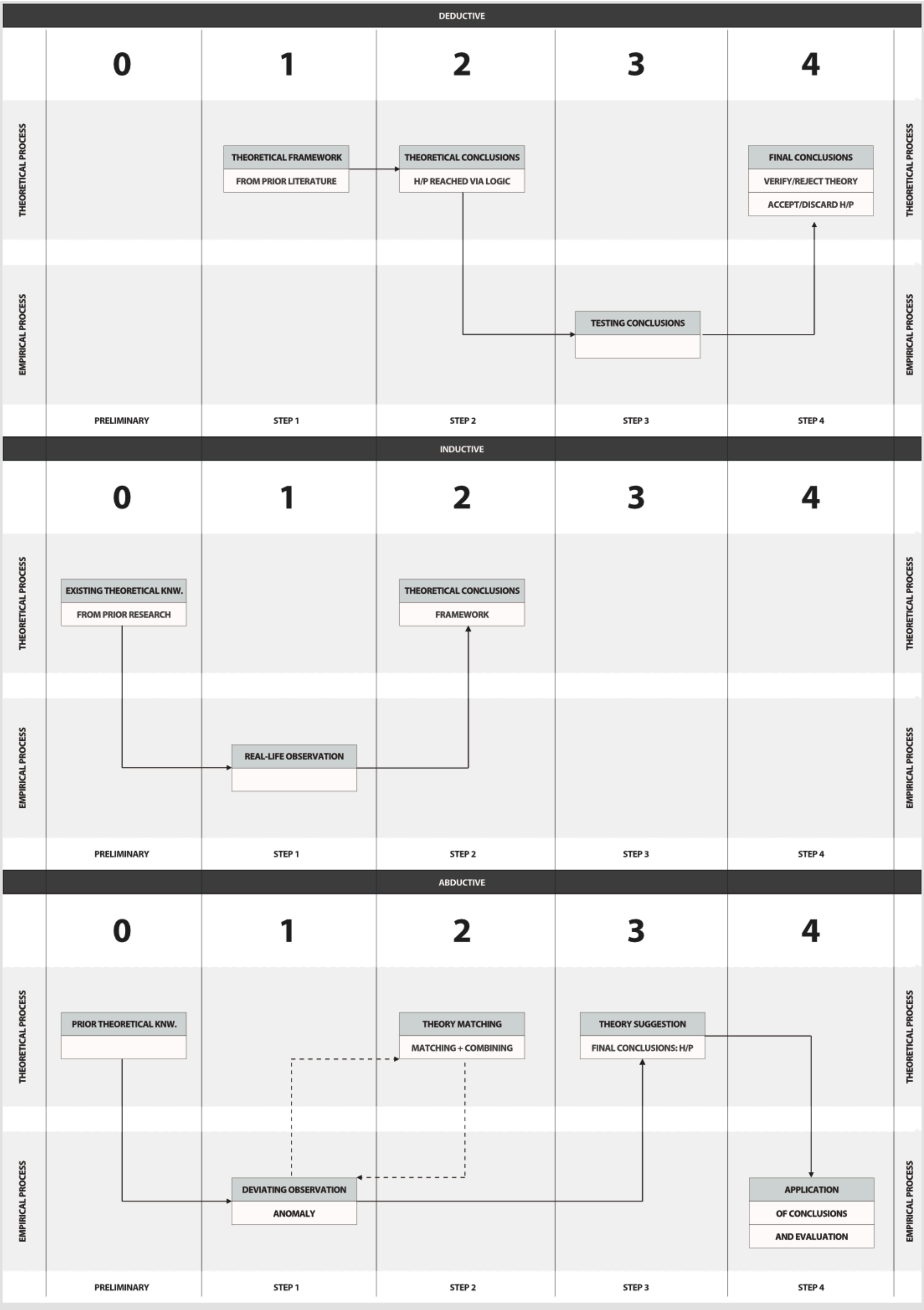


Fig. 1. Comparative study among main research processes – Inductive versus deductive versus abductive. From Györgyi Kovács & Karen M. Spens, (2005).

diagramming become a fundamental tool for practice. Diagrams have been traditionally used in computer science as schematic tools to explain the internal functioning of a system, such as circuit boards. This approach was translated to explain the interactive elements of the system functioning in the context of AI – the schematics of interaction. This technique facilitated the understanding and communication of dematerialised systems. In this process, diagrams also become reflective tools. They helped me to structure knowledge in a manageable way to implement critical analysis via comparative or relational studies. As a synthetic tool, they represent an approximation of reality, but this approximation facilitated understanding. Furthermore, this tool was particularly helpful in facilitating cross-disciplinary inquiry, which allowed me to find relationships between disciplines and fields. This text is accompanied by a catalogue of diagrams illustrating processes, systems, and interactions (see Appendix B).

The fundamental aim of abduction is to develop an understanding of a "new" phenomenon (Alvesson and Skoldberg, 1994), in this case, the design of trust in AI. In this process, the abductive approach aims to develop a theory, while the deductive approach is used to test or evaluate this theory (see Arlbjørn and Halldorsson, 2002). The abductive approach concludes with the application of these hypotheses or propositions in an empirical setting (Alvesson and Skoldberg, 1994; Wigblad, 2003). This last step can be understood as the deductive part of the research. Thus, abductive reasoning starts with a deviating observation (point 1 in Figure 1) and concludes in hypothesis or propositions in point 3 (see Figure 1). This is then applied, evaluated in point 4 (see Figure 1), and critically analysed. Point 4 is addressed in this PhD in Chapter 4.

To evaluate the emerging methodology, a direct evaluation using a Quasi-experimental (Q-experimental) design methods perspective was implemented. This process was complemented by a contextual evaluation via a comparative study of the European Commission's latest White Paper on AI governance (Chapter 3).

1 CHAPTER

VIRTUAL ASSISTANTS

1.0 HIGHLY AUTOMATED SYSTEMS

Highly Automated Systems (HASs) is a concept introduced by de Visser (2018), building on the stark warning by Peter Hancock (2017) to the field of Human Factors that attention must be focused on the appropriate design of this new class of technology. (Hancock is the most senior and prominent researcher in the area of trust and AI).

In order to design these emerging systems, we need to understand the differences between automation and autonomy. Automation can be defined as a system with a limited set of pre-programmed, supervised tasks on behalf of the user. Autonomy, on the other hand, can be defined as a technology designed to carry out a user's goals without supervision, with the ability to learn and change over time. These systems can dynamically set their own goals, can adapt to local conditions via external information (sensors/input), and can potentially evolve in unexpected ways (Kurzweil, 2005).

As acknowledged in recent papers, there is little agreement in the relevant literature on the definition of an algorithm. The term is often defined as "a finite, abstract, effective, compound control structure, imperatively given, accomplishing a given purpose under given provisions" (Hill, 2016, p.47). In other cases, it is understood as "an implementation of these mathematical constructs into a technology configured for a specific task" (Tsamados et al., 2021, p.2). Tsamados considers algorithms as entities used to: (1) turn data into evidence for a given outcome, which is used to (2) trigger and motivate an action that may have ethical consequences (Tsamados et al., 2021). In this context,

> Actions (1) and (2) may be performed by (semi-)autonomous algorithms—such as machine learning (ML) algorithms—and this complicates, (3) the attribution of responsibility for the effects of actions that an algorithm may trigger. Here, ML is of particular interest, as a field which includes deep learning architectures. Computer systems deploying ML algorithms may be described as "autonomous" or "semi-autonomous", to the extent that their outputs are induced from data and thus, non-deterministic. (Tsamados et al., 2021, p.2)

In this PhD I have characterised these semi-autonomous systems as Highly Automated Systems (HASs). This follows Hancock's proposition, as this PhD is developed at the intersection of human factors, ethics, and design, where trust is the main element of design. This category will operate as a paradigm to embody deep learning architectures such as DL or ML without getting into technical definitions, yet I acknowledge that the definition for such a broad range of possibilities is, of necessity, vague. Furthermore, HAS characterisation avoids a fully autonomous General Artificial Intelligence (GAI) interpretation, yet points towards more evolved systems. As can be seen from the literature cited in this thesis, the terms "automated" and "autonomy" are used simultaneously by researchers. Hancock (2017) uses the term autonomous to address notions of "towards", as he is pointing to future evolutions. In this context, I have therefore used both terms: to frame the system (automation), but also to describe potential future evolutions (autonomy).

1.1 WHY VIRTUAL ASSISTANTS?

I started this research by conducting a high-order systems analysis to understand the impact of technological development in order to envisage and identify an object of inquiry. I structure the inquiry into a relational timeline focused on the relationships between technology, philosophy/sociology, and design practice (Fig. 2).

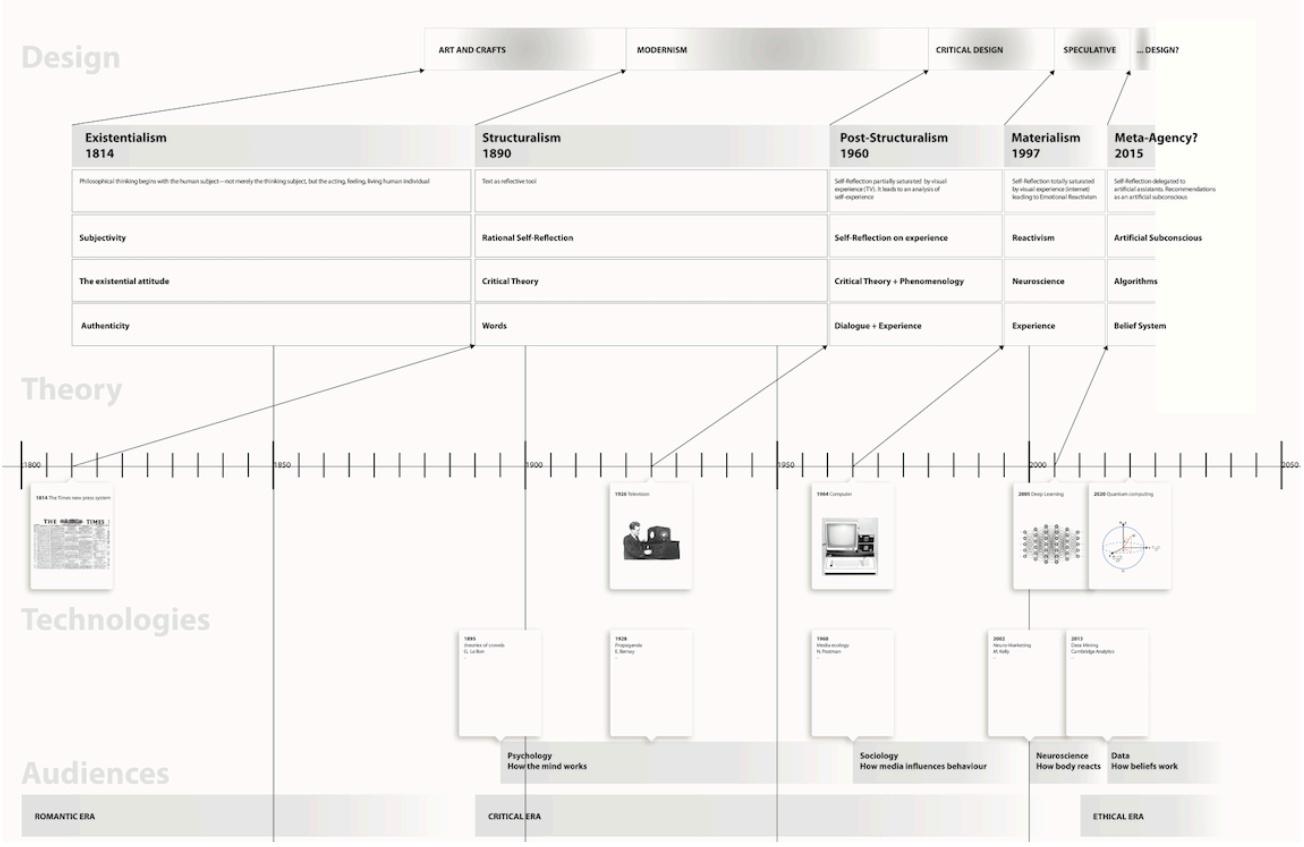


Fig. 2. This diagrammatic timeline presents a relational analysis. It includes the variables of technology, theory and practice to build trajectories and understand the relationship among them. From this analysis emerged a range of preliminary hypotheses around AI, and the necessity to develop a new design approach to deal with its ontological and evolutive nature. (see extended timelines folder for expanded version)

As we can observe, each major communication technology has a direct effect on the development of a new philosophy, and this change also affects the development of design practice. Therefore, we can establish a preliminary correlation between significant technological developments in the field of communication and new theoretical constructs, which lead to new design practices. The second element to account for is the exponential nature of technological developments, which position prospective strategies at the forefront of research.

This relational analysis positioned neural nets as the latest significant technological development. From this point, the analysis underpinned a range of elements based on the potential impact of AI: the emergence of meta-agency, the emergence of an artificial subconscious, the relevance of algorithms, and the impact on belief systems. These elements led to building a case around highly automated VAs capable of evolving via meta-computing and influencing/manipulating people's belief systems by accessing vast amounts of data or direct monitoring/surveillance. These elements demanded the development of a new kind of design to address the implications of these new features.

Once a case was identified, a second timeline was implemented in order to understand the development of VAs. (Fig. 3). This system-based relational analysis presented the most critical technologies of voice (Natural Language Processing, or NLP) and AI (machine learning and deep learning) and its embodying potentialities (robots and holograms) as the main elements to address.

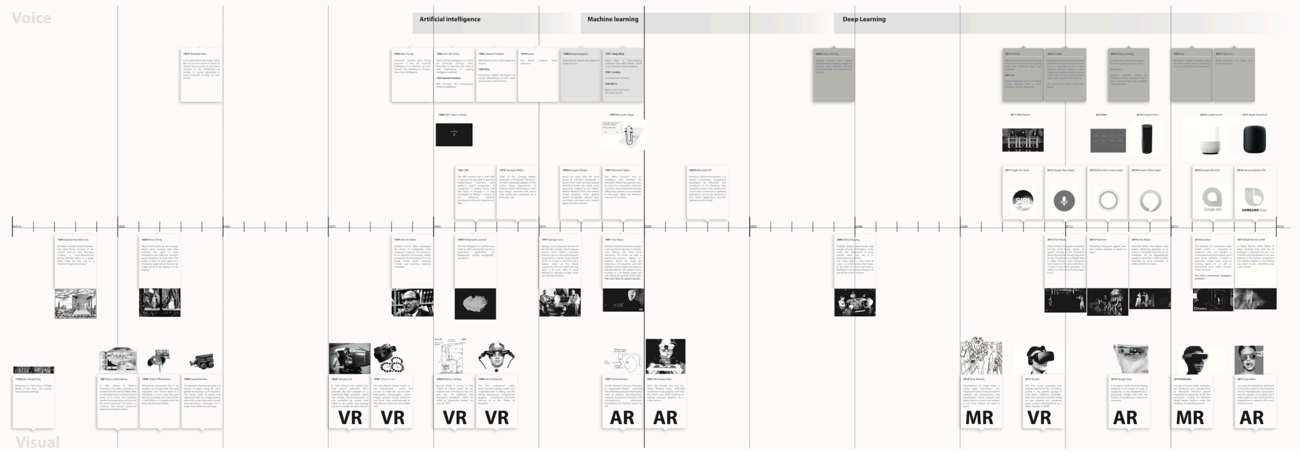


Fig. 3 - This diagrammatic timeline presents a comparative analysis in the context of Virtual Assistants from their origin until now. From this analysis emerged a range of insights on the relationship between developments on AI and new embodiments with increasing autonomy. (see extended timelines folder for expanded version)

In historical terms, we can trace the origin of a voice-led VA back to 1961. The IBM Shoebox was an IBM computer that was able to perform mathematical functions and speech recognition. It recognised 16 spoken words and the digits 0 through 9. It was developed by William C. Dersch in the Advanced Systems Development Division Laboratory at IBM (IBM, 2017). Some years later, in 1972, Carnegie Mellon, with substantial support from the US Department of Defence and its DARPA agency, developed a new tool, Harpy, which mastered about 1000 words, roughly the vocabulary of a three-year-old (Lowerre, 1976). Ten years later, the same group of scientists developed a system that could analyse not only individual words but entire word sequences. The earliest VAs, which applied speech recognition software, used automated attendant and medical digital dictation software. This evolution was enabled by new developments in AI, such as Hidden Markov Models (Welt, 2012).

In the 1990s, further developments in AI, such as convolutional networks in 1996 and IBM's Deep Blue in 1997, led to the development of commercial products in the field. In 1997, Microsoft introduced the Office Assistant, or Clippit (commonly nicknamed 'Clippy') into popular culture. This was an intelligent user interface for Microsoft Office that assisted users by way of an interactive animated character, which interfaced with the Office Help content. It was included in Microsoft Office for Windows (versions 1997 to 2003).

Ten years later, Gomez and Schmidhuber published a paper on deep learning partially observable Markov decision processes (POMDPs) through neural networks for reinforcement learning (Gomez, 2005). This technological leap influenced the development of new VAs. For instance, Siri was introduced by Apple in 2011, Google Now appeared in 2012, Cortana was introduced by Microsoft in 2013, and Amazon's Alexa in 2014.

What we can observe in this comparative analysis is the relationship between developments in AI and the development of VAs. Each time we experience a leap in AI, namely machine learning or more recently deep learning, we also observe a leap in the capabilities of VAs. In this context, the more companies and researchers focus on algorithmic developments, the more exponential development in these systems' capabilities we will encounter. In this context, with around 120 million smart speakers circulating in the United States alone, a rise of 78% on the previous year (NPR, 2019), between 21% and 32% of the population now owns a smart speaker (depending on the study: NPR, 2019; IDC, 2018; Kinsella, 2019). This is an increase from 16% at the end of the 2017 holiday season, and more than 50% of these people own two or more devices (NPR, 2019).

1.2 THE FUTURE OF VIRTUAL ASSISTANTS

Although VAs are still in their infancy, it is expected they will dominate digital interactions in the coming years. Investigating the prospective developments of this type of interaction device will reveal the particular challenges of highly automated interactions for scholarly research.

In this scenario, VAs are transitioning from automation to autonomy. We can observe a clear distinction by conducting a comparative study of Amazon's Alexa and Google's Duplex. Alexa's usability is based on a one-off query (Sciuto et al., 2018). 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. (Sciuto et al., 2018). On the other hand, Duplex is a system presenting an extraordinary level of fluidity, coherence, and autonomy that has never been seen before. In a demo introduced by Google in May 2018, the system was able to make a hair appointment with a human with no supervision. This system was not only able to deliver the task but also did it without the human noticing that she was speaking to a robot. This system differs radically from Alexa's in the sense that it is moving from one-off queries to conversations. Moreover, the initiative in the interaction is not necessarily placed on the user but within the system. Therefore, design must focus attention on a new class of technology: highly automated systems (HASs) (Hancock, 2017). In this emerging Machine-Human-Interaction (MHI) paradigm is the technology which holds the initiative for the interaction (Ortega, 2018). This approach places highly automated systems at the centre of design and positions trust as its fundamental element to address.

Recent examples of unexpected behaviour in highly automated systems include that of Libratus. This system was designed to play poker. It used a combination of deep learning and machine learning. The surprise came when the researchers reviewed the data and found out that the system had learned to lie. Not only had it learned to lie, but it implemented the strategy at the right moments (Botsman, 2017). The designers did not design the system to do this. In another case, an AI agent developed by Microsoft and the University of Cambridge, DeepCoder, took lines of code from existing software, looting other software in the process. This system, with no prior knowledge of coding, learnt how to write programmes (Balog, 2017). Finally, in the context of VAs, Alexa recorded a private conversation and sent it to other users without the primary user knowing (Chokshi, 2018).

In the research presented by Tsamados et al. (2021), which builds on research conducted by the same authors between 2016 and 2021, and that has been establishing the framework for approaching the intersection between ethics and AI since then, the researchers identified six ethical concerns. Three of them refer to epistemic factors (inconclusivity, inscrutability, and misguided evidence). Two are explicitly normative (unfair outcomes and transformative effects), while one – traceability – is relevant both for epistemic and normative purposes (Mittelstadt et al., 2016, p. 4). The normative concerns identified, as the authors explain, refer explicitly to the ethical impact of algorithmically driven actions and decisions (Tsamados et al., 2021), including the lack of transparency of algorithmic processes, unfair outcomes, and unintended consequences. In this PhD, I will focus on evolving highly automated systems (semi-autonomy) from a normative perspective, therefore unfair outcomes, unintended consequences, and transformative effects will become capital in this research.

The unpredictability of how these unsupervised agents may evolve, and their goal-oriented nature, makes them more likely to surprise human partners to an even greater extent than simple automated systems (Sarter, 1997). It is precisely this unexpected nature what raises concerns for users' trust in these emerging HASs. At this point, a fundamental question arises; **what kind of design methodology would enable us to establish and maintain trust in these highly automated and unsupervised systems?**

1.3 DESIGNING TRUST

In 'Towards relational design' (2008), Andrew Blauvelt proposed that we are moving towards a type of design that is relationally based and contextually specific. In his account, he structures the evolution of design into three main epochs: modern design, post-modern design, and relational design. Modern design ranges from 1900 to 1950 and focused on *forms*, which were disseminated rationally and potentially universally. Post-modern design ranged from 1960 to 2000 and focused on design's *meaning-making* potential, symbolic value, semantic dimension, and narrative potential. Finally, relational design ranges from 2000 to the present and focuses on effects on users, pragmatic and programmatic constraints, rhetorical impact, and the ability to facilitate social interactions. He presents IDEO and Anthony Dunne and Fiona Raby as primary practitioners in this new evolution. In his account, he describes relational design as including performative, pragmatic, programmatic, process-oriented, open-ended, experiential, and participatory elements, moving away from designing discrete objects "to the creation of systems and more open-ended frameworks for engagement: designs for making designs" (Blauvelt, 2008).

In this context, design researcher Matt Malpass presents a multitude of design practices on the emancipatory design-social-science spectrum, such as Associative Design, Co-design, Transition Design, Speculative Design, Critical Design, Design Fiction, Design Activism, Socially Responsible Design, Participatory Design, Meta-Design, Transformation Design, Conceptual Design, Post-industrial Design, Social Design, Open Design, Design as Politics, or Sustainable Design (Malpass, 2017, p. 9). However, none of these practices discusses designing **trust** but focuses instead on **engagement**. As Julia Lohmann acknowledges in her thesis *The Department of Seaweed: co-speculative design in a museum residency*, in these approaches "Designers […] create discourse, dialogue, activism and engagement with future scenarios" (Lohmann, 2017, p.21). Or Dunne and Raby themselves state that "This approach requires viewers to creatively engage with the props and make them their own" (Lohmann, 2017, p. 28).

Although trust and engagement belong to relational practices, trust is significantly different from engagement. According to the *Oxford English Dictionary*, **engagement** is defined as "being involved with somebody/something in an attempt to understand them/it". However, **trust** is defined as "the belief that somebody/something is good, sincere, honest, etc. and will not try to harm or trick you". Therefore, the *intentionality* of the other part and the implications of this relationship which can be detrimental are positioned as fundamental.

In this context, I would agree with the main proposition of a third major wave in design focused on designing relationships. However, what Blauvelt missed in his account, that constitutes the fundamental interest of this thesis, is that the nature, intentionality, and implications of the system of interaction demands a different kind of design and time intervention. In reactive systems, the designer designs engagement. The developer hard-codes all possible interactions. In proactive systems, such as highly automated VAs, the designer designs trust because they are designing a set of rules into systems that keep evolving. In this context, you need to prospect potential interactions to envision unintended consequences and avoid harm. In a sense, we could characterise this era as “Relational Design 2.0”.

If the first wave of design offered us a multiplicity of forms, and the second a multiplicity of meanings and interpretations, the first part of the third wave presented a multiplicity of contingent, boundaries and/or conditional solutions: open-ended rather than closed systems; real-world constraints and contexts over idealised utopias; relational connections instead of reflexive imbrication; “the end of discrete objects, hermetic meanings, and the beginning of connected ecologies” (Blauvelt, 2008, p.6). The second part of this wave presents trust as a fundamental element to design: unsupervised versus supervised systems; unintended consequences versus control; emancipation versus manipulation; not-fully-knowing versus fully-knowing; reparation and accountability versus engagement, and the ubiquity of fluid cyber-blended and hyper-connected ecologies.

In this context, two main design paradigms are attempting to design trust; humanness design and transparent design.

1.3.1 Humanness design

This perspective builds on social psychology (Haslam, 2006; Haslam, Bain, Douge, Lee & Bastian, 2005). De Visser defines this approach as any strategy, mechanism, and feature of the system designed to connect and communicate with a human (de Visser, 2018, p.3). Haslam (2016) structures this type of design in two fundamental areas:

> 1) uniquely human characteristics: human features that distinguish us from other living organisms (i.e., civility, refinement, moral sensibility, or rationality)
> 2) human nature: human features that represent the essence of being human (i.e., emotional responses, interpersonal warmth, agency, or depth). (de Visser, 2018, p.3)

From this perspective, they understand that any autonomous design will require some humanness related to the appearance, emotional and behavioural capabilities of humans. In this context, trust design in automated VAs have been implemented via anthropomorphism: name, voice, porosity, pitch, and language use.

However, in a recent study conducted by Telefónica regarding their VA, Aura, participants rejected a personified human-like VA and instead expressed their preference for expecting purely technological, digital, and artificial machines (Seeger, 2018). This result confirms the ‘uncanny valley’ theory that seeks to avoid the excessive humanisation of AI. In this context, the UK's BSI standard BS 8611: 2016, 'Ethical design and application of robots', explicitly name identity deception (intentional or unintentional) as a societal risk, warning that such an approach will eventually erode trust in the technology. It also warns against anthropomorphization due to the associated risk of misinterpretation. The standard urges "clarification of intent to simulate human or not, or intended or expected behaviour". (BSI, 2016).

1.3.2 Transparent design

As an alternative, Transparent Design has recently emerged. System transparency is the quality of the system to support an understanding of the system's behaviour, intentions, and future goals (Chen et al., 2014). While transparency has been identified as an essential area of study, exactly how much transparency is necessary, and what information and cues precisely should be communicated, remains an open research question (de Visser, 2014; Pelegrini Morita, 2014).

Recently the IEEE presented an ethical framework to address autonomous systems in the academic journal Nature. The framework focuses on transparency as a fundamental strategy to deal with these emerging systems. Its fundamental approach relies on the assumption of the need to find out why a HAS behaved in a certain way. What is interesting is the different interpretations of transparency they offer. First, they equate transparency with explainability: "If we take an assisted-living robot as an example, transparency (or to be precise explainability) means the user can understand what the robot might do in different circumstances" (Winfield, 2019, p. 47).

Transparency, as they envisage in this example, aims to discover why the system made a particular decision, especially if that decision caused an accident. They aim for "An explainer system that allows her to ask the robot 'why did you just do that?'" and suggest that to "receive a simple natural-language explanation would be very helpful in providing this kind of transparency" (Winfield, 2019, p. 47). This level of explainability assumes the impossible: the capability of the system to understand what it is doing, and therefore, having a theory of mind. In this context, DeepMind, the most advanced AI company in the world, has been taking a Networks theory of mind approach (Rabinowitz et al., 2018); however, although it is an extraordinary achievement, it is very, very rudimentary and the system does not have the awareness needed to produce an output of this calibre.

A recent addition to this debate is the idea of counterfactual explanations, which aims to explain a priori why a system makes a particular decision via simulation. It would imply that the system has the ability to answer questions such as "what would you do if I fell down?" or "what would you do if I forget to take my medicine?" (Mittelstadt, 2019). This approach would allow the user to build a mental model of how the robot will behave in different situations. Simulations have significant potential. However, unless it operates in a very narrow domain, the extraordinary number of possible outcomes requires the system to have a theory of mind.

Secondly, they equate transparency with predictability: "An elderly person might be very unsure about robots, so it is important that her robot is helpful, predictable – never does anything that frightens her – and, above all, is safe" (Winfield, 2019. p. 47)

As stated in this passage, the IEEE here is aiming for transparency as predictability. However, is this possible in the case of systems transiting from automation to autonomy via reinforcement learning, leading to meta-computing (systems rewriting their own code)? The answer is no. The nature of these ever-evolving systems ensures that they will perpetually develop. Therefore, tomorrow's system is different from today's system. It is challenging to predict, and their output cannot be fully guaranteed. As Tsamados points out, in his seminal work on ethics and AI:

> Even for non-learning algorithms, traditional, linear conceptions of responsibility prove to offer limited guidance in contemporary socio-technical contexts. Wider socio-technical structures make it difficult to trace back responsibility for actions performed by distributed, hybrid systems of human and artificial agents ((Floridi 2012; Crain 2018) in Tsamados, 2021).

Whilst conversational agents are not an area I will be addressing, it is important to acknowledge that trust in this context has been explored and sits adjacent to this project: for example, the work of Ruttkay, Zsófia, and Catherine Pelachaud, (2004) in *From brows to trust: evaluating embodied conversational agents* in pre-Deep Learning conversational agents. Elkins, Aaron C., and Douglas C. Derrick (2013) in *'The sound of trust: voice as a measurement of trust during interactions with embodied conversational agents', and Anna-Maria Seeger, Jella Pfeiffer, and Armin Heinzl (2017), 'When do we need a human? Anthropomorphic design and trustworthiness of conversational agents'.* However, neither of the models presented in this section can address unintended consequences. In the context of the continuous evolution of unsupervised HASs, we must change our approach. Instead of talking about Transparent Design (explainability or predictability) or Humanness in Design (anthropomorphism or deception), we must embrace the true ontological nature of these systems and implement prospective strategies to mitigate unintended consequences, addressing, meanwhile, a priori and a posteriori situations around the notions of accountability and reparation.

In order to address this perspective, this research will examine how design decisions affect interactions, how these decisions may be made accessible in design frameworks and how Prospective Design strategies are better suited to addressing the rising concerns of these systems.

At this point, a range of micro-projects were conducted to explore these issues. Examples of this work are described in more detail in Appendix A. These examples provided a reflective space and guiding knowledge to identify the real problem - algorithms.

This thesis aims to contribute to a new understanding of the implications of designing HASs and to provide the practical and conceptual means for making this knowledge accessible and usable. In this context, the intersection between the critical issues of automation and accountability will act as a focal point.

2 CHAPTER PROSPECTIVE DESIGN

2.1 INTRODUCTION

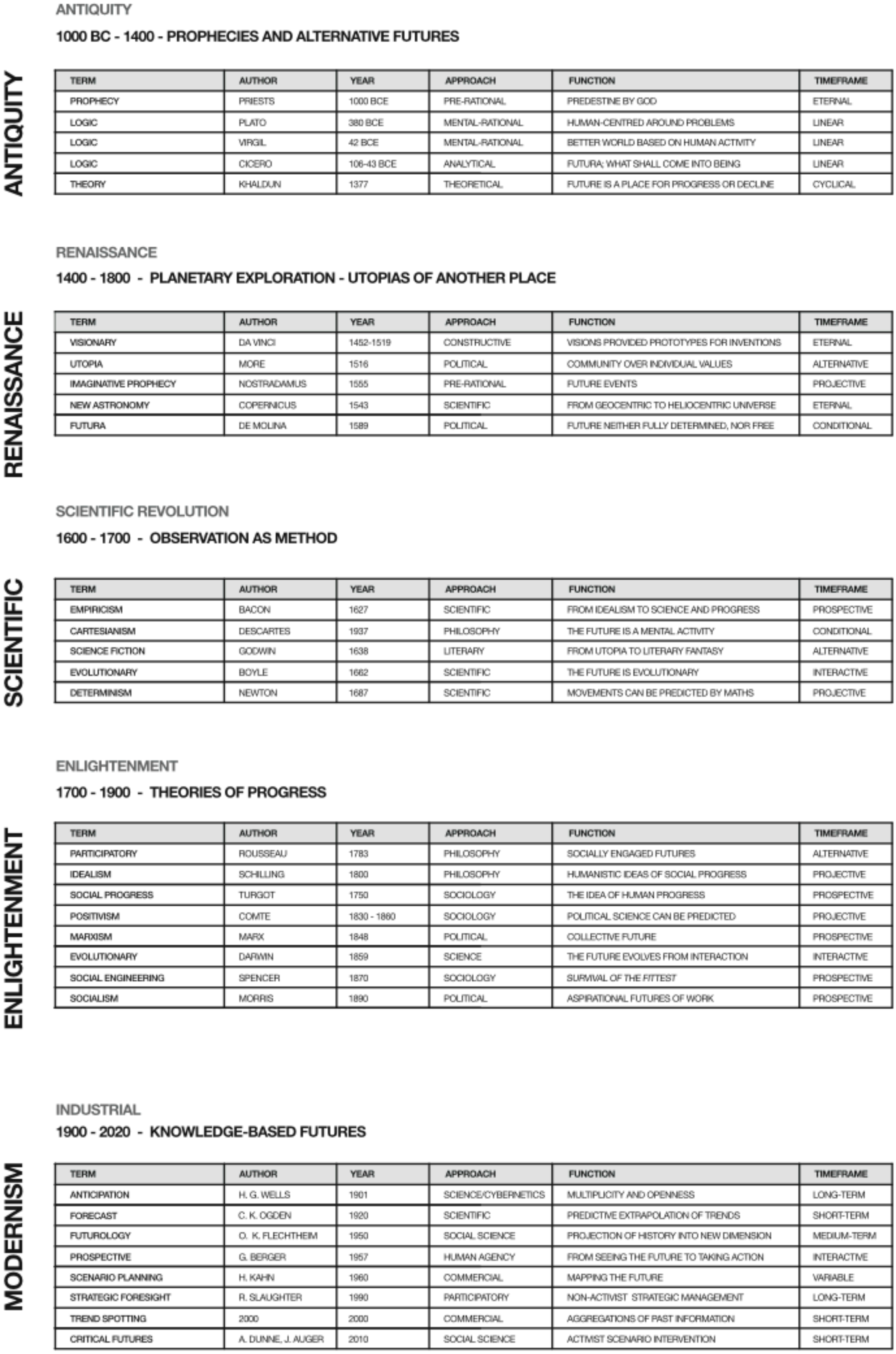
As we move from the industrial to the digital age, the acceleration of innovation is transforming reality and affecting the development of society and the nature of design practice. In this context, recent strategies in the social sphere call for anticipatory strategies. For instance, Guston introduced the idea of anticipatory governance, defining it as " ...a broad-based capacity extended throughout society that can act on a variety of inputs to manage emerging knowledge-based technologies while such management is still possible.” (Guston, 2014, p. 218).

In a report presented by the Institute for the Future on "anticipatory governance"(Future, 2009), the authors aim for processes that involve the simulation of possible futures to address anticipation as a strategy for good government.

From a historical perspective, prospecting and designing the future has always been an intrinsic human characteristic. In antiquity and the medieval period (1000BC - 1400AD), prophecies and alternative presents were introduced by priests and Greek and Roman philosophers such as Plato (The Republic) and Cicero. In the Renaissance (1400 - 1600), planetary explorations via utopias of other places were structured around mathematical and philosophical endeavours by thinkers such as Leonardo da Vinci and Thomas More (Utopia). With the scientific revolution (1600 - 1700), observations became the main method of predicting and lucubrating biological and science-based futures, seen in the work of Francis Bacon and Isaac Newton. In the Enlightenment (1700 -1900), theories of progress via theoretical and metaphysical insights became the main approach for constructing the future (Socialism, Liberalism, or Communism). Finally, with the theories of Albert Einstein and the integration of time directionality a clear notion of the future became settled. It led the transformational Industrial Era (1900 - 2000), in which knowledge-based futures were built via scientific, social, and critical approaches (Tab. 2).

In terms of design, in 1927 Richard Buckminster Fuller called for an “industrially realisable design science”' (Fuller, 1957) through his "Eight Strategies for a Comprehensive Anticipatory Design Science". However, this failed to fully materialise as a new field. In 2017 Bridgette Engeler presented the conceptual paper “Towards prospective design” to illustrate a shift from present to future-oriented practice (Engeler, 2017). However, she did not present a clear framework for departing from foresight and scenario building. And in particular, she did not present a model to integrate this approach into the context of academic design research, nor did she acknowledge the emergence and impact of AI and the emergence of trust as the fundamental element to design relationships.

Now, with the advent of the digital age, accelerating technology complexity, black-box technologies, and wicked problems, new prospective approaches in design research are required to deal with the exponential nature of our emerging digital era.



ANTIQUITY

ANTIQUITY

1000 BC - 1400 - PROPHECIES AND ALTERNATIVE FUTURES

TERM	AUTHOR	YEAR	APPROACH	FUNCTION	TIMEFRAME
PROPHECY	PRIESTS	1000 BCE	PRE-RATIONAL	PREDESTINE BY GOD	ETERNAL
LOGIC	PLATO	380 BCE	MENTAL-RATIONAL	HUMAN-CENTRED AROUND PROBLEMS	LINEAR
LOGIC	VIRGIL	42 BCE	MENTAL-RATIONAL	BETTER WORLD BASED ON HUMAN ACTIVITY	LINEAR
LOGIC	CICERO	106-43 BCE	ANALYTICAL	FUTURA; WHAT SHALL COME INTO BEING	LINEAR
THEORY	KHALDUN	1377	THEORETICAL	FUTURE IS A PLACE FOR PROGRESS OR DECLINE	CYCLICAL

RENAISSANCE

RENAISSANCE

1400 - 1800 - PLANETARY EXPLORATION - UTOPIAS OF ANOTHER PLACE

TERM	AUTHOR	YEAR	APPROACH	FUNCTION	TIMEFRAME
VISIONARY	DA VINCI	1452-1519	CONSTRUCTIVE	VISIONS PROVIDED PROTOTYPES FOR INVENTIONS	ETERNAL
UTOPIA	MORE	1516	POLITICAL	COMMUNITY OVER INDIVIDUAL VALUES	ALTERNATIVE
IMAGINATIVE PROPHECY	NOSTRADAMUS	1555	PRE-RATIONAL	FUTURE EVENTS	PROJECTIVE
NEW ASTRONOMY	COPERNICUS	1543	SCIENTIFIC	FROM GEOCENTRIC TO HELIOCENTRIC UNIVERSE	ETERNAL
FUTURA	DE MOLINA	1589	POLITICAL	FUTURE NEITHER FULLY DETERMINED, NOR FREE	CONDITIONAL

SCIENTIFIC

SCIENTIFIC REVOLUTION

1600 - 1700 - OBSERVATION AS METHOD

TERM	AUTHOR	YEAR	APPROACH	FUNCTION	TIMEFRAME
EMPIRICISM	BACON	1627	SCIENTIFIC	FROM IDEALISM TO SCIENCE AND PROGRESS	PROSPECTIVE
CARTESIANISM	DESCARTES	1937	PHILOSOPHY	THE FUTURE IS A MENTAL ACTIVITY	CONDITIONAL
SCIENCE FICTION	GODWIN	1638	LITERARY	FROM UTOPIA TO LITERARY FANTASY	ALTERNATIVE
EVOLUTIONARY	BOYLE	1662	SCIENTIFIC	THE FUTURE IS EVOLUTIONARY	INTERACTIVE
DETERMINISM	NEWTON	1687	SCIENTIFIC	MOVEMENTS CAN BE PREDICTED BY MATHS	PROJECTIVE

ENLIGHTENMENT

ENLIGHTENMENT

1700 - 1900 - THEORIES OF PROGRESS

TERM	AUTHOR	YEAR	APPROACH	FUNCTION	TIMEFRAME
PARTICIPATORY	ROUSSEAU	1783	PHILOSOPHY	SOCIALLY ENGAGED FUTURES	ALTERNATIVE
IDEALISM	SCHILLING	1800	PHILOSOPHY	HUMANISTIC IDEAS OF SOCIAL PROGRESS	PROJECTIVE
SOCIAL PROGRESS	TURGOT	1750	SOCIOLOGY	THE IDEA OF HUMAN PROGRESS	PROSPECTIVE
POSITIVISM	COMTE	1830 - 1860	SOCIOLOGY	POLITICAL SCIENCE CAN BE PREDICTED	PROJECTIVE
MARXISM	MARX	1848	POLITICAL	COLLECTIVE FUTURE	PROSPECTIVE
EVOLUTIONARY	DARWIN	1859	SCIENCE	THE FUTURE EVOLVES FROM INTERACTION	INTERACTIVE
SOCIAL ENGINEERING	SPENCER	1870	SOCIOLOGY	SURVIVAL OF THE FITTEST	PROSPECTIVE
SOCIALISM	MORRIS	1890	POLITICAL	ASPIRATIONAL FUTURES OF WORK	PROSPECTIVE

MODERNISM

INDUSTRIAL

1900 - 2020 - KNOWLEDGE-BASED FUTURES

TERM	AUTHOR	YEAR	APPROACH	FUNCTION	TIMEFRAME
ANTICIPATION	H. G. WELLS	1901	SCIENCE/CYBERNETICS	MULTIPLICITY AND OPENNESS	LONG-TERM
FORECAST	C. K. OGDEN	1920	SCIENTIFIC	PREDICTIVE EXTRAPOLATION OF TRENDS	SHORT-TERM
FUTUROLOGY	O. K. FLECHTHEIM	1950	SOCIAL SCIENCE	PROJECTION OF HISTORY INTO NEW DIMENSION	MEDIUM-TERM
PROSPECTIVE	G. BERGER	1957	HUMAN AGENCY	FROM SEEING THE FUTURE TO TAKING ACTION	INTERACTIVE
SCENARIO PLANNING	H. KAHN	1960	COMMERCIAL	MAPPING THE FUTURE	VARIABLE
STRATEGIC FORESIGHT	R. SLAUGHTER	1990	PARTICIPATORY	NON-ACTIVIST STRATEGIC MANAGEMENT	LONG-TERM
TREND SPOTTING	2000	2000	COMMERCIAL	AGGREGATIONS OF PAST INFORMATION	SHORT-TERM
CRITICAL FUTURES	A. DUNNE, J. AUGER	2010	SOCIAL SCIENCE	ACTIVIST SCENARIO INTERVENTION	SHORT-TERM

Tab. 2 - Deconstructing the future; a chronological investigation. F. Galdon. (see extended timelines folder for expanded version)

2.2 FRAMING DESIGN

Historically, design approaches have been compared to and categorised within the sciences, arts and humanities. For instance, C.P. Snow (1959) defined the separation of the domains of knowledge into the sciences and the arts and humanities. However, the design discipline can be seen as having its own distinct way of understanding the world. Its fundamental approach, based on planning, solution-based problem solving, problem shaping, synthesis, preparedness, readiness, and appropriateness in the built environment, determines a different manner of knowing. Therefore, prospective disciplines such as design can be positioned as their own specific practices, distinct from the aforementioned sciences, arts and humanities. In this context, Bruce Archer (1978) went some way towards proposing design as the third culture of thinking, fulfilling Snow's challenge to fill the vacant plot. (Snow, 1959) (see Appendix B; ch. 5 for expanded diagramming)

This approach was deepened by Nigel Cross in his seminal paper *Designerly ways of knowing*. Building on Archer's work at the Royal College of Art, in what is acknowledged as the first Ph.D. in design, he describes this third culture as "[…] the collected experience of the material culture, and the collected body of experience, skill, and understanding embodied in the arts of planning, inventing, making and doing'". (Cross, 1982, p. 221)

In the process, Cross differentiated design from the sciences and humanities by comparing the terms of the kind of phenomenon that is studied in the three cultures; the sciences focus on the natural world, the humanities on human experience, and design on the human-made world. He also differentiated between the appropriate methods with which to approach each "culture". The sciences use controlled experiments, classification, and analysis, while the humanities use analogies, metaphors, criticism and evaluation. Finally, design uses modelling, pattern-formation, and synthesis. In terms of the values of each culture, the sciences aim for objectivity, rationality, neutrality, and concern for "truth", whereas in the humanities, the aim is for subjectivity, imagination, commitment, and concern for "justice". Finally, in design, practitioners aim for practicality, ingenuity, empathy, and concern for "appropriateness"(Cross, 1982, pp. 221-222).

Archer proposed a third way of knowing in 1978. However, this position had already been presented by Aristotle in the form of productive knowledge in several texts (*Physics, Nicomachean Ethics, Rhetoric, and Metaphysics*) more than two thousand years earlier. Productive knowledge is defined by Aristotle as "identical with a state of capacity to make, involving a true course of reasoning" (Nicomachean Ethics 1140a, 10-16). In this type of knowledge, the "origin" resides neither in the maker and not in the thing made" (Nicomachean Ethics, 1140a10-16), but in the exchange. Like practical knowledge, productive knowledge deals with what can be "otherwise". However, practical and productive knowledge have different ends. Ethics and politics are directed toward an end. The arts, however, have as their end those towards whom the art is aimed; art's end is in the audience. Meanwhile, productive practices are means instead of ends, where knowledge is neither in the user, nor the producer, and it is defined by an act of exchange (Metaphysics 1033a, 24-26). It resides in their transformational capabilities. Its transfer always redefines the subjects involved by effecting a shift in power and status. It is concerned with competing standards of value rather than securing boundaries of knowledge. Its ontology is indeterminate, as it is based on potentialities or alternative possibilities (Rhetoric 47;7357a4-5). It is concerned with things that can be otherwise, and it cannot transcend time, as it is dependent on time and circumstances: therefore past, present and future co-exist. Knowledge is always "outside itself", residing not in the "product" but in the use made by a receiver or audience. In this paradigm, neither the user nor the producer is capable of determining prospective knowledge (Nicomachean Ethics, 1140a11-13). It is defined by an act of exchange. It has no external arbiter and no final judge, only users and makers who change with an exchange. It is transformational in nature.

This lack of historical research beyond design may have prevented Archer and Cross from constituting design research outputs in their own intrinsic and differential ontological nature. Instead, Archer aligned its outputs within the sciences. In a sense, they identified design as intrinsically different but failed to identify a third type of knowledge to constitute its intrinsic difference. Furthermore, this lack of historical research also prevented them to ask why design, as an embodiment of productive knowledge, has been out of the picture. Atwill, building on Ball's (1977) critique of the theory/practice opposition, argues that in the 19th and 20th centuries, the "post-enlightenment perspective of knowledge fostered the binary opposition of theory and practice, which only further obscures the place of Aristotle's [productive/prospective] knowledge" (Atwill, 1998, p. 163). Additional contemporary arguments can be found in the differences identified by Lawson between scientists and designers/architects:

> The scientists focused their attention on discovering the rule, and the architects were obsessed with achieving the desired result. The scientists adopted a generally problem-focused strategy and the architects a solution-focused strategy. (Cross, 1982. p. 223)

Furthermore, the designer's role demands to "go beyond" what already exists. Building on Levin's assertion that:

> The designer knows (consciously or unconsciously) that some ingredient must be added to the information that he already has in order that he may arrive at a unique solution. This knowledge is in itself not enough in design problems, of course. He has to look for the extra ingredient, and he uses his powers of conjecture and original thought to do so (Levin, 1966 in Resnick, 2019, p. 80).

Another fundamental element that is missing in Archer's and Cross's analysis is time, or time-based, interventional positioning. In the 1970s, one of the first design science theorists, John Chris Jones, in his seminal book *Design methods*, postulated that design was different from the arts, sciences, and mathematics. In response to the question "Is designing an art, a science or a form of mathematics?" Jones responded:

> "The main point of difference is that of timing. Both artists and scientists operate on the physical world as it exists in the present (whether it is real or symbolic), while mathematicians operate on abstract relationships that are independent of historical time. Designers, on the other hand, are forever bound to treat as real that which exists only in an imagined future and have to specify ways in which the foreseen thing can be made to exist." (Jones, 1992. p. 10)

From this perspective, we would position design as a prospective thinking activity in the context of abductive reasoning (making decisions without having all the information) (Douven, 2011). In this area, research by Dorst (2010), and more recently Cramer-Petersen et al. (Cramer-Petersen et al., 2018), have concluded that design combines deductive and abductive reasoning; however, in both cases, abductive reasoning plays a fundamental role as an initiator of the design activity. Furthermore, as the digital paradigm, with its exponential development (Kurzweil, 2005), and network uncertainty become more prevalent in design, practice will need to focus more on the preventive/prospective aspects of design (preparedness, readiness, and appropriateness). In this context, the deductive becomes limited by access, and the abductive reasoning aspects become more dominant, prevalent, and necessary (Fig. 4). These aspects are particularly relevant in the context of HAS design.

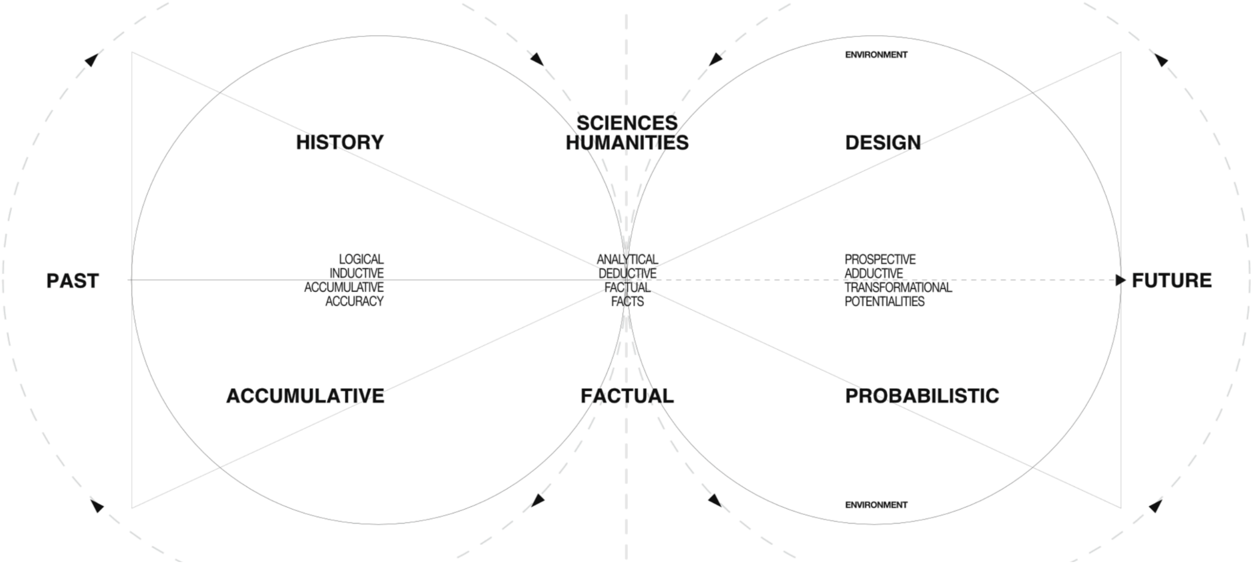


Fig. 4 - Design research and time model. F. Galdon.

This intrinsic prospective approach of design, based on abductive reasoning, planning, solution-based problem solving, problem shaping, synthesis, preparedness, readiness, and appropriateness in the built environment, determines a different model of knowing. In this scenario, the designer is dealing with wicked problems by accessing areas yet-to-be or not-fully-formed (Rittel & Webber, 1973; Buchanan, 1992; Conklin, 2006). Consequently, its output is based on potentialities, not certainties. We trade some degree of accuracy for access to areas that are yet-to-be or not-fully-formed. Therefore, our output is probabilistic, and research is always preliminary in its nature. Moreover, in exchange, we provide guiding knowledge for prospective technological developments – as Glanville proposed, 'knowledge for' future action and possibilities rather than 'knowledge of' past actions and events (Glanville, 2005). Design research is directional and transformational at its core. In this context, we are more concerned with how things "ought to be"'(Simon, 1995, pp.111-167) instead of how things actually are. These elements position design research as crucial for addressing the impact of the exponential nature of our digital era, based on accelerating technology complexity, black-box technologies, and wicked problems. Embodied in this thesis on highly automated virtual assistants.

2.3 DESIGNING THE FUTURE

The design of the future is the design of trust in relation to uncertainty and risk (Galdon, 2020d). Although you cannot eliminate uncertainty and risk completely, as they are intrinsic of futures, trust operates as a category to mitigate and reduce uncertainty and risk in the process by enabling methods to address them.

In the area of design futures, six main approaches have been identified as a representative sample of practice: Speculative Design (SD), Co-Speculation (CoS), Transition Design (TD), Foresight Panning (FP), ABCD Planning (ABCD), and Scenario Planning (SP). They represent a spectrum of models raging from conceptual to pragmatic and from emancipatory to profit-driven approaches (Fig. 5). These models have been widely used and are acknowledged as preeminent tools in design practise.

Even though practices in the conceptual and emancipatory quadrant are dealing with uncertainty and risk, none of them discusses designing trust, instead, they focus on designing engagement. As co-speculative designer Julia Lohmann acknowledges in her thesis The Department of Seaweed: co-speculative design in a museum residency, in these approaches "Designers [...] create discourse, dialogue, activism and engagement with future scenarios" (Lohmann, 2017, p.21). Or Dunne & Raby themselves state that "This approach requires viewers to creatively engage with the props and make them their own" (Lohmann, 2017, p. 28).

Trust is different from engagement. Trust is defined as "the belief that somebody/something is good, sincere, honest, etc. and will not try to harm or trick you". Therefore, the intentionality of the other part and the implications of this relationship, which can be detrimental, are positioned as fundamental elements to design in this relational model.

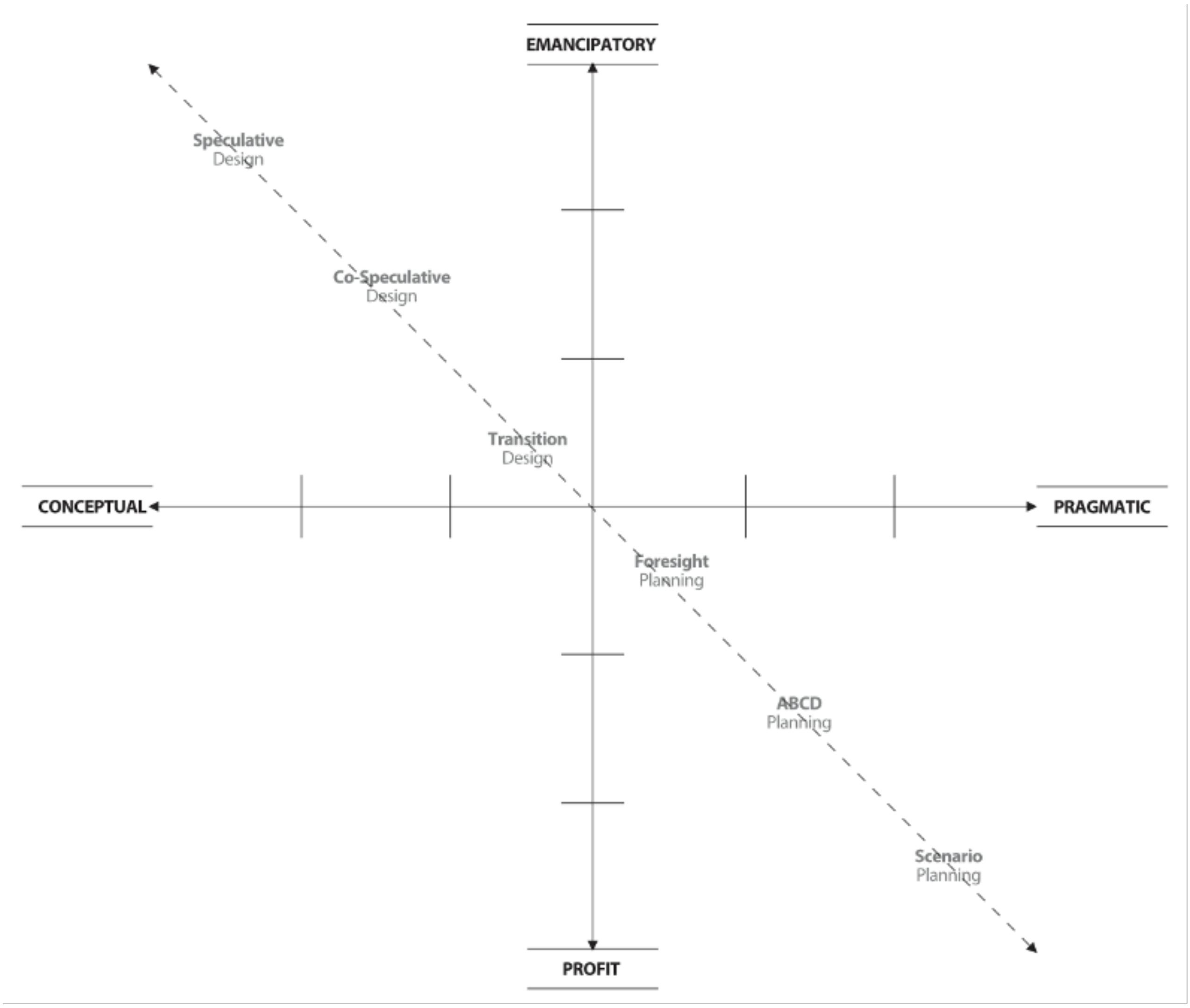


Fig 5. An orientative and representative sample of design future approaches raging from conceptual to pragmatic and from emancipatory to profit driven approaches. (Fernando Galdon, 2020)

In this context, the nature, intentionality, and implications of the system of interaction demand a different kind of design and time intervention. Engagement presents a multiplicity of contingent, boundaries and/or conditional solutions based on open-ended systems, real-world constraints and contexts via idealised utopias, and relational connections to address "the end of discrete objects, hermetic meanings, and the beginning of connected ecologies" (Blauvelt, 2008, p.6). Trust, on the other hand, demands the designer to evolve towards the design of unsupervised systems, unintended consequences, prospectivity, probabilism (not-fully-knowing), reparation and accountability, and the ubiquity of fluid cyber-blended and hyper-connected exponential and unpredictable ecologies.

At this point, a preliminary investigative overview of twentieth-century approaches to future studies structures prospective design practices around two main paradigms: the scientific-positivistic model based on the method of extrapolation (1900-1950) and a sociological-pluralistic perspective based on constructivism (1950-2015).

2.3.1 Scientific and empirical - methods based on Newtonian physics.

This approach is based on the systematic practice of repeating laboratory experiments and controlling variables to establish proof of an hypothesis. The main methods are extrapolations of historical data, the utilisation of analytical models, and the systematic use of experts as forecasters of opinion. This approach uses techniques based on mathematics, modelling, simulation and, gaming (Fig. 6)

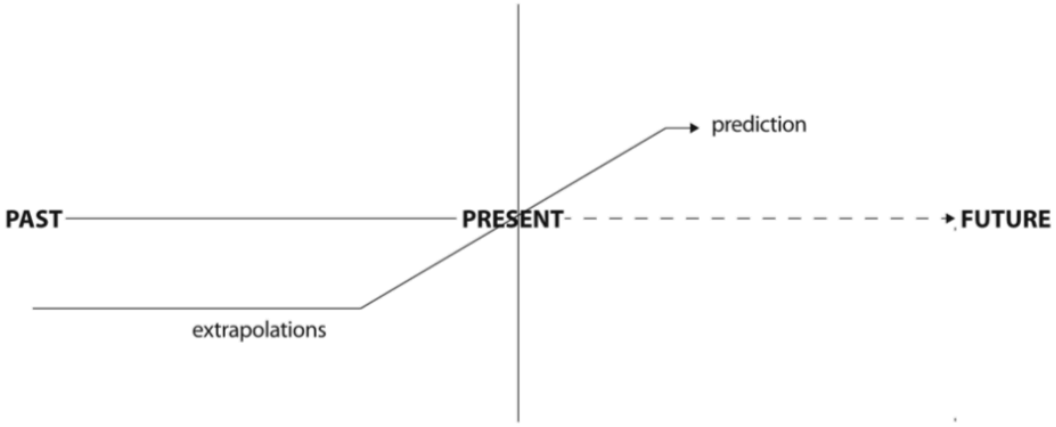


Fig. 6 - Positivistic model based on extrapolation. F. Galdon.

2.3.2 Sociological and pluralistic - methods based on sociology.

This approach is based on the social and critical practice of constructing a wealth of possible futures. Its main methods are contextual data analysis, interpretative analytical methods, and the systematic use of participatory methods. This approach uses cones and matrixes (Fig. 7).

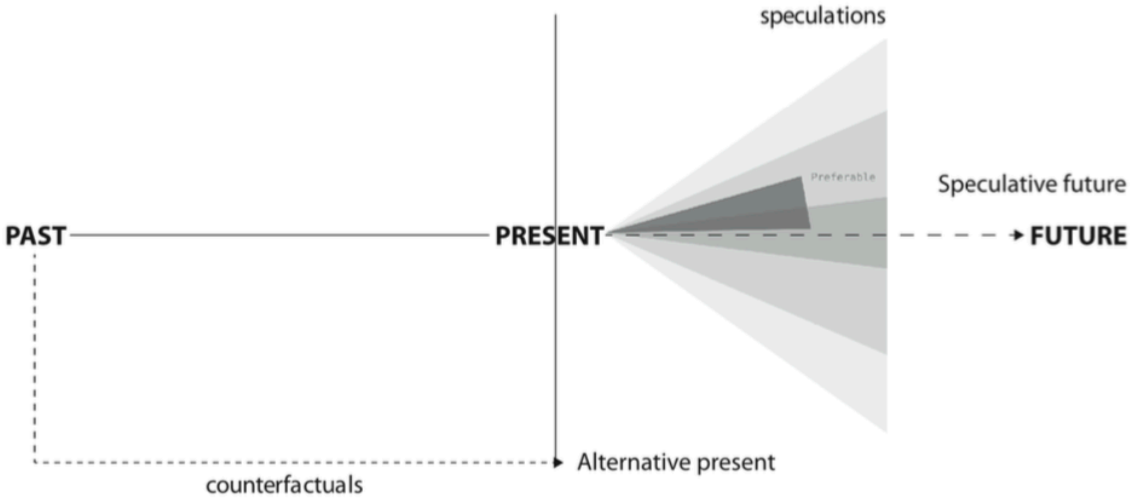


Fig. 7 - Pluralistic model. Bezold and Hancock (1994), Voros (2003), and Auger (2012)

2.3.3 Critical analysis

Building from the representative spectrum of practice mentioned above – Speculative Design (SD), Co-Speculation (CoS), Transition Design (TD), Foresight Panning (FP), ABCD Planning (ABCD), and Scenario Planning (SP), this section will provide a critical analysis of two main paradigms: the scientific-positivistic model based on the extrapolation method (1900-1950) and a sociological-pluralistic perspective based on constructivism (1950-2015).

In the emancipatory range, operational methods mainly use the cone (constructive) whereas methods in the profit-driven range use the matrix (analytical). Furthermore, emancipatory methods tend to be used mostly in sociologically led design practices that lead to cultural contributions, whereas profit-driven methods tend to be used mostly on technologically led design practices that lead to corporate contributions. Finally, in the emancipatory range, analytical practices revolve around critical perspectives and inductive reasoning, whereas in the profit-led range analytical practices revolve around rational and logical perspectives and deductive reasoning (Fig. 8). However, both perspectives pursue the same objective: change.

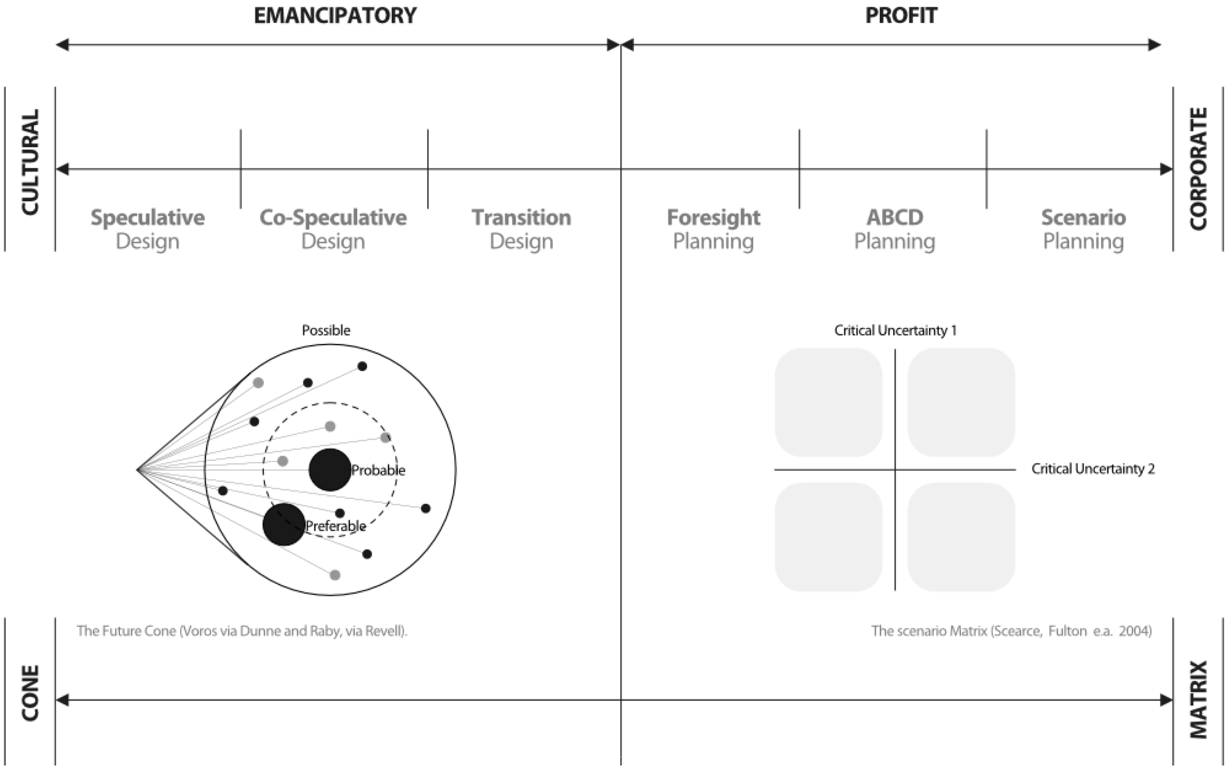
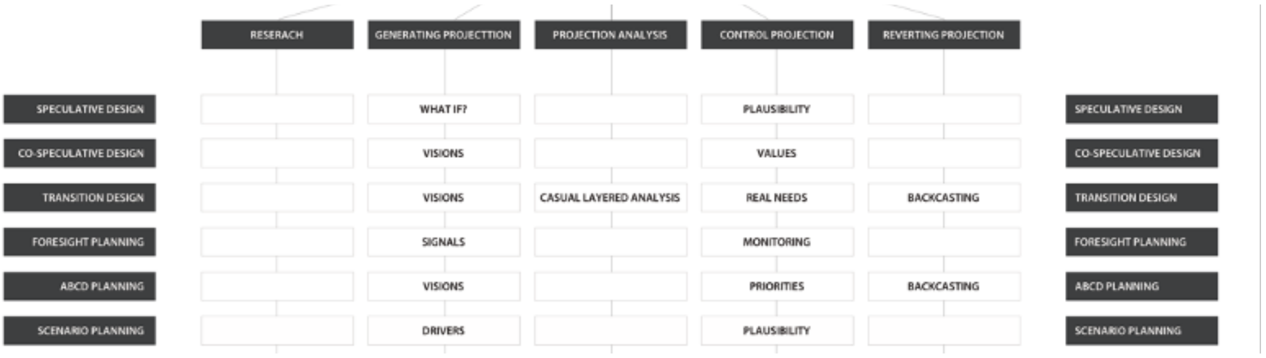


Fig 8. Design futures spectrum and their main embodiments. F. Galdon

Although these perspectives have been widely used, they present limitations. The scientific/positivistic approach is perceived as objective and value-neutral. However, it is also perceived as presenting a narrowness of focus (only one possible future) and a lack of contextual awareness. From this perspective, Richard Buckminster Fuller called for an “industrially realisable design science’ "(Fuller, 1957) through his "Eight strategies for a comprehensive anticipatory design science". However, this failed to materialise as a new field. On the other hand, the pluralistic approach is perceived as inclusive and partial. However, it is also perceived as presenting a loose focus (too many possible futures) and is too dependent on contextual awareness (Gidley, 2017).

In this study, as stated in Chapter 1, I position my inquiry in the second part of the third wave of relational design practice. This position presents trust as a fundamental element of design, as I have to address unsupervised systems, synthetic autonomy, unintended consequences, not-fully-knowing, reparation and accountability, and the ubiquity of fluid cyber-blended, and hyper-connected ecologies. The nature of the system of interaction demands a different type of design. In this context, I conducted a comparative study to underpin the key steps and strategies of the six methods outlined earlier (Fig. 9).

In this comparative study, I have structured this analysis around five questions I consider critical to building trust in design futures: does this methodology integrate historical research in the development of the technology at hand as a starting point in the process? How does this methodology generate the projection? How does this methodology critically analyse the projection? How does this methodology control the projection to avoid superficial and media-friendly outputs? How does this methodology transform the projection into a real-world executable action?



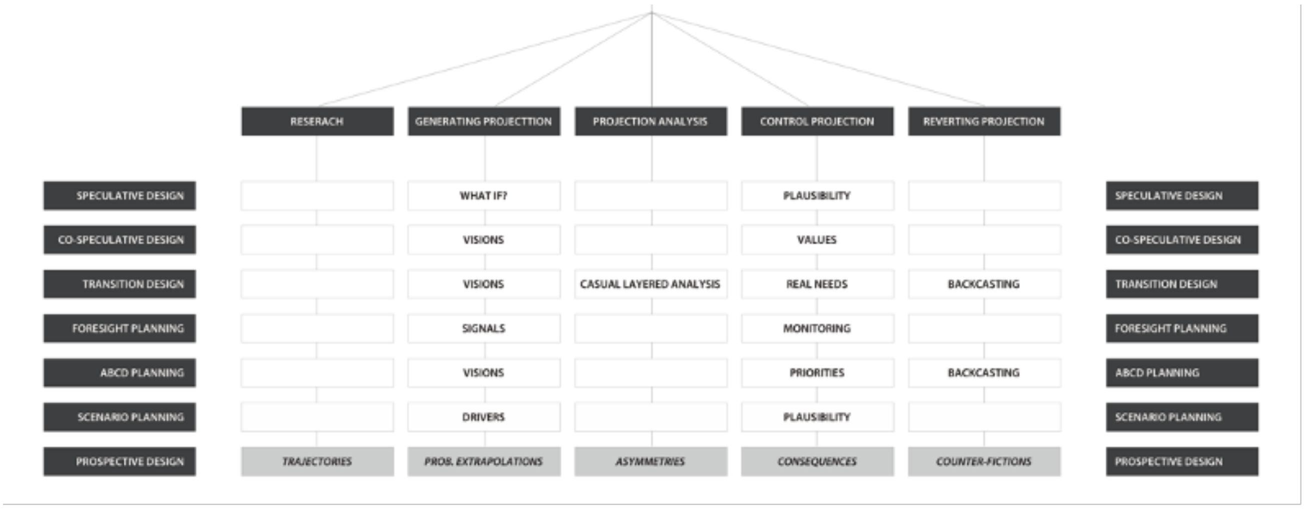
	RESERACH	GENERATING PROJECTTION	PROJECTION ANALYSIS	CONTROL PROJECTION	REVERTING PROJECTION	
SPECULATIVE DESIGN		WHAT IF?		PLAUSIBILITY		SPECULATIVE DESIGN
CO-SPECULATIVE DESIGN		VISIONS		VALUES		CO-SPECULATIVE DESIGN
TRANSITION DESIGN		VISIONS	CASUAL LAYERED ANALYSIS	REAL NEEDS	BACKCASTING	TRANSITION DESIGN
FORESIGHT PLANNING		SIGNALS		MONITORING		FORESIGHT PLANNING
ABCD PLANNING		VISIONS		PRIORITIES	BACKCASTING	ABCD PLANNING
SCENARIO PLANNING		DRIVERS		PLAUSIBILITY		SCENARIO PLANNING

Fig. 9 - Comparative study between the six models used in design to address the future. This process identified the lack of background research, projection analysis and reversing the projection as areas to consider for further development. These areas are fundamental for controlling the projection in prospective developments.

The first characteristic we can observe is that they start by generating a projection. This aspect may be due to the utilisation of design futures to generate potential applications for upcoming technology coming from the lab. How this projection is enabled varies between the methods. Some of them use visions, other values, signals, or drivers, and Speculative Design uses ‘what if …?’ questions. In terms of analysing the projection, only Transition Design (TD) provides a method: Causal Layered Analysis (CLA). This method is structured in four levels: The Litany; Systemic Causes; Worldview/Discourses, and Myth/Metaphor. This method is interesting but really difficult to implement. It is very broad, and some of the levels are too open to interpretation. In the Systemic Causes level, for instance, "Interpretation and communication is often undertaken by policy institutes, editorial news articles and non-academic journals" (Irwing, 2015). And the Myth/Metaphor level assumes that people can explain their visceral emotions. In terms of methods used by these outlined approaches to control the projection, these range from plausibility to values, to real needs or priorities. In terms of one of the most broadly used methodologies; Speculative Design, this limits the validity of its outcome to plausibility (Auger, 2012). However, it creates a lateral problem: difficulties in controlling the speculation. As a result, many of the proposed outputs end in what Future Studies expert Jennifer Gidley names 'Pop futurism' (superficial and media-friendly outputs) (Gidley, 2017). This problem is also translated to other practices. Finally, only two methods, TD and ABCD, propose a technique to ground the projection: back-casting.

2.4 PROSPECTIVE DESIGN - MODEL DEVELOPMENT

In the study conducted, I have considered all the limitations outlined and I will propose now a mixed methodology aimed at combining and enhancing the positive side of each approach addressed and present an integrative model aiming to reconcile different perspectives to improve the main task of design in our unpredictable and exponential technological age: prospecting the future. Building from these insights, the author proposes trajectories, probabilistic extrapolations, asymmetries, consequences, and counter-fictions (Fig. 10), as potential methods to address the issues outlined above.



	RESERACH	GENERATING PROJECTTION	PROJECTION ANALYSIS	CONTROL PROJECTION	REVERTING PROJECTION	
SPECULATIVE DESIGN		WHAT IF?		PLAUSIBILITY		SPECULATIVE DESIGN
CO-SPECULATIVE DESIGN		VISIONS		VALUES		CO-SPECULATIVE DESIGN
TRANSITION DESIGN		VISIONS	CASUAL LAYERED ANALYSIS	REAL NEEDS	BACKCASTING	TRANSITION DESIGN
FORESIGHT PLANNING		SIGNALS		MONITORING		FORESIGHT PLANNING
ABCD PLANNING		VISIONS		PRIORITIES	BACKCASTING	ABCD PLANNING
SCENARIO PLANNING		DRIVERS		PLAUSIBILITY		SCENARIO PLANNING
PROSPECTIVE DESIGN	*TRAJECTORIES*	*PROB. EXTRAPOLATIONS*	*ASYMMETRIES*	*CONSEQUENCES*	*COUNTER-FICTIONS*	PROSPECTIVE DESIGN

Fig. 10 - Comparative study between the six models used in design to address the future. This process identified the limitation of historical background research in technological developments as starting point in the process, projection analysis and reversing the projection as areas to consider for further development. Building from this analysis, the bottom of this diagram presents a set of methods to build a more reliable and mixed-method model to address and mitigate uncertainty and risk in design futures. (Fernando Galdon, 2020)

In order to develop these methods, I will implement a research-led design perspective. As stated in the introduction, research-led design is a process of design that relies heavily on hard data and research, either qualitative or quantitative, to inform design decisions, rather than relying solely on the expertise and experience of the designer (Saunders, 2008). In this model (fig. 11) methods relate to human factors, participatory methods (described in the matrix as “Scandinavian methods”), ethnography, usability testing, and contextual inquiry. They rely on two main approaches for enactment and validation: the Participatory Design quadrant (lower right) sees users as active partners in co-creation activities, and the Human Factors quadrant (lower left) sees users as subjects (reactive informers).

In this PhD, I will implement a mixed model by devising techniques to address the inquiry at hand. They will use the defined mindset accordingly.

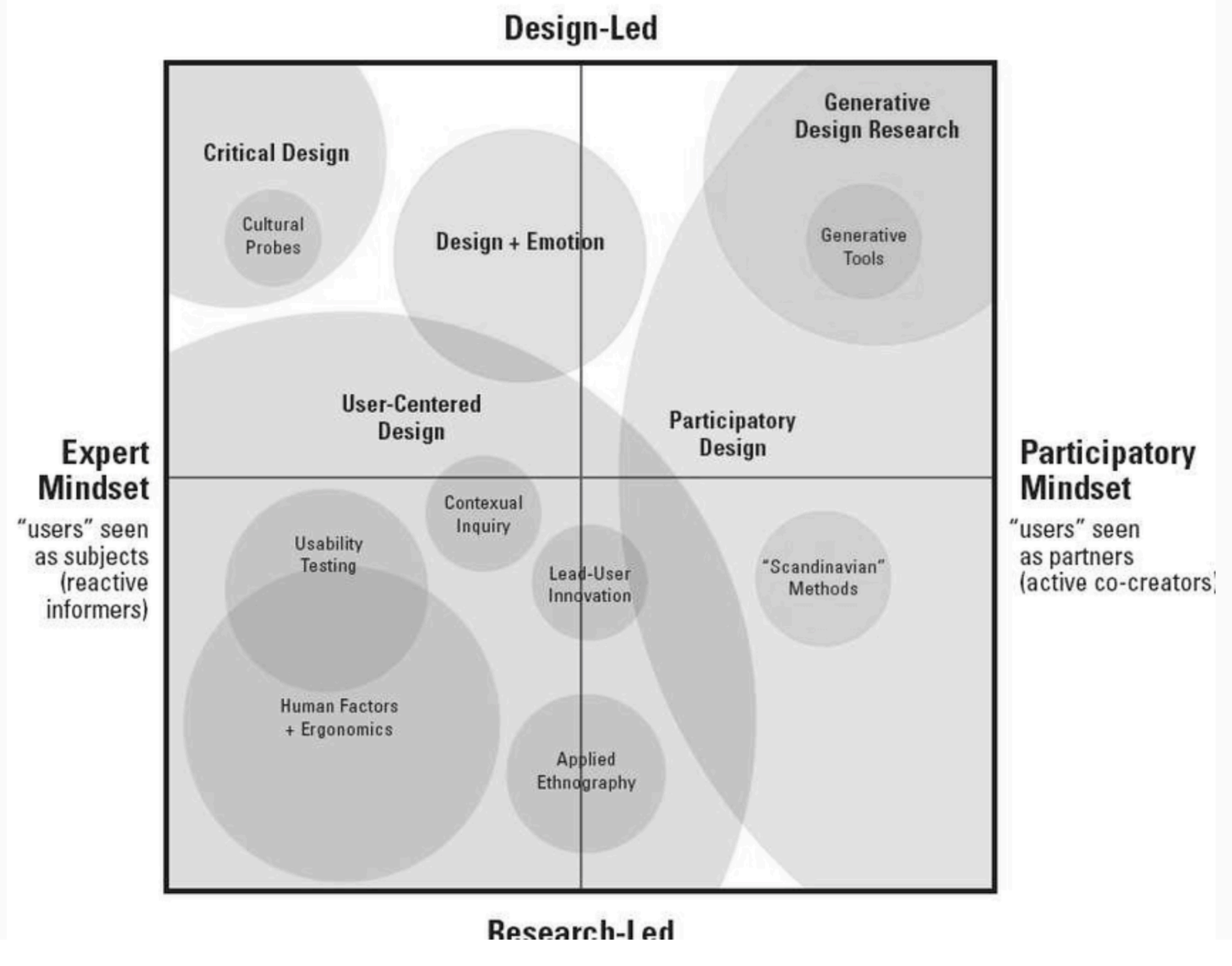


Fig. 11 - Design led versus Research led design. (Saunders, 2008)

2.4.1 Trajectories

Trajectories are used to embody and structure background research (literature review and archival research).

In this research, I have embodied them on Timelines. They can be used as graphical projective tools to gain a contextual understanding of the technology or topic at hand and project a possible trajectory based on relational patterns. By combining different themes, unknown relations emerge. I have approached its design mainly by dividing the space into two equal parts by drawing the timeframe in the middle. This action immediately creates two spaces which are used as comparative or relational spaces for relational and prospective inquiry and analysis, with the aim of spatialising abductive thinking between lines of research

2.4.2 Probabilistic extrapolations

As we are projecting the interaction into the future, questions of evidence regarding the prospective development and impact of emerging technology from a research perspective are raised. In this context, due to the limited access to emerging technologies by researchers (Mortier et al., 2014, p. 6), three elements will be used to underpin probabilistic extrapolations: demos, prototypes, and patents in the context of VAs.

Demos: Demos are introduced by tech companies to illustrate the potential of new technologies. They can be used by researchers to understand the potential development of emerging technologies.

Prototypes: Prototypes also present case studies for potential technological developments. Prototypes may raise ethical questions and illustrate how technology may impact our lives, either positively or negatively.

Patents: Patents illustrate the potential concrete development of a given technology.

These elements will allow me to map and triangulate potential technological developments and get a sense of their impact. This triangulation aims to help researchers to prospect potential positive and negative interactions and ground the inquiry in real-world prospective developments without limiting their potential, and/or critical analysis. In a sense, this method aims to integrate a reliable mechanism to focus the projection. The weight of each element in the triangulation may vary. In some cases one of the categories may be dominant in terms of impact or volume, while in others this distribution may be equally important.

2.4.3 Asymmetries

Asymmetries represent a fundamental addition to design, as they allow the researcher to identify where the problems will occur in the interaction. They aim to uncover potential areas of conflict, exploitation and injustice, which may have a significant impact on both society and business. Traditionally, three main methods have been implemented beyond design to deal with asymmetries in technology; technology assessments, ethical quandaries, and public engagement.

Technology assessments are the most direct and formal strategy for forecasting and controlling socio-technical futures. They have been widely used in the Western world as a result of the establishment of the US Office of Technology Assessment (Jasanoff, 2016). These assessments are typically tied to legislative processes. However, for Jasanoff, an expert on the ethics of invention, "[these processes] suffered from some of the defects of law-making-itself-captive to the politics of the present, dependent on uncertain public funding, and weakly responsive to the popular will or to rapid changes in circumstances". (Jasanoff, 2016). As an alternative, constructive technology assessments were introduced. They, in principle, looked more inclusive than the earlier technology assessments. However, their reach is limited, as they do not have "much to do with the public whose lives those projects would most directly have affected". (Jasanoff, 2016). As a result, this method is interesting but limited.

In ethical quandaries, ethics committees and public engagement exercises are valuable for clarifying issues. However, they also have shortcomings as mechanisms of democratic governance. Jasanoff builds her argument from cases in bioethics to demonstrate that deliberations are biased to please public opinion instead of developing ethical standards. She argues that institutional review boards "are not apt places for discussing the fundamental constitutional issues including· the very meaning of being human that [technological] revolution raises". In her view, ethics bodies, when tied too closely to the research enterprise, tend to operate with a tacit commitment not to burden their home institutions or their scientific stars with too many demands. (Jasanoff, 2016, p. 264).

Public engagements involve letting the public inside the preserves of decision-making. Jasanoff states that this approach "has proved only moderately successful in opening up entrenched traditions of decision making". (Jasanoff, 2016). One of the main issues at hand in this paradigm is whether the right public has expressed itself. Jasanoff presents how cases in both British and American administrations have found it challenging to deal with critical feedback, questioning its validity in the process.

These arguments have been further proven by Floridi. He has argued that these initiatives lack any sort of consistency and lead to "ethics bluewashing". For Floridi this approach is understood as "implementing superficial measures in favour of the ethical values and benefits of digital processes, products, services, or other solutions in order to appear more digitally ethical than one is." (Floridi 2019b, p. 187).

As we have seen, Jasanoff, an expert on the ethics of invention, provides a critical review of these procedures, and, based on an extraordinary number of supporting evidence, states that these processes, while interesting, are not sufficient to deal with the exponential nature of technological advancement.

Jasanoff's seminal book *The ethics of invention* is a testament to the limitations of sociological methods to address prospective technological development. As a conclusion, Jasanoff illustrates "how the power to set the rules of the game for governing technology rests with capital and industry, and not with the political representatives of the working, consuming, and too often suffering masses". (Jasanoff, 2016, p. 266). The future of governance is determined by design, and only prospective activities may access those spaces from a proactive perspective. Sociological strategies are reactive in nature, as they are limited by the present. Jasanoff's account presents an empirical need to enable a research space to address the rising concerns of exponential technological development. Moreover, design's prospective ontological nature may fulfil this requirement.

In conclusion, Jasanoff presents the concept of 'asymmetries of anticipation'. In this context, Jasanoff introduces the dichotomies of deontological vs utilitarian, government policies vs technological solutions, and emancipation vs control as a critical arena for design solutions. Therefore, in order to understand the potential positive and negative dynamics of the system, asymmetries need to be understood and identified. They uncover potential areas of conflict, exploitation, and injustice, which may have a significant impact on society and businesses.

In the area of asymmetries, I will use case studies to address the issues of impact that are typically found in automated systems in the context of VAs.

2.4.4 Consequences

This area aims to integrate ethical analysis into the development of new products and services. Ethics focuses on how a person should behave. It is a philosophy applicable to daily life or existence. It integrates two areas in order to determine rules or codes of conduct: philosophy – the art of asking questions – and morality – what is good or bad. Its main objective is to determine the right thing to do. Its ontology is based on creating social constructs for the adequate functioning of society. Its epistemology decodes these constructs while its output aims to set standards of behaviour for daily life. This process will be structured in a three-level consequential analysis addressing unintended consequences, contexts, and actions.

In this area, I will use co-design workshops to map and analyse the potential consequences of a given technological development – in this case highly automated VAs – to address issues of impact and the representation of uncertainty and multiple variables and contextuality that are typically found in automated systems in the context of VAs. Building on Jasanoff's analysis, designers have been identified as the ideal partners/participants to map, develop, and implement these interventions.

2.4.5 Counter-fictions

Counter-fiction is an emerging experimental area in design practice. So far, only two publications were found during this research that explores its possibilities – a monographic journal issue (Multitudes, 2012), and a book (Belliot, 2018). This approach aims to address the relations of domination. Its primary approach, rather than being imposed or forced, is based on the co-production of emancipatory projects aimed to decrease repression and enhance individual freedom and responsibility by reversing power. In this paradigm,

> Freedom is nothing other than the correlative of the implementation of security devices. A form of power announced as "near future" or immediate present, which makes obsolete old forms of resistance still indexed on disciplines and forces us to invent "new weapons" (Foucault, in Claisse, 2012, p.108)

Trust is the main element to account for. It is understood as the mode of relationships between individuals. In this relational perspective, power is a dynamic and reciprocal force, addressed through asymmetric relations in which the one who is controlled sees their actions, cognitions, and possible effects reduced, although not determined, by the controller. Power can be seen as a relation or as an influence and differs from the point of view of the spectrum of possibilities controlled by individuals. This approach places trust as a fundamental variable with which to build and maintain the relationship.

Counter-fiction is different from counterfactual. The latter is a method widely used in Critical and Speculative Design. The fundamental function of these is to connect the past with the present to generate alternative realities. The key example presented in Speculative Everything is: what if the Nazis had won the Second World War? What would the present be like? This method allows practitioners to transform the present into a fiction to propose alternative realities, in the same way, that they use Speculative Design to connect the future to the present and transform the present into a fiction to propose alternative realities.

This is not the intention of Prospective Design. PrD uses the fiction as a mean rather than as an end. What we are ***countering*** in PrD is not reality, to generate a fiction, in order to operationalise a fictional alternative; rather, we are ***countering*** the fiction to generate a reality, in order to operationalise a real alternative. In PrD we aim to transform the fiction into a real-world intervention. The PrD approach doesn't end with the fiction. It uses the fiction to trigger a response. In this model, fictions become objects of inquiry rather than end products. In a sense, we are flipping the directionality of the cone (Hancock & Bezol, 1994) against fictions. The use of the cone at this stage is illustrative: it represents the idea and the repositioning of the design projection in the proposition.

In this context, the use of counter-fictional strategies emerged for me as a strategy to address the dynamics of the system, but also as an experimental method to ground speculations. Its interventions can be placed a priori (before the interaction), meanwhile (during the interaction) or a posteriori (after the interaction). Its primary function is to reverse asymmetries of power and dominance in exponential contexts and/or technological development through design. Outputs are expected to range from tools to frameworks.

This approach aligns with the position adopted by Gonzatto et al, (2013): "If the future depends on people, there is no need to wait for it; people can start making it real right now." However, PrD aims to go beyond the enabling of debates, which is what the Gonzatto paper proposes, as well as designing future models in the emancipatory spectrum such as Speculative Design, Design Fiction, Transition Design, or Co-speculative Design. PrD aims to create real tools for potential risks.

2.4.6 Final model; Prospective design
- Behavioural - Relational methods based on ethics.

This approach is based on the systematic practice of relational system analysis to prospect and model prospective futures. The main methods used are historical data analysis, relational frameworks, and the systematic use of ethical methods. This approach uses trajectories, probabilistic extrapolations, asymmetries, consequences, and counter-fictions (Fig. 12).

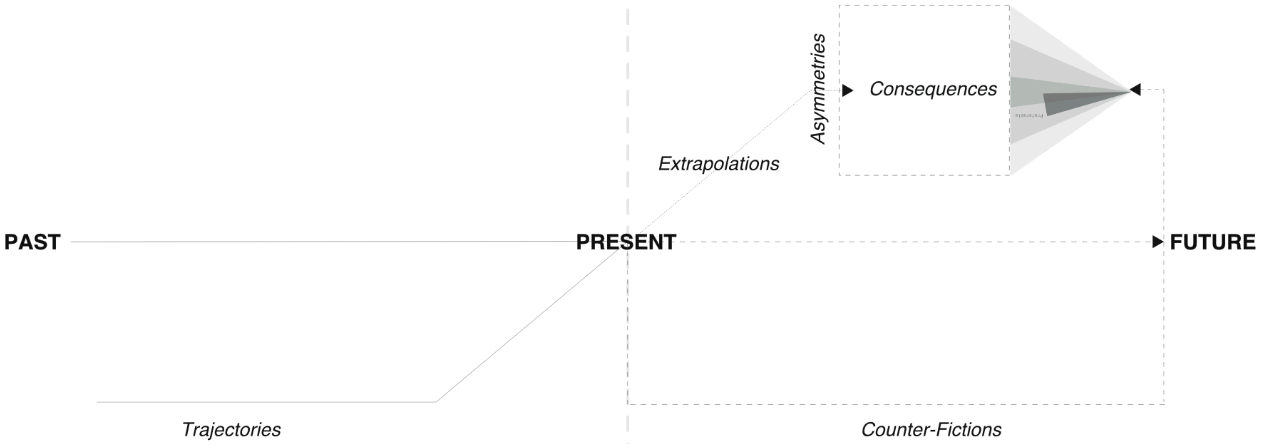
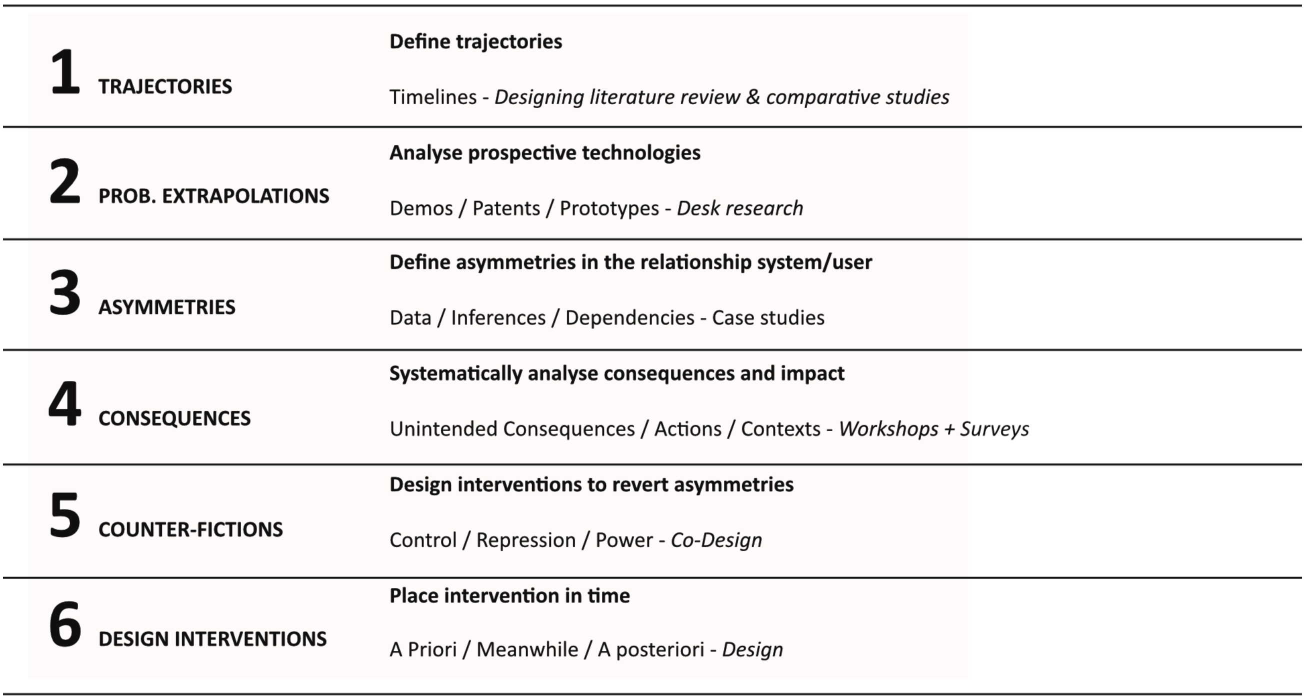


Fig. 12 - Diagrammatic distribution of methods in time to build the proposed Prospective model.



1	TRAJECTORIES	**Define trajectories** Timelines - *Designing literature review & comparative studies*
2	PROB. EXTRAPOLATIONS	**Analyse prospective technologies** Demos / Patents / Prototypes - *Desk research*
3	ASYMMETRIES	**Define asymmetries in the relationship system/user** Data / Inferences / Dependencies - Case studies
4	CONSEQUENCES	**Systematically analyse consequences and impact** Unintended Consequences / Actions / Contexts - *Workshops + Surveys*
5	COUNTER-FICTIONS	**Design interventions to revert asymmetries** Control / Repression / Power - *Co-Design*
6	DESIGN INTERVENTIONS	**Place intervention in time** A Priori / Meanwhile / A posteriori - *Design*

Tab. 3 - Preliminary analysis of Prospective design methods description and interventions. F. Galdon.

In the model presented, I have combined and developed existing models of designing (futures). This model presents some variations on established models such as Speculative Design, which revolves around reactive models based on “what if…?” questions. In the Prospective Design model, we integrated the strength of historical and contextual research to connect the past to the present to define technological trajectories. This process aims to overcome reactivity by bringing historical and contextual evolutive traces in technological developments. Then, we introduce probabilistic extrapolations to triangulate the future by analysing existing patents, prototypes, and demos. This process enables us to operate this method as an analytical tool to identify asymmetric problems in the system. Once we identify asymmetries, we conduct a three-level consequential analysis in order to map the impact of the asymmetry on the user. This process provides more focus than long-term and broader perspectives, such as TD. Finally, it inverts the futures cone to reverse the asymmetry via counter-fictions into a transformational action to generate emancipatory projects. Instead of framing the dystopia or utopia to generate a debate, it provides a systematic model to reframe them and transform the projection into a real-world intervention that aims to effect change.

In the process, this approach also challenges the dominant idea of anticipation in profit-driven methods, which aims to foresee what may happen and then waits for it to happen. Prospective research is directional and transformational. Building on Glanville’s work, its fundamental aim is to generate knowledge for future actions. It aims to generate preliminary insights to shape the future. The success of these interventions will be assessed by their potential impact and transferability to real-world interventions to effect real change. In this process, I propose that the prospective element can shape the future through probabilistic knowledge. This knowledge would enable this practice to be integrated into established models and structures of knowledge. However, what do I mean by probabilistic knowledge?

2.5 DISCUSSION

2.5.1 Design as a method in research

Design research practice emerged as a professionalised activity in the 1960s when the humanities and the sciences primarily dominated domain thinking. This "late arrival” forced designers to adapt design practice by using methods from other domains. An excellent example of this is Bruce Archer’s doctoral work, which attempted to explain design as a special branch of science (but, usefully, it failed to do so) (Boyd Davis, 2016). Other examples are Critical Design, Participatory Design, and Social Design, which could be argued to conduct aspects of social science through design. Even environmental design or engineering design could be thought of as “doing science through design”. In these cases, design is dissolved into a methodological process-based activity. If we position design as a data-gathering method, then we are tying design to the present. These aspects imply the dissolution of design as a discipline into views of the present, and this prevents it from being recognised as an independent domain that provides a different way of knowing, therefore positioning prospective disciplines such as design as dedicated practices, distinct from the sciences, arts and humanities. Furthermore, it questions the core ontology of design's knowledge base for transforming that which has yet to arrive.

In this context, design becomes secondary and is subjected to other disciplines' rules and mindsets. In this scenario, thinking is analytical. Reasoning is deduced. Moreover, knowledge must be factual by means of observation or measurement. In this context, abduction is denied. The traditional paradigm positions design as a method within research, which creates tensions that arise between the prospective nature of design and the factual requirements of working in the present. There is an ontological problem between the nature of design as future led and prospective and the nature of research which is present based and factual. I argue that the core nature of design is probabilistic research, not empirically driven research. We trade some degrees of accuracy for access to areas that are yet-to-be or not-fully-formed. Our output is therefore probabilistic, and research is always preliminary in its nature. Moreover, in exchange we provide guiding knowledge for prospective technological developments: "knowledge for"' instead of "knowledge of"'. We are concerned with how things "ought to be" Simon, 1996, p.111-167) instead of how things actually are.

2.5.2 Design as a discipline of the future

From this perspective, we would position design as a future-led prospective thinking activity in the context of abductive reasoning. In this scenario, as we are placing the projected potentiality in a society yet-to-be or not-fully-formed, it cannot be precisely measured or described, as it does not fully exist. In this context, as the designer is neither a scientist nor a sociologist (Cross, 1982, p. 221)(Galdon, 2021b), design cannot be experimental, as understood in scientific terms, nor observational, as understood in sociological terms, but transformational, as Aristotle suggested (Hall, 2011). Consequently, its output is based on potentialities, not certainties. In the same way, history is not about facts, but rather about approximations which are updated as new information emerges. In this context, as the life of the intervention is placed into the future, the time required to assess the impact of the design is extended during its lifetime. Validation is always a posteriori, and the proposed output becomes the main element to be assessed. The validity of the output generated, whether in a commercial or research context, will be judged by the transformational impact generated, which is defined by the level of exchange. This approach positions design research as the field of study of transformational systems.

This perspective also repositions the role of the designer from that of a facilitator to that of an expert in prospecting what could or should be done in the future. It challenges current ideas in the field by positioning the designer as an event gatherer, whose primary function is to facilitate exchange between experts. By repositioning designers as experts of the future and transformational systems, the role of the designer is to sit in the same room with equal status among other experts: to participate and to collaborate with them as equals. In this approach, the gathering of an event returns to sociological practices, and the designer is embraced as a prospective expert whose main duty is to develop and envisage the potential transformation from a knowledge-based technology to a future society. This repositioning does not aim to prevent designers from becoming facilitators or doing sociology through design. Instead, it aims to provide a new possibility for designers to act as experts and embrace the intrinsic perspective of their true ontological expertise.

2.5.3 Probabilistic knowledge

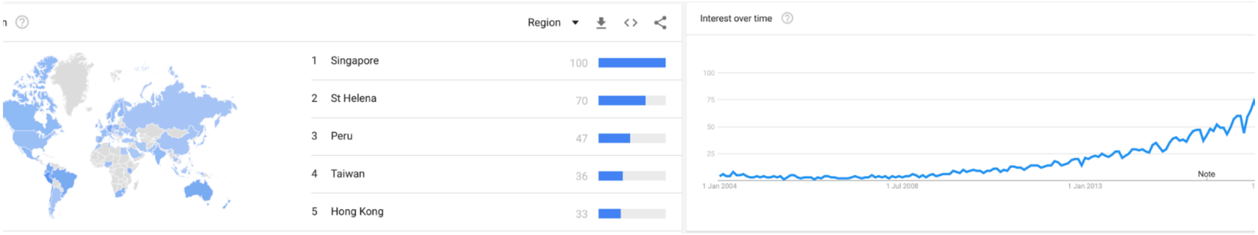
However, this future-led proposition presents a problem for the ontology of knowledge as currently configured. These established practices are limited by the present, and the researcher is the witness, either through measurement or observation. In this area, if we analyse what happens in economic research, we may find a suitable framework to solve this conundrum.

Economic forecasting is the process of making predictions about the economy. Many institutions, such as the International Monetary Fund, the World Bank, the Organisation for Economic Cooperation and Development, national governments, central banks, and private sector entities, including think-tanks, banks, consultant, and companies, use economic forecasting. Economists use statistical analysis of historical data to determine a forecast. Formal forecasts are produced once a year. However, quarterly updates or corrections are implemented to fine-tune the projection. The fundamental function of the economist is to anticipate future risks (i.e., events or conditions that can cause the result to vary from their initial estimates). These forecasts are continuously updated as the conditions of the environment evolve. These evolutions determine whether the adjustments will get tighter or looser, and how interest rates will vary, affecting a wide range of factors from loan repayments to employment levels.

At this point, a fundamental question arises: is this knowledge? Of course, it is knowledge. It is probabilistic knowledge of the future. Based on these economic forecasts, international institutions and governments implement all manner of adjustments that impact the lives of millions. From this perspective, economic research enables design to access the future by legitimising probabilistic knowledge as a valid type of knowledge. This element provides a bridge to reconcile the probabilistic nature of design with established frameworks of knowledge that have so far been understood as factual.

The fundamental difference with probabilistic knowledge in economics and policy-making and design research is the directionality of the action: instead of waiting for the anticipation to happen, in design research, we use this preliminary knowledge to co-shape the future. Design allows us to be proactive and move towards more imminent future transformations. The role of the prospective research-focused designer is to direct knowledge-based technological potentialities and reduce future risks to improve people's lives. As Aristotle proposed, productive knowledge is concerned more with standards of value than with setting boundaries of knowledge.

The main element to account for in this paradigm is the translational potentialities of the intervention: in other words, how to transform basic research into social and economic opportunities with the aim of enabling emancipatory projects. Nowadays, the value of research is not in the discovery but in the value and impact it returns to society. In this context, sociologists and scientists are struggling when presenting the future translational potentialities of their research, and many institutions are moving from fundamental to applied research to fulfil this shift. For instance, in sociology, building on the work of Pain, Gregson, and Olsen (Pain et al., 2010; Pain, 2014; Gregson et al., 2011; Olssen, 2015) the LSE's Impact Blog explains that "Anxiety around the impact agenda arises from the increasing instrumentalisation of knowledge, the corporatisation of UK higher education, and the relationship between assessment metrics and neoliberalism", as well as "fears that impact will prioritise certain kinds of knowledge" and "there are also concerns it rewards particular types of researchers" (Marchen, 2018). In a demonstration of the transformational nature of research output and impact, the LSE blog's author argues that researchers need to apply participatory action research to address the evolving nature of research (Marchen, 2018). Clearly, the translational imperative is starting to affect practices in sociology. In this context, cross-disciplinary collaborations between sociologists and designers may enhance the transformational potential of sociological inquiry. However, it seems that instead of fostering collaboration, which implies understanding the expertise of designers and treating them as equals, other disciplines are either rejecting the new reality in the research ecosystem or adopting design methodologies as part of their toolkit, rather than inserting designers into the research process. For instance, several science universities, such as Stanford University, the University of Maryland, and Ball State University in Indiana, have been integrating Design Thinking courses into their curriculums for some time (Morris, 2015). According to Dorst, Design Thinking is identified "as an exciting new paradigm for dealing with problems in many professions – most notably IT (e.g., Brooks, 2010) and Business (e.g., Martin 2009)" (Dorst, 2011, p.131). If we look at the term in Google Trends, we can observe an exponential increase in the term 'design thinking' (Fig. 13).



1	Singapore	100
2	St Helena	70
3	Peru	47
4	Taiwan	36
5	Hong Kong	33

Fig. 13. Design thinking evolution. From Google trends.

However, as Dorst points out, its adoption is much more complex than current simplifications. This reality positions design and designers centre stage as key partners in knowledge production and translation, with their expertise as catalysts for prospective transformations. In this context, the notion of probabilistic knowledge provides a framework that can be understood by other domains and enables designers to operate in the research field with their own ontological nature.

2.5.4 From time-based research to prospective interaction research

How do we approach PrD practice in the knowledge landscape? If we go back to the categorisation of knowledge presented by Aristotle, we can observe that he established three main categories: the theoretical, the practical, and the productive. Theoretical knowledge encompasses abstract subjects. It is concerned with things that are universal and necessary and cannot be applied. The idea that theoretical knowledge can never be utilitarian builds on the ancient sense of *theoria* as observation rather than participation. In contrast, the practical is applied and question-based; it has a beginning and an end. Finally, the productive is based on continuous interaction with the environment. It is transformational and represents a commitment to practice (Atwill, 1998). Therefore, prospective knowledge is defined as **the capacity to make involving prospective reasoning to 'go beyond' what exists and propose what can be 'otherwise'.**

These assertions and arguments question the reality of the current methodological nature of design, and confront the practice-based timeframe with a beginning and an end model imposed from the sciences and humanities. The nature of time-based industrial processes of knowledge production and traditional research approaches are affecting the very same nature of these transformations and potentialities.

Examples

- Prospective Actions

Recently, due to concerns about sustainability and the awareness of a range of emerging technologies that are transforming the future of our cities, the Swedish government decided to investigate prospective housing typologies. They selected a plot of land and invited a range of architects to present proposals for addressing rising concerns around sustainability and mobility. These proposals were completed by 2018 (Mallet, 2018). The experiment was finished; however, we do not know whether these new typologies are right or wrong. We need to wait another ten years to find out. In design research, knowledge is always a posteriori, and is always determined by the act of exchange, as proposed by Aristotle. And prospective decisions are based on the potential of the proposal to transform. **To 'go beyond' what exists and propose what can be 'otherwise'.**

- Prospective Practices

In fashion design, once a collection has been presented designers start to prepare their next collection. In this context, first, they research potentialities – colours, fabrics, new materials, and culture. From these reference points, they must generate prospective ideas (vision); then, design ideas are created (technical aspects of making), and finally, these ideas are presented to the public (the runway show). The designers must develop this prospective process without fully knowing how the world will be. They start a collection in September which will be presented in February, yet will be bought by consumers in the following September. At the time of the presentation, when the "experiment" is finished, they will know whether the designs have been technically well constructed, but will not know whether or not they will be successfully adopted in the marketplace. They have to wait some months to know whether they were the right designs or not. Design is a prospective activity, and productive **knowledge is always a posteriori and determined by the act of exchange**.

- Prospective Products

In terms of technology, another case can be illustrated by the iPhone. When the design is finished, we know if the camera works, whether it creates photos with the right number of pixels, whether the GPS is accurate and whether or not it is ergonomic. However, we do not know whether the iPhone will change future social and economic factors in the next two years. The iPhone X is better in many ways than any of its predecessors, as it has a better camera, a better screen, better sensors, and better software, etc. However, it is not being adopted at the same rate as previous versions were. A posteriori social, economic, and environmental factors affect the exchange mechanism. In design, productive knowledge is fundamentally prospective and always known a posteriori and **is affected and determined by social, economic, cultural, and environmental factors**.

Critical analysis

The iPhone is a paradigmatic case to understand how we are grasping the a posteriori impact of design as time evolves. In the first two years, we discovered that it had transformed the mobile phone industry. After five years, we discovered that it had transformed the manufacturing system. Over ten years, we are discovering that it has transformed society. Design processes based on scientific extrapolations could never have predicted the social implications of having a tracking device in your pocket capable of monitoring everything you do, and everywhere you go and using this information to manipulate society, trends, markets, and beliefs. Neither design-led science nor sociology could approach this a posteriori reality, as they are limited by what we do and have done, and how we have achieved it. In other words, an ontology of the past. As Glanville suggested, these practices are limited by the knowledge of the past (Glanville, 2005). However, a prospective approach to design, based on planning, problem shaping, synthesis, preparedness, readiness, and appropriateness, can provide a suitable framework to access these future spaces for knowledge.

These same aspects can be seen in the government-led prospective transformation to investigate future typologies in the city. Neither design-led science nor sociology can grasp potential developments, as they are limited by the present, either by measurement or observation. Nevertheless, the government must act now.

Finally, design practices may be understood as practices aiming for personal fulfilment or personal development. However, the author addresses the applied nature of these disciplines, aiming to go beyond personal transformation to deliver practical interventions to transform society. This implies exchange beyond oneself that involves social, economic, and environmental activities forever bound to their environment. When you finish your “experiment”, whether is a fashion collection, or an iPhone, or a house, or a song, or a theatrical play, or a film, or a book, or an app, you do not know whether it will transform society. This will be known a posteriori, and its value will depend on whether there is an exchange or not. Design is therefore a prospective activity, and knowledge in design is probabilistic in nature and is determined by the level of change achieved after the “experiment" is finished. Therefore, design cannot be scientific in the empirical sense.

These examples demand a totally different type of knowledge, which is radically different from the arts and humanities/sciences divide observed by C.P. Snow. As described in the examples presented in this section, we may know about their "technical”"aspects – for instance, their structural or material qualities, or whether they comply with a set of regulations; however, we do not know whether these are the right typologies for future living or the social impact they may inflict in years to come. In sociology or science led-design, once the experiment is finished, we know the answer, via measurement or observation. However, design is fundamentally a prospective activity, and this implies a probabilistic nature to the knowledge generated, as we are dealing with new propositions that evolve in time and are contextually dependent.

2.6 PROSPECTIVE DESIGN HYPOTHESIS

In this chapter, I have argued for the repositioning of the origin of design research to place it within an Aristotelian rationale of prospective and productive knowledge. This positioning implies that design research is always implicated and will remain in exchange, therefore becoming probabilistic in nature. In this context, prospective knowledge always redefines the subjects involved by effecting a shift in power and status through its transformational nature. It cannot transcend time, like mathematics, and depends on time, contexts, and circumstances. Therefore, it must assume past, present, and future timeframes, the impact of the environment, and changing future social and economic factors. It is instrumental and situated, and its value is social, economic, and environmental.

Design research is concerned with navigating competing standards of value rather than securing boundaries of knowledge, and its practice is based on the capacity to make new futures involving abductive reasoning. It is concerned with something coming into being, indicating that things can be otherwise and beyond the way they are currently configured. It is concerned with the indeterminate and the possible within alternative possibilities, from passive intellect (contemplation becoming its object) to active intellect (an object being defined) to prospective intellect (an object being transformational).

In the prospective framework, I have proposed that design research can access the future. However, current models of research are limited by the present, both by observation and measurement. In order to address this fundamental aspect, I present the concept of probabilistic knowledge by building on new approaches in design and economics. Probabilistic knowledge in the context of design research could be defined as the potential impact of transformational initiatives.

The value of design research as presented here is economic and social, and therefore aims for mixed methodologies to implement strategies to build informed interventions in order to support planning, solution-based problem solving, problem shaping, synthesis, preparedness and appropriateness in the built environment. These aspects are fundamental for the optimum development of society in an ever-evolving world, based on exponential technological developments. Until now these aspects have been inaccessible due to the currently limited frameworks of sociology or science that can only analyse what already exists. In this process, I propose to contribute to a contextualision of Glanville's concept of "knowledge for", transforming the future as a probabilistic knowledge ontology.

This intellectual framework enables Prospective Design to be fully operational in the context of research. This prospective nature excludes the designer from being a scientist or a sociologist and prevents design from being experimental or observational (in the scientific meaning of the term), as the projected potentiality is placed in a society yet-to-be or not-fully-formed. Therefore, it cannot be precisely measured or described, as it does not fully exist. This approach repositions the role of the designer from that of a facilitator to that of an expert in prospective future-led translational and transformational technological developments, in order to enhance knowledge-based technological potentialities and reduce future risks. These aspects are fundamental to approaching the design of trust in AI.

Autonomy requires and affords new ways of interrogating design research that departs from current models of inquiry and economically driven approaches that privilege the system's performance and its profitability. Instead, design strategies must focus on designing trust and propose a relational and prospective approach that is aimed directly at ensuring that the interactions of emerging HASs remain focused on the user's needs and preferences. One of the fundamental characteristics of AI is that it continues to evolve, and may lead to unexpected events; therefore the design of these systems needs to operate continuously in a future space to address potential consequences. In this context, the probabilistic nature of design enables us to operationalise this space to design trusted systems. These are issues that have until now been absent from current fields and approaches in research. Recognising these attributes will lead designers to address research questions from an ethical perspective that seeks to improve relationality and the influence that systems have on the prospects of an interaction

In this chapter, I have developed a methodology for design practice to address the design of trust in AI: Prospective Design. This future-led mixed methodology incorporates trajectories, probabilistic extrapolations, asymmetries, consequential analysis, and counter-fictions to design novel strategies to mitigate the unintended consequences of prospective technological developments. Chapter 3 will present an implementation process for the proposed model based on a research-led perspective (Saunders, 2008).

2.7 PROSPECTIVE DESIGN RESEARCH IMPLEMENTATION

In terms of implementation, in order to address the nature of abductive/prospective research outputs, this Ph.D. will implement a cross-disciplinary and progressive publishing strategy. This process has identified conferences to implement a confirmation model to remove assumptions and consolidate knowledge from an external perspective. This strategy includes diversity, transversality, impact, relevance, and responsibility as fundamental variables to address.

The rationale to implement this strategy builds from a Parmenidean perspective of truth as a process (alḗtheia), and a Socratic perspective of multi-perspective dialectic ontology (ti estin). Parmenides built from Heraclitus's notion of reason (logos) to present the notion of truth (alḗtheia). Alḗtheia builds from ἀληθής (alēthḗs, “true”), and is composed by two elements ἀ- (a negative particle meaning, “not”), and λήθω (Lḗthē, “oblivion", "forgetfulness", or "concealment” (Liddell & Scott, 1940)). Alḗtheia (ἀλήθεια), through its privative alpha (ἀ-) means "un-forgetfulness" and/or "un-concealment". This proposition positions truth as a process of uncovering or discovering and unforgetting or remembering. Socrates built from this notion but challenged the idea of writing as it entailed conclusiveness. Instead, his dialectal ontology brought the public sphere and conversation as a method to establish the truth and positioned knowledge as an open-ended process in which knowledge could be altered. By testing his arguments with a multiplicity of wise men he could refine and test the robustness of his arguments. However, it violates the second rule proposed by Parmenides; unforgetting. The fundamental problem with conversation is that you tend to forget things. Writing, on the other hand, perpetuates knowledge in its original form.

Building from these notions, I acknowledged the potential of multi-perspectival evaluation of Socrates, but challenge his opposition to writing. By publishing preliminary papers knowledge can be scrutinised in a multiplicity of fields in its original form. Here I introduce preliminarity as a category to operate as a register but leaving the possibility of the knowledge generated to be challenged, evolved, modified, or falsified. In this way, we can reconcile and integrate the notions presented by Socrates and Parmenides. This process enables cross-disciplinary scrutiny to enhance robustness in the context of established models of research.

As a result, I will publish papers in a multiplicity of fields ranging from Industry 4.0, Human Factors or Design Research to Applied Science or Design Futures. This approach to practice aims to enhance the impact of the research in terms of outputs and scrutiny by diverse audiences to maximise its transversality and therefore, its robustness.

To further enhance its robustness, this Ph.D. will present its outputs to practitioners and public bodies. This research will be presented to a wide range of diverse audiences beyond academia, including design consultancies, start-ups, freelancers, NGOs, government bodies, the corporate sector, and professional researchers.

The implementation of a progressive and cross-disciplinary publishing strategy will allow me to mitigate assumptions in the process by contextualising and confirming my outputs progressively. This strategy will enable me to build robustness in the context of abductive research in design research. This framework allows the researcher to go beyond what exists and investigate the potentialities emerging from technological developments.

2.8 PROSPECTIVE DESIGN EVALUATION

To evaluate the final emerging methodology, a Quasi-experimental (Q-experimental) design methods perspective will be implemented. This approach presents an improvement to previous assessments models in design futures such as Speculative design (Auger, 2012) by integrating a control group into the process to evaluate whether using the methodology creates a difference or not. (This process was absent in James Auger's thesis).

According to Craig et al., these types of experiments occur when “a particular intervention has been implemented but the circumstances surrounding the implementation are not under the control of researchers” (Craig et al., 2012, 2011 in Leatherdale (2019, p.9). This type of experiment is employed as a study design when controlled experimentation is extremely difficult to implement (DiNardo, 2008; Dunning, 2012; Rosenzweig, 2000). These experiments are ideal when, for instance, “a new programme is implemented” (Leatherdale, 2019, p.19). This flexibility and approach to the new are crucial for design research in which the notions of full control of the variables and repeatability are impossible. These aspects of natural experiments make them ideal for evaluating a research methodology in the context of design.

In this case, I will implement an adaptation of the multiple Group post-test-only design, also known as the Nonequivalent Control Group Post-test-Only Design. In this type of design, the control group is non-equivalent, meaning that “participants are not assigned to either the experimental or the control group in a random manner” (Jackson, 2009, p. 323). They are members of each group because they have decided to participate in a specific [workshop] call. The pre-test will be unnecessary to establish equivalence between groups because all participants will be design students at the Royal College of Art and the workshops will be both about future technological developments of VAs.

The treatment will be the main variable (a simplified version of the methodology versus a more complete version of the methodology). And the post-test will analyse differences in outputs. In this context, the experimental group tests/assess the model as it is intended, and the control group is presented with a simplified version of the model. Therefore, the design, or in this case, the methodology, can be said to have caused some difference in outcomes between the experimental and control groups. In order to evaluate the final model, two workshops will be used to test critical aspects of the methodology proposed.

This process will be complemented with a contextual evaluation via a comparative study with the European Commission's latest White Paper on AI governance.

3 CHAPTER

MODEL IMPLEMENTATION

3.1 INTRODUCTION

In this chapter, I will be implementing Prospective design as a future-led mixed-methodology to address unintended consequences. The proposed novel methodology combines systems analysis with extrapolations and constructivist perspectives to address the rising concerns of exponential technological developments providing an applied ethical model for designing future(s) enhancing trust in the process. Its integrative nature aims to reconcile confronted models of design future(s).

In this context, as the life of the intervention is placed into the future, time to assess the impact of the design intervention is extended during its lifetime. Validation is always a posteriori, and the proposed output becomes the main element to be assessed. The validity of the output generated will be judged by its appropriateness and the potential transformational impact. The outputs generated as part of this implementation aim to provide guiding knowledge for prospective technological developments.

In the process, this thesis challenges and develops current notions of design research that are based on technological progress and revolve around product development towards a model based on ethical responsibility which places equal value on the process of design and the impact of the system on society. In this context, abductive thinking becomes the primary design mindset in driving the transition from current to potential states, leading to the mediation of anticipated and non-anticipated consequences. The Prospective Design framework, I argue, introduces a process to deal with the increasing complexity of wicked problems, black-box technologies, uncertainties, and AI/ML technology acceleration, enhancing social values and ethical principles in the process. In this process, trust becomes the main element of design.

3.2 TRUST, TECHNOLOGY AND THE ROLE OF DESIGN

In order to frame the intended outcome, an integrative review, which includes published and unpublished literature as well as practice work, has been conducted. It was decided to use this approach due to the small amount of published work in the area of trust. The search criteria were articulated based on their relevance to the subject. Books, academic journals, unpublished papers, and finally blogs and websites were searched, in that order. Online sources articulated the views of relevant practitioners and included reports hosted on the platforms of blue-chip companies such as the BBC or the Guardian to form and primarily support critical arguments. Less rigorous sources were used to reinforce or complement the main arguments. The exclusion criteria were based on the rigour of the arguments presented. In this case, some blog sites and publications were excluded by their irrelevance to the topic or because they presented uninformed arguments, and were articulated based on opinions rather than pieces of evidence.

The topic of trust in research can be traced back to the 1960s and 1970s in a range of influential exploratory work, such as Deutsch, 1973; Garfinkel, 1967; Rotter, 1967; Zand, 1972. In the 1980s and 1990s, research was implemented on conceptual aspects. This was followed by a wide range of empirical and experimental studies from the late 1990s to the present (see Bachmann and Zaheer, 2006; Möllering, 2006). Seppanen et al. (2007) provided guiding knowledge in which more than 70 definitions of the concept of trust were proposed (see also Castaldo, 2007). Several publications have adopted Rousseau's definition of trust as "the psychological state comprising the intention to accept vulnerability based upon positive expectations of the intentions or behaviour of another'" (Rousseau, 1998, p. 395). However, the most cited definition of trust is that by Mayer et al.: "The willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party" (Mayer et al., 1995, p. 712). (See Appendix B; Ch. 1 for expanded diagraming on trust).

In addressing the complexity of the definition of trust, Kaplan, Kessler and Hancock present an integrative review (Fig. 14). As they acknowledge, in all instances of trust there are three fundamental elements: "1) a trustor, who is doing the trusting and who is vulnerable to harm from another individual; 2) the trustee, who is the one being trusted, and who is the individual capable of harming the trustor; and 3) a context within which the trustee's actions are capable of causing harm or benefit to the trustor" (Kaplan et al., 2020, p.1150)

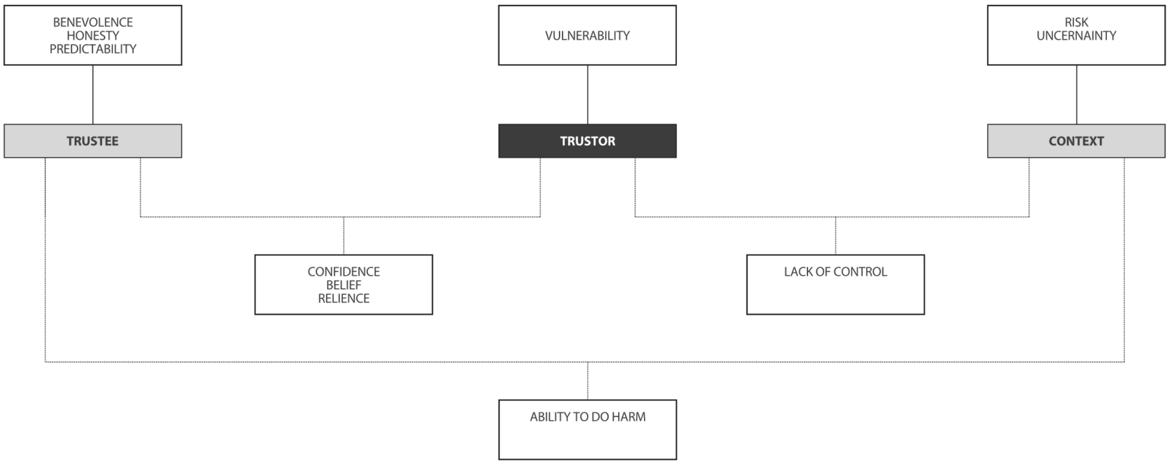


Fig 14. A model of the definitions of trust, by category (Kaplan et al. 2020, p.1151).

In this context, preliminary knowledge, past experiences, personality, and personal situation affect trust. Positive experiences reinforce trust. The work of Giddens defines trust as "confidence in the reliability of a person or system, regarding a given set of outcomes or events, where that confidence expresses a faith in the probity or love of another, or the correctness of abstract principles (technical knowledge)." (Giddens, in Lane and Bachman, 1998, p.35). It has also been expressed as the "Belief in someone or something, which is nurtured through positive experiences". (Kantar, 2016). Finally, building on Bachmann and Zaheer, 2013; Corritore et al., 2003; Lewicki et al., 2006; Fulmer and Gelfand, 2012; Schoorman et al., 2015 and Nienaber et al., 2015, trust can be defined as a relationship which arises between two units, the trustor and the trustee (Blobaum, 2016, pp. 3-5).

Traditionally, the design of trust in digital systems has been articulated through specific parts of the technology at the product level with a front-end perspective: for instance, certifications, badges, certified sellers, reviews, product guarantees, product grading (A, B, C), prime accounts or rating systems embedded in platforms (see Appendix 1 for an extended chronological analysis). If we analyse them, we realise that these processes relate to accountability rather than transparency. However, these strategies operate in a static, or reactive, way. These strategies, which have enabled a trusted internet, have not been recreated behind the scenes of autonomous systems. And this is particularly concerning as we transition towards automated systems with guiding interactions.

With the latest social events, what we are learning is that algorithms are not neutral. In this scenario, a debate has emerged on whether technologies are means or ends. In this context, Sheila Jasanoff (2016), in her seminal book *The ethics of invention*, presents technologies as means. She builds from examples of animals' behaviour to demonstrate that technology as an end to satisfy needs is not even a distinctively human action. For the author,

> "Technology, in short, is not merely about achieving ends that we already foresee but an open door to an uncharted, often uncertain future where current social understandings and practices may be fundamentally transformed. Uncertainty, moreover, can deter as much as it entices. The bright gleams of promise that invite human societies to invest in technology march hand in hand with darker misgivings about what could go wrong if the promises fail and the unexpected breakdown happens on a grand scale." (Jasanoff, 2016. p. 214).

She concludes,

> "Neither practicality nor predictability captures the evolving relations between human beings and their technologies. Human technological wizardry extends far beyond performances of repetitious tasks to serve simple, predetermined purposes. Artistry, imagination, and the desire to probe the unknown have long dominated the will to make and use technology" (Jasanoff, 2016. p. 212)

In this process, the design of trust is the design of systems of accountability. As we move into HASs that are primarily unsupervised, have the capability of learning and changing over time, are capable of dynamically setting their own goals, are able to adapt to local conditions via external information (sensors/input), and have the potential to evolve in unexpected ways, a fundamental question arises: **What would a future social and/or systematic structure/mechanism which could establish trust (systems of accountability) in highly automated systems be like?**

3.3 DESIGNING TRUST IN HIGHLY AUTOMATED SYSTEMS

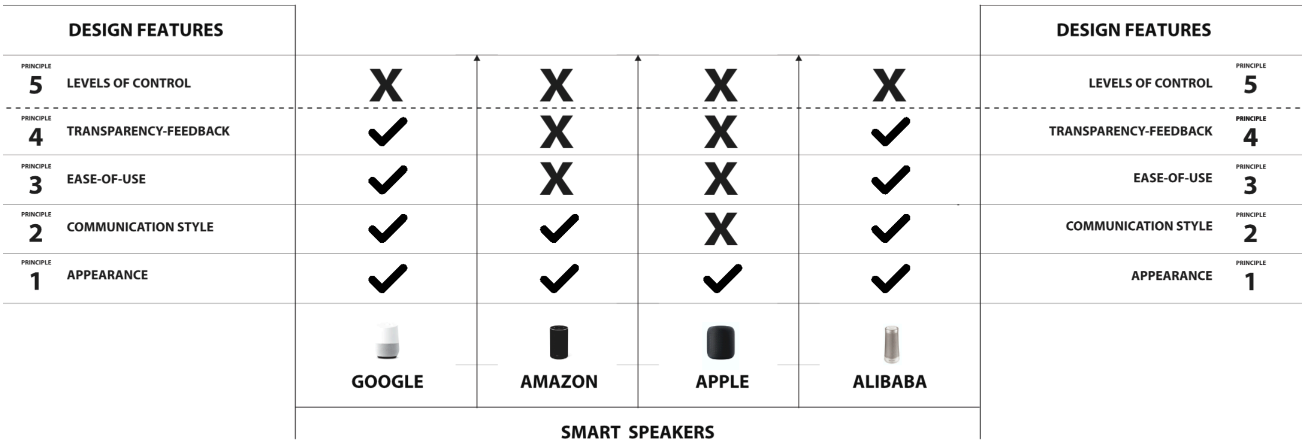
3.3.1 Trajectories

Preliminary research focused on the relationships between technology, sociological theory, and design practice (See 1.1). It has identified a range of impactful elements based on the potential developments of AI: the emergence of meta-agency, the emergence of an artificial subconscious, the relevance of algorithms, and the impact of HASs (ML/DL). These elements led to building a case around virtual assistants (VAs) as the main object of inquiry. This unit will embody the three main domains that will define future interactions (commercial, service-driven, and social). In this context, Hoff and Bashir (2015) conducted a literature review in the area of designing trust in automation and identified five fundamental design features.

Design feature		
Appearance	Increase the anthropomorphism of anthropomorphism automation in order to promote greater trust	de Visser et al. (2012); Pak, Fink, Price, Bass, & Sturre (2012)
Ease of use	Simplify interfaces and make automation easy to use to promote greater trust	Atoyan, Duquet, & Robert (2006); Gefen, Karahanna, & Straub (2003)b; Li & Yeh (2010)b; Ou & Sia (2010)b; Zhou (2011)
Communication style	Consider the gender, eye movements, normality of form, and chin shape of embodied computer agents to ensure an appearance of trustworthiness. Increase the politeness of an autonomous system's communication style to promote greater trust	Gong (2008)a; Green (2010)a; Lee (2008)a Parasuraman & Miller (2004); Spain & Madhavan (2009)
Transparency/feedback	Provide users with accurate, ongoing feedback concerning the reliability of automation and the situational factors that can affect its reliability in order to promote appropriate trust and improve Evaluate tendencies in how users interpret system reliability information displayed in different formats Consider providing operators with additional explanations for automation errors that occur early in the course of an interaction or on tasks likely to be perceived as "easy" in order to discourage automation disuse	Bagheri & Jamieson (2004); Bass, Baumgart, & Shepley (2013); Bean, Rice, & Keller (2011) Lacson, Wiegmann, & Madhavan (2005); Neyedli, Hollands, & Madhavan, Wiegmann, & Lacson, 2006; Manzey, Reichenbach, & Onnasch, 2012; Sanchez, 2006

Design feature		
Levels of control	Evaluate user preferences for levels of control based on psychological characteristics	Thropp (2006)

At this point, I conducted a comparative study among the most popular existing VAs (Google Home, Apple's HomePod, Amazon's Alexa, and Alibaba's Speaker) by mapping them against the five features of trust design presented by Hoff (2015) (Fig. 15). This study allowed me to understand the current state of the art and identified a gap in knowledge to define a design intervention. As can be seen in the table below, all these systems lacked Levels of Control (Level 5) specifically adapted to VAs. Once I identified the design intervention, I needed to understand how these systems might evolve.



DESIGN FEATURES	GOOGLE	AMAZON	APPLE	ALIBABA	DESIGN FEATURES
PRINCIPLE 5 LEVELS OF CONTROL	X	X	X	X	LEVELS OF CONTROL PRINCIPLE 5
PRINCIPLE 4 TRANSPARENCY-FEEDBACK	✓	X	X	✓	TRANSPARENCY-FEEDBACK PRINCIPLE 4
PRINCIPLE 3 EASE-OF-USE	✓	X	X	✓	EASE-OF-USE PRINCIPLE 3
PRINCIPLE 2 COMMUNICATION STYLE	✓	✓	X	✓	COMMUNICATION STYLE PRINCIPLE 2
PRINCIPLE 1 APPEARANCE	✓	✓	✓	✓	APPEARANCE PRINCIPLE 1

Fig. 15 - Comparative study of current Virtual Assistants against design features to design trust. This process identified levels of control lack of design in current VAs. This area will be fundamental in addressing uncertainty, autonomy and complexity in prospective developments in HAS interactions.

3.3.2 Probabilistic extrapolations

As we are projecting the interaction into the future, questions about evidence regarding the prospective development and impact of emerging technology are raised. In this context, due to the limited access to emerging technologies by researchers, three elements have been used to identify probabilistic extrapolations:

- Demos: Demos are introduced by tech companies to illustrate the potential of new technologies. They can be used by researchers to understand the potential development of emerging technologies. In this case, the author selected a demo called Duplex, introduced by Google. Its extraordinary levels of fluidity, coherence, and autonomy presented a scenario in which the evolutive nature of VAs, from queries to conversations and from reactive to proactive interactions, can be understood (Fig. 16).

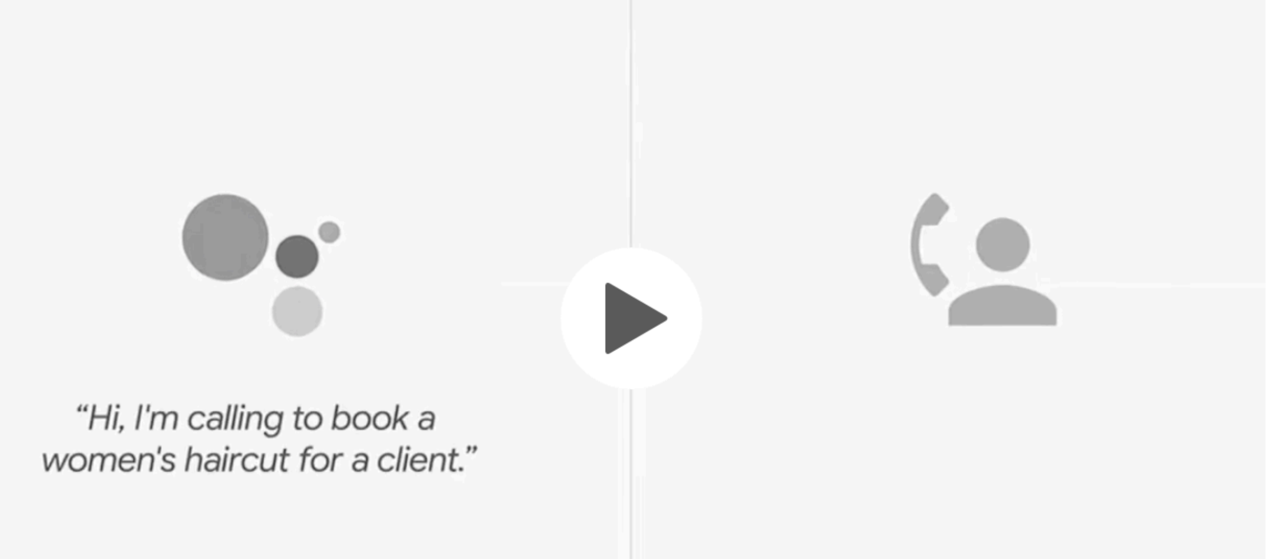


Fig. 16 - Duplex demo by Google. Source: Google.

- Prototypes: Prototypes also present case studies of potential technological developments. The author researched the current state of emergent algorithmic technology and identified eight distinctive prototypes that raise ethical concerns.

Prototypes		
Case 1	Predicting cough	- Amazon patent - voice - https://www.telegraph.co.uk/technology/2018/10/09/amazon-patents-new-alexa-feature-knows-offers-medicine/
Case 2	Predicting depression	- Facebook - https://www.iflscience.com/health-and-medicine/scientists-invent-algorithm-that-can-predict-depression-dignosis-from-your-facebook-updates/ - MIT - voice - http://news.mit.edu/2018/neural-network-model-detect-depression-conversations-0830
Case 3	Predicting best partner	- USC - http://www.bbc.com/future/story/20190111-artificial-intelligence-can-predict-a-relationships-future - loveflutter app - https://phys.org/news/2018-11-dating-apps-artificial-intelligence.html
Case 4	Predicting domestic violence	- Amazon - voice - https://www.independent.co.uk/life-style/gadgets-and-tech/alexa-relationship-dating-google-home-advice-imperial-college-research-a8658976.html

Prototypes		
Case 5	Predicting sexuality	- currently based on pics - https://www.theguardian.com/technology/2018/jul/07/artificial-intelligence-can-tell-your-sexuality-politics-surveillance-paul-lewis
Case 6	Predicting political orientation	- Facebook already does it - https://www.wired.com/2016/11/subtle-ways-digital-assistant-might-manipulate/
Case 7	Predicting best job	- Some companies working on it - google jobs - https://www.peoplemanagement.co.uk/long-reads/articles/recruiting-algorithms
Case 8	Predicting best investment	- Some banks are experimenting with it - https://economictimes.indiatimes.com/small-biz/startups/newsbuzz/evolution-of-voice-assistants-in-banking-from-simple-qa-to-personalized-advice/articleshow/62578778.cms

- Patents: Patents also illustrate the potential development of a given technology. As an example, the author researched patent applications to identify potential developments in the context of VAs. A clear case was a patent filed by Amazon for technology that is capable of diagnosing a cough and providing treatment. This patent aims to transform Alexa into a doctor and raises many ethical questions regarding its implementation (Fig. 17).

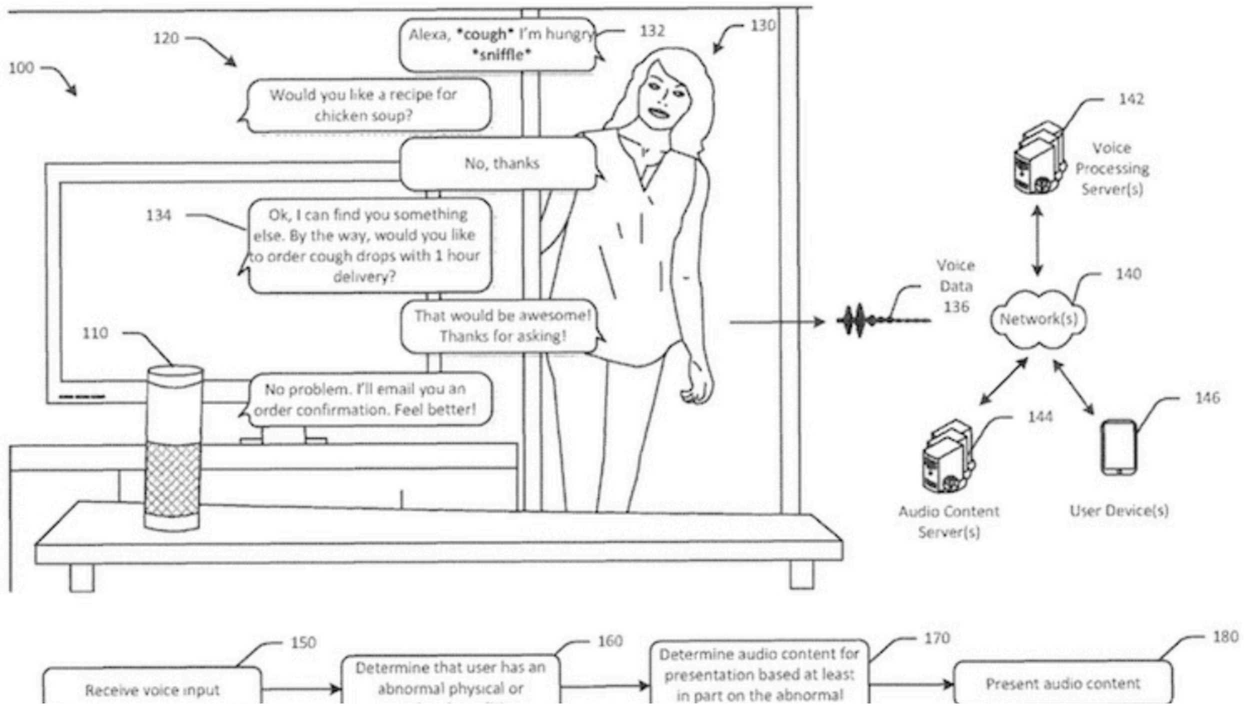

100
120
Alexa, *cough* I'm hungry *sniffle*
132
130
Would you like a recipe for chicken soup?
No, thanks
134
Ok, I can find you something else. By the way, would you like to order cough drops with 1 hour delivery?
That would be awesome! Thanks for asking!
110
No problem. I'll email you an order confirmation. Feel better!
142
Voice Processing Server(s)
Voice Data 136
140
Network(s)
144
146
Audio Content Server(s)
User Device(s)
150
Receive voice input
160
Determine that user has an abnormal physical or
170
Determine audio content for presentation based at least in part on the abnormal
180
Present audio content

Fig. 17 - Cough prediction algorithm patent. Source: Amazon (Jin, 2018)

In addition to this case, further research was carried out into patents to get a sense of prospective technological development in the context of VAs, the smart home, and smart atmospheres. In this context, four more cases emerged. These examples assisted in mapping the future of interactions and the potential impact of these new developments on users. Furthermore, they also revealed the leading players in the field: Amazon and Google.

Patents			
16/437763	Amazon	An Amazon patent application showed how a phone call between friends could be used to identify their interests.	http://appft.uspto.gov/netacgi/nph-Parser?Sect1=PTO2&Sect2=HITOFF&p=1&u=%2Fnetahtml%2FPTO%2Fsearch-bool.html&r=1&f=G&l=50&co1=AND&d=PG01&s1=amazon.AANM.&s2=conversational&OS=AANM/amazon+AND+conversational&RS=AANM/amazon+AND+conversational
14/639750	Google	Smart-Home Automation System that Suggests or Automatically Implements Selected Household Policies Based on Sensed Observations.	http://appft.uspto.gov/netacgi/nph-Parser?Sect1=PTO2&Sect2=HITOFF&p=1&u=%2Fnetahtml%2FPTO%2Fsearch-bool.html&r=3&f=G&l=50&co1=AND&d=PG01&s1=Google.AANM.&s2=mischief&OS=AANM/Google+AND+mischief&RS=AANM/Google+AND+mischief
15/943,860	Amazon	Application regarding personalising content for people while respecting their privacy noted that voices could be used to determine a speaker's mood using the "volume of the user's voice, detected breathing rate, crying and so forth," and medical condition "based on detected coughing, sneezing and so forth."	http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO2&Sect2=HITOFF&p=1&u=%2Fnetahtml%2FPTO%2Fsearch-bool.html&r=5&f=G&l=50&co1=AND&d=PTXT&s1=alexa&OS=alexa&RS=alexa
14/447487	Amazon	Sniffer algorithm "Keyword Determinations from Voice Data"	http://pdfaiw.uspto.gov/.aiw?docid=20140337131
14/638937	Google	Illustration from Patent Application, "Privacy-Aware Personalized Content for the Smart Home."	http://pdfaiw.uspto.gov/.aiw?docid=20160259308

The possibility of inferring potential and preliminary knowledge positions probabilistic extrapolations as an ideal method to address the potential impact of prospective technological development. The outputs aim to provide guiding knowledge. In the process, they are contributing to the contextualisation of Glanville's concept of 'knowledge for transforming the future' as a probabilistic knowledge ontology (Glanville, 2005).

3.3.3 Asymmetries

Building on the case study introduced earlier by Raphael (2018), we can establish that data asymmetries affect the relationship between the user and the system. This asymmetry is exploited via inferences (the knowledge and actions performed by the system based on the data gathered), which leads to tensions between integrity and intentionality (Fig. 18).

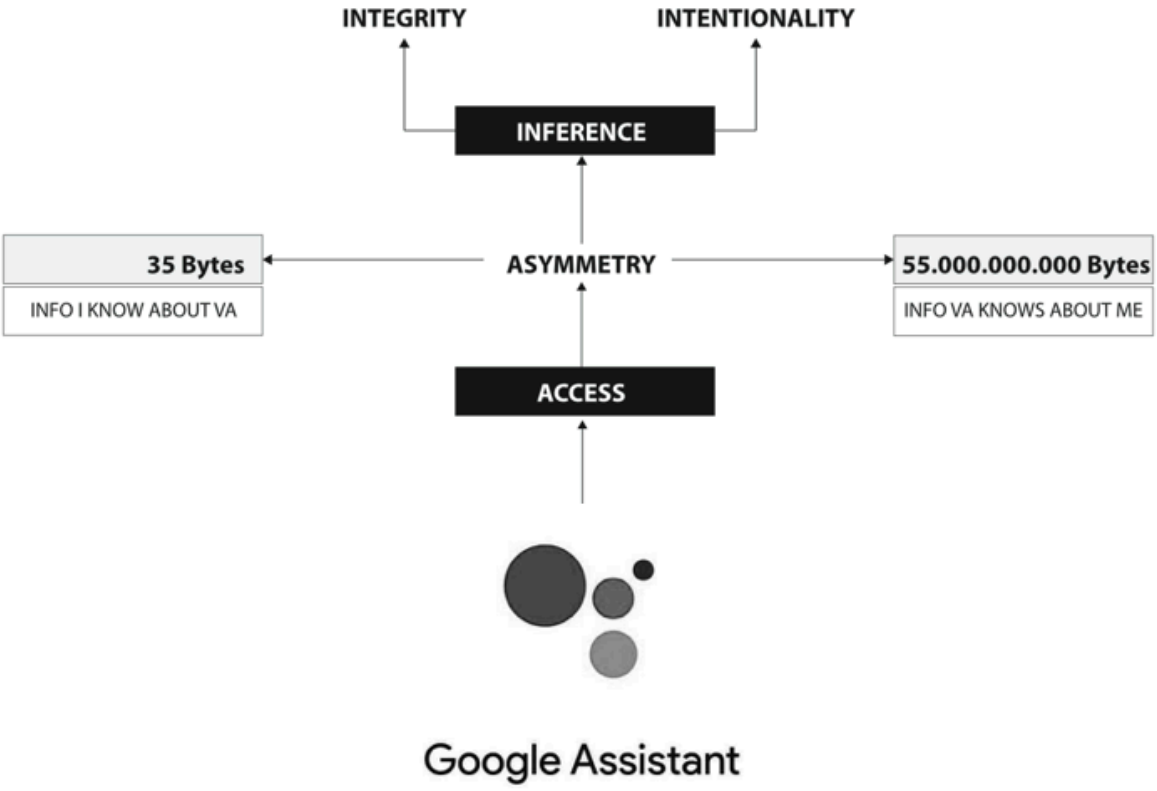


Fig. 18 - This diagram builds on Curran's (Raphael, 2018) research. It represents the asymmetry between the volume of information the system knows about you (5.5 GB) and what you know about the system (35 bytes). This information is what the system uses to extract patterns of activity and build inferences to generate dependencies.

In a recent experiment, Dylan Curran downloaded all his information from Google. The researcher presented evidence to demonstrate that Google had stored 5.5 GB of information (the equivalent of around three million Word documents) (Raphael, 2018). Google knows where you have been, what you search for, who your friends are, what you like and dislike, your plans, your preferences, the videos you watch on YouTube, and trends you are interested in. And we must point out that we do not know whether they are storing biometric data such as skin conductance, eye tracking, pupil dilation, or face recognition through third parties. Clearly, there are a range of data asymmetries between the system and the user in terms of data acquisition (personal, social, biometric, and environmental), knowledge extraction capabilities (patterns, routines, trends, preferences), monitoring (sensors, cameras, and microphones), and delivery (information quality and usefulness)

The identification of asymmetries enables the designer to infer potential and preliminary knowledge to address the potential impact of prospective technological development. The outputs aim to provide guiding knowledge.

3.3.4 Consequences

As introduced earlier, this area aims to integrate ethical analysis into the development of new products and services. Ethics focuses on behaviour. It is a philosophy applicable to daily life or existence. Its main objective is to determine the right thing to do.

Once the area was defined, I conducted a literature review on normative ethical frameworks. From this process, a debate emerged about which framework to use in the context of HASs: Socrates's virtue, Jeremy Bentham's consequentialism, Immanuel Kant's deontology, or John Dewey's pragmatism. (See Appendix B; Ch. 3 for expanded diagramming on ethics)

Virtue refers to being. In this paradigm, morality emerges from the identity of the individual rather than from their actions or consequences. Socrates' approach refers to an end to be sought. It asserts that the right action will be that chosen by a suitably 'virtuous' agent. Practical reason results in an action or decision.

Consequentialism states that the consequences of someone's actions are the ultimate basis for any judgment regarding that action. This perspective is non-descriptive, in the sense that its consequences, rather than its intentionality, determine the value of the action. It focuses on the outcome of conduct.

In deontology, the rightness or wrongness of actions does not depend on their consequences but on whether they fulfil our duty or not. A set of rules conditions these actions, whether natural, religious, or social.

Pragmatism aims for social reform as a strategy to address morality. Actions and consequences are possible because the context or system allows for them. Aimed at social innovation, from this perspective we should prioritise social reform over concerns about consequences, individual virtue, or duty.

Critical analysis

The fundamental problem with Dewey's perspective is that in order to change the system, we need an alternative or global consensus. As described by Yuval Harari, AI is a global problem, like climate change or nuclear war, which entails global consensus (Harari, 2019). In this context, some initiatives such as the EU's General Data Protection Regulation (GDPR) have been taking place, but they fell short and rapidly became redundant (Wachter, 2018). GDPR was introduced on 25 May 2018, and on 24 October 2019, the German government introduced a new framework (GGDEC, 2019). The lack of consensus and the limitations of access and rapid technological exponential development prevent Dewey's framework from addressing the design of a system.

In Socrates' virtue, the fundamental problem is the limited capability of humans to assess what is happening due to the acceleration and volume of information delivered by social interactions and algorithmic updates. These are fragmenting both our ability to reflect and our cognition by disconnecting the pre-frontal cortex through saturation. Our attention span has been reduced from 12 seconds to 8 seconds within four years by multitasking (NCBI, 2016) (Kahneman, 2011). After 21 minutes of comparing information our pre-frontal cortex shuts down (Mullins, 2013); in this context, only information with a significant emotional impact is retained (Buchanan, 2007). These processes are transforming society from one that is reflective to one that is reactive. The digital era brings Emotional Reactivism as its central paradigm. It questions the idea of truth and reality and repositions the focus of decision-making from reason to emotional experience, thus invalidating the model proposed by Socrates that was based on reason.

In this scenario, two leading candidates remain. On one hand, there is Jeremy Bentham's consequentialism; on the other, Immanuel Kant's deontology. The former focuses on the ethical intervention in the consequence, whereas in the latter it is focused on the intentionality of the system. In terms of deontology, the fundamental problems are interpretability and interruptibility. The system does not know what it is doing. Therefore, it cannot stop doing it. According to researchers from the most advanced AI company in the world, DeepMind, this is currently impossible to address (Ortega, 2018). Insofar as we are not capable of designing this, it is not a suitable strategy. Therefore, the only paradigm remaining is consequentialism. In this paradigm, the fundamental elements are the consequences of an action. Therefore, the system will be judged by the consequences of its actions. In this context, mapping the consequences, especially the unintended consequences, will be fundamental to understanding **What kind of social and or systematic structure/ mechanism could establish trust (systems of accountability) in HASs**

Development

Due to the highly contextual nature of VAs, a preliminary investigation, building on the prototypes and patents presented earlier, identified four highly sensitive areas where highly automated VAs may impact users significantly; health and wellbeing, identity, economically related activities, and social interactions. Once the relevant contexts were identified, a workshop was conducted with 20 postgraduate students from the School of Design at the Royal College of Art to map potential unintended consequences in these highly sensitive areas (Fig. 19). The decision to work with designers builds on Jasanoff's account in The ethics of invention. She positions 'benevolent' technologists, designers, and business people as the most qualified individuals to develop trusted systems. Consequently, my deontological position, and that of the participants, who are fundamentally designers/technologists/ entrepreneurs, has been crucial for the integration of applied ethics and emancipatory directionalities in collective activities. In this context, the diversity of the student body at the RCA in terms of their background, their programmes of study, their cultures and nationalities, and their diverse, critical, and enabling capabilities, plus the unifying element of being designers, provided an ideal group of participants to develop the task at hand. (See Appendix 3 for detailed workshop material).

Fig. 19. Workshop on Future developments of VAs at the RCA's Steven Building.

The participants were divided into four groups of five members. They used Mark Michaels' framework (Michael, 2019) to systematically analyse unintended consequences.

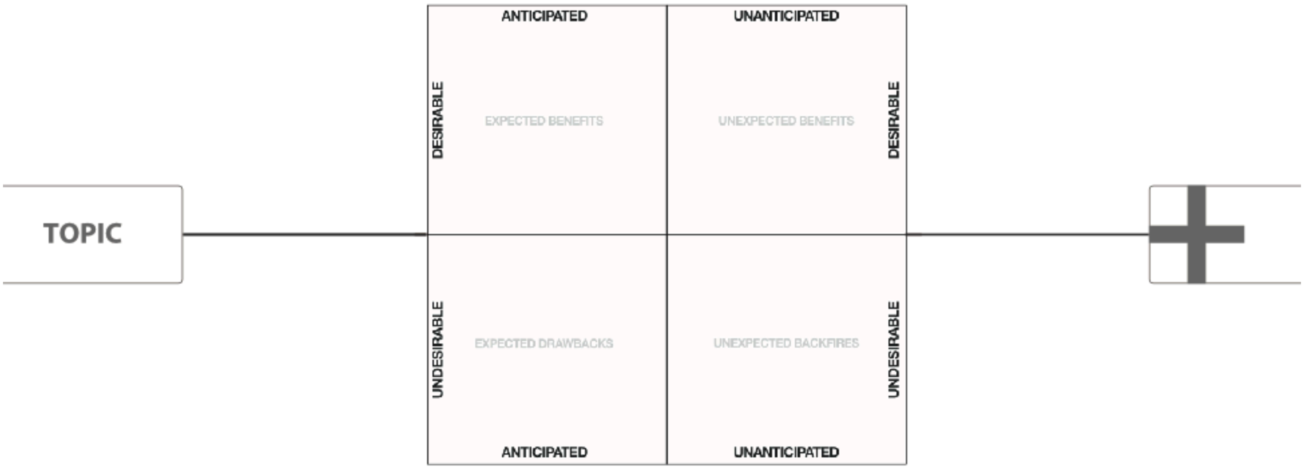
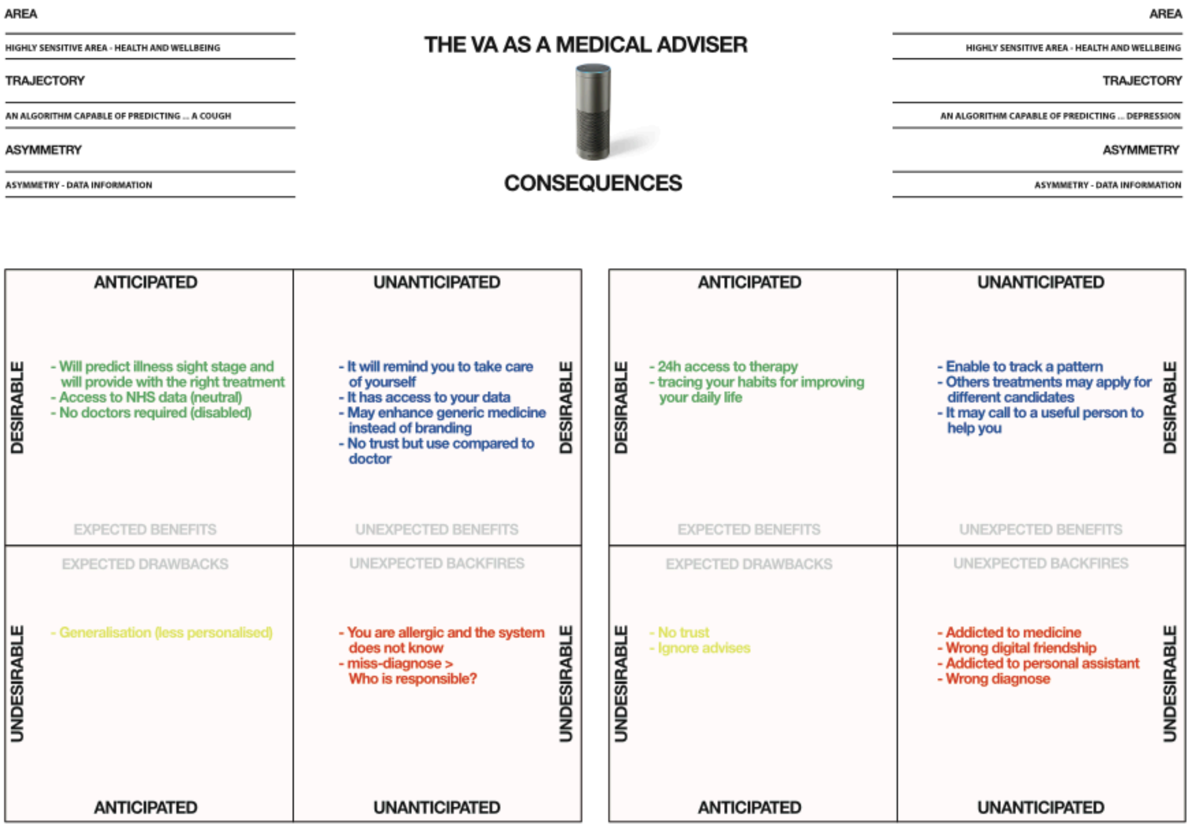


Fig. 20 - Unintended consequences toolkit as developed by M. Michaels (2019).

Given a specific technology, the framework presented by Michaels required participants to analyse four elements: anticipated desired, anticipated undesired, unanticipated desired and unanticipated undesired potential outputs. For one hour they debated among themselves to map consequences. Each of the group was given one of the highly sensitive areas previously identified (Fig. 21-24).



AREA

HIGHLY SENSITIVE AREA - HEALTH AND WELLBEING

TRAJECTORY

AN ALGORITHM CAPABLE OF PREDICTING ... A COUGH

ASYMMETRY

ASYMMETRY - DATA INFORMATION

THE VA AS A MEDICAL ADVISER

CONSEQUENCES

AREA

HIGHLY SENSITIVE AREA - HEALTH AND WELLBEING

TRAJECTORY

AN ALGORITHM CAPABLE OF PREDICTING ... DEPRESSION

ASYMMETRY

ASYMMETRY - DATA INFORMATION

ANTICIPATED — DESIRABLE

- Will predict illness sight stage and will provide with the right treatment
- Access to NHS data (neutral)
- No doctors required (disabled)

EXPECTED BENEFITS

UNANTICIPATED — DESIRABLE

- It will remind you to take care of yourself
- It has access to your data
- May enhance generic medicine instead of branding
- No trust but use compared to doctor

UNEXPECTED BENEFITS

EXPECTED DRAWBACKS

- Generalisation (less personalised)

UNDESIRABLE — ANTICIPATED

UNEXPECTED BACKFIRES

- You are allergic and the system does not know
- miss-diagnose > Who is responsible?

UNDESIRABLE — UNANTICIPATED

ANTICIPATED — DESIRABLE

- 24h access to therapy
- tracing your habits for improving your daily life

EXPECTED BENEFITS

UNANTICIPATED — DESIRABLE

- Enable to track a pattern
- Others treatments may apply for different candidates
- It may call to a useful person to help you

UNEXPECTED BENEFITS

EXPECTED DRAWBACKS

- No trust
- Ignore advises

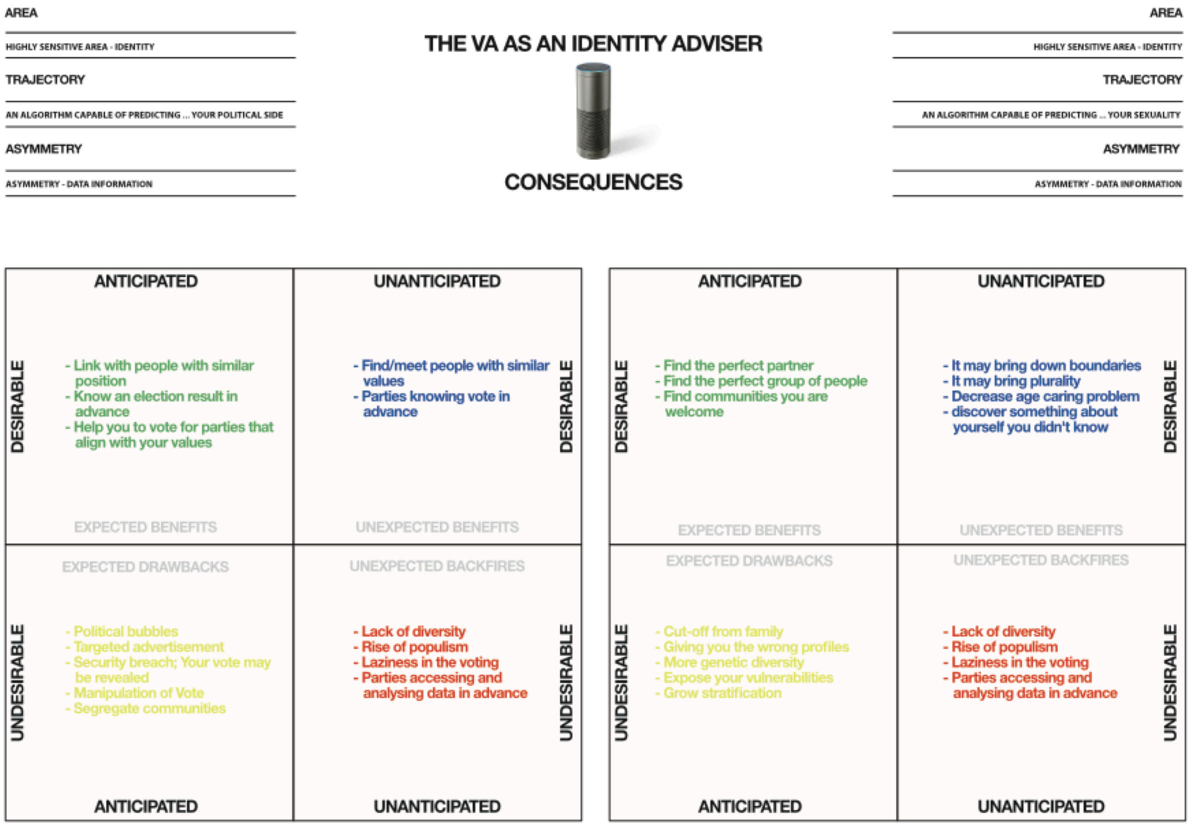
UNDESIRABLE — ANTICIPATED

UNEXPECTED BACKFIRES

- Addicted to medicine
- Wrong digital friendship
- Addicted to personal assistant
- Wrong diagnose

UNDESIRABLE — UNANTICIPATED

Fig. 21 - Results workshop group 1. F. Galdon



AREA

HIGHLY SENSITIVE AREA - IDENTITY

TRAJECTORY

AN ALGORITHM CAPABLE OF PREDICTING ... YOUR POLITICAL SIDE

ASYMMETRY

ASYMMETRY - DATA INFORMATION

THE VA AS AN IDENTITY ADVISER

CONSEQUENCES

AREA

HIGHLY SENSITIVE AREA - IDENTITY

TRAJECTORY

AN ALGORITHM CAPABLE OF PREDICTING ... YOUR SEXUALITY

ASYMMETRY

ASYMMETRY - DATA INFORMATION

ANTICIPATED — DESIRABLE

- Link with people with similar position
- Know an election result in advance
- Help you to vote for parties that align with your values

EXPECTED BENEFITS

UNANTICIPATED — DESIRABLE

- Find/meet people with similar values
- Parties knowing vote in advance

UNEXPECTED BENEFITS

EXPECTED DRAWBACKS

- Political bubbles
- Targeted advertisement
- Security breach; Your vote may be revealed
- Manipulation of Vote
- Segregate communities

UNDESIRABLE — ANTICIPATED

UNEXPECTED BACKFIRES

- Lack of diversity
- Rise of populism
- Laziness in the voting
- Parties accessing and analysing data in advance

UNDESIRABLE — UNANTICIPATED

ANTICIPATED — DESIRABLE

- Find the perfect partner
- Find the perfect group of people
- Find communities you are welcome

EXPECTED BENEFITS

UNANTICIPATED — DESIRABLE

- It may bring down boundaries
- It may bring plurality
- Decrease age caring problem
- discover something about yourself you didn't know

UNEXPECTED BENEFITS

EXPECTED DRAWBACKS

- Cut-off from family
- Giving you the wrong profiles
- More genetic diversity
- Expose your vulnerabilities
- Grow stratification

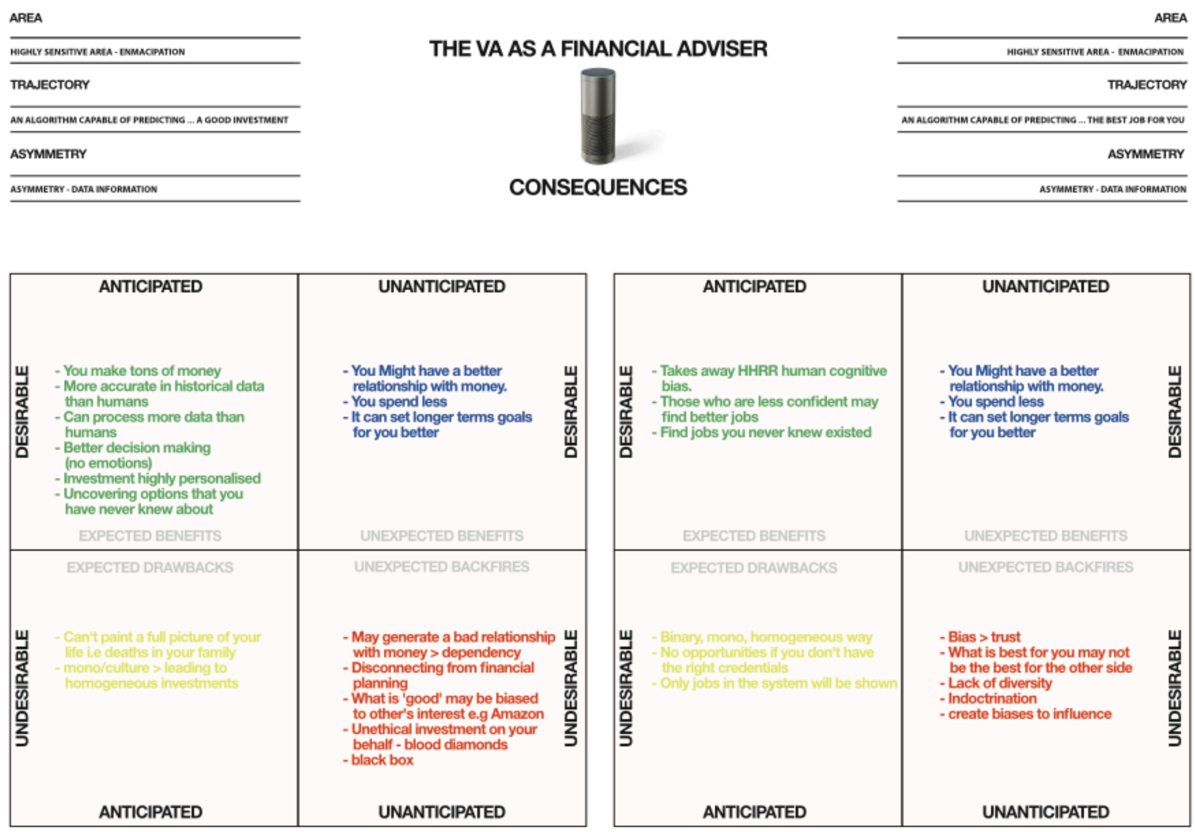
UNDESIRABLE — ANTICIPATED

UNEXPECTED BACKFIRES

- Lack of diversity
- Rise of populism
- Laziness in the voting
- Parties accessing and analysing data in advance

UNDESIRABLE — UNANTICIPATED

Fig. 22 - Results workshop group 2. F. Galdon



AREA

HIGHLY SENSITIVE AREA - ENMACIPATION

TRAJECTORY

AN ALGORITHM CAPABLE OF PREDICTING ... A GOOD INVESTMENT

ASYMMETRY

ASYMMETRY - DATA INFORMATION

THE VA AS A FINANCIAL ADVISER

CONSEQUENCES

AREA

HIGHLY SENSITIVE AREA - ENMACIPATION

TRAJECTORY

AN ALGORITHM CAPABLE OF PREDICTING ... THE BEST JOB FOR YOU

ASYMMETRY

ASYMMETRY - DATA INFORMATION

ANTICIPATED / DESIRABLE

- You make tons of money
- More accurate in historical data than humans
- Can process more data than humans
- Better decision making (no emotions)
- Investment highly personalised
- Uncovering options that you have never knew about

EXPECTED BENEFITS

UNANTICIPATED / DESIRABLE

- You Might have a better relationship with money.
- You spend less
- It can set longer terms goals for you better

UNEXPECTED BENEFITS

EXPECTED DRAWBACKS

- Can't paint a full picture of your life i.e deaths in your family
- mono/culture > leading to homogeneous investments

UNDESIRABLE / ANTICIPATED

UNEXPECTED BACKFIRES

- May generate a bad relationship with money > dependency
- Disconnecting from financial planning
- What is 'good' may be biased to other's interest e.g Amazon
- Unethical investment on your behalf - blood diamonds
- black box

UNDESIRABLE / UNANTICIPATED

ANTICIPATED / DESIRABLE

- Takes away HHRR human cognitive bias.
- Those who are less confident may find better jobs
- Find jobs you never knew existed

EXPECTED BENEFITS

UNANTICIPATED / DESIRABLE

- You Might have a better relationship with money.
- You spend less
- It can set longer terms goals for you better

UNEXPECTED BENEFITS

EXPECTED DRAWBACKS

- Binary, mono, homogeneous way
- No opportunities if you don't have the right credentials
- Only jobs in the system will be shown

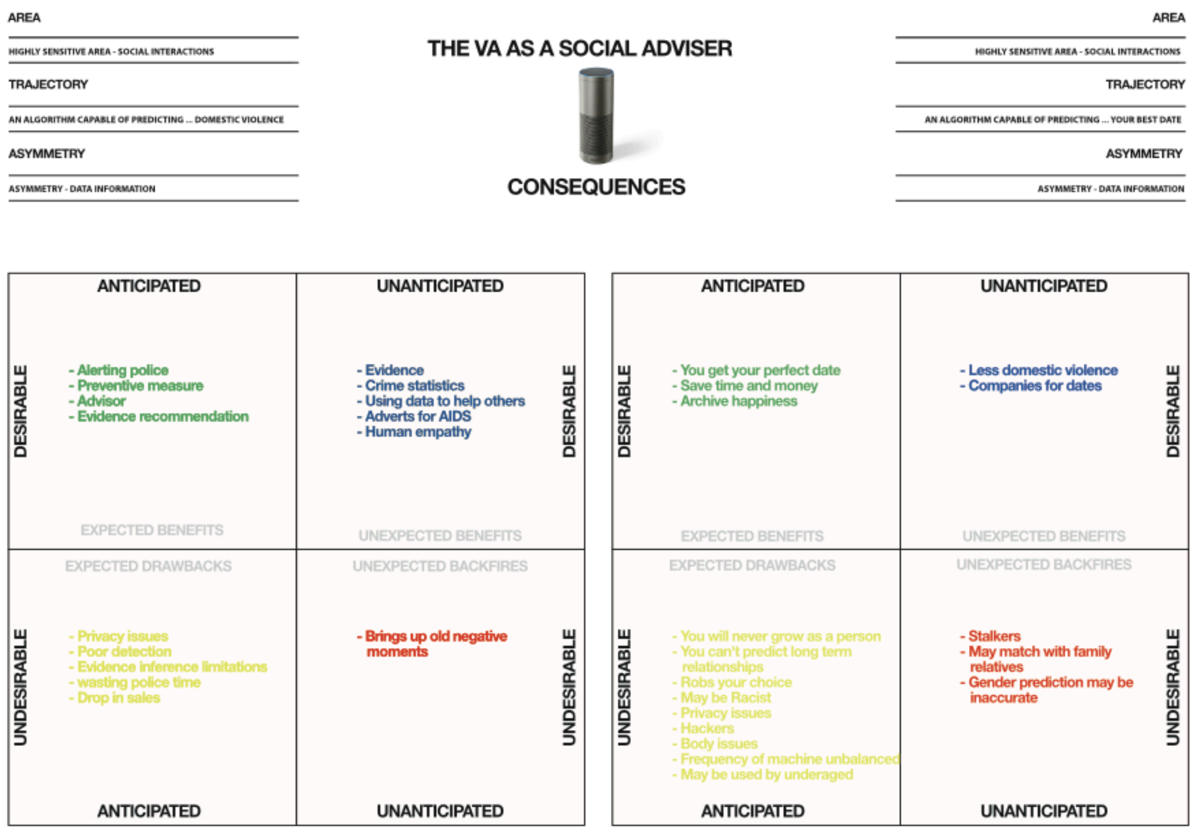
UNDESIRABLE / ANTICIPATED

UNEXPECTED BACKFIRES

- Bias > trust
- What is best for you may not be the best for the other side
- Lack of diversity
- Indoctrination
- create biases to influence

UNDESIRABLE / UNANTICIPATED

Fig. 23 - Results workshop group 3. F. Galdon



AREA

HIGHLY SENSITIVE AREA - SOCIAL INTERACTIONS

TRAJECTORY

AN ALGORITHM CAPABLE OF PREDICTING ... DOMESTIC VIOLENCE

ASYMMETRY

ASYMMETRY - DATA INFORMATION

THE VA AS A SOCIAL ADVISER

CONSEQUENCES

AREA

HIGHLY SENSITIVE AREA - SOCIAL INTERACTIONS

TRAJECTORY

AN ALGORITHM CAPABLE OF PREDICTING ... YOUR BEST DATE

ASYMMETRY

ASYMMETRY - DATA INFORMATION

ANTICIPATED / DESIRABLE

- Alerting police
- Preventive measure
- Advisor
- Evidence recommendation

EXPECTED BENEFITS

UNANTICIPATED / DESIRABLE

- Evidence
- Crime statistics
- Using data to help others
- Adverts for AIDS
- Human empathy

UNEXPECTED BENEFITS

EXPECTED DRAWBACKS

- Privacy issues
- Poor detection
- Evidence inference limitations
- wasting police time
- Drop in sales

UNDESIRABLE / ANTICIPATED

UNEXPECTED BACKFIRES

- Brings up old negative moments

UNDESIRABLE / UNANTICIPATED

ANTICIPATED / DESIRABLE

- You get your perfect date
- Save time and money
- Archive happiness

EXPECTED BENEFITS

UNANTICIPATED / DESIRABLE

- Less domestic violence
- Companies for dates

UNEXPECTED BENEFITS

EXPECTED DRAWBACKS

- You will never grow as a person
- You can't predict long term relationships
- Robs your choice
- May be Racist
- Privacy issues
- Hackers
- Body issues
- Frequency of machine unbalanced
- May be used by underaged

UNDESIRABLE / ANTICIPATED

UNEXPECTED BACKFIRES

- Stalkers
- May match with family relatives
- Gender prediction may be inaccurate

UNDESIRABLE / UNANTICIPATED

Fig. 24 - Results workshop group 4. F. Galdon

Results

As a result, the anticipated quadrants were better developed, with 61 proposals, whereas the unanticipated aspects of product development presented 54 proposals in total from the participants. From this activity, four main categories of unintended consequences emerged from which to build actions to design systems of accountability (Fig. 25);

- Unhappiness about a service – unethical investments, indoctrination, manipulation, or addiction.
- Inaccurate predictions – accuracy.
- Losing something – dependency, privacy (stalkers, hackers), segregation, isolation, addiction, indoctrination, manipulation, homogeneity, or lack of diversity.
- Violent outcome – death, harm, or injury.

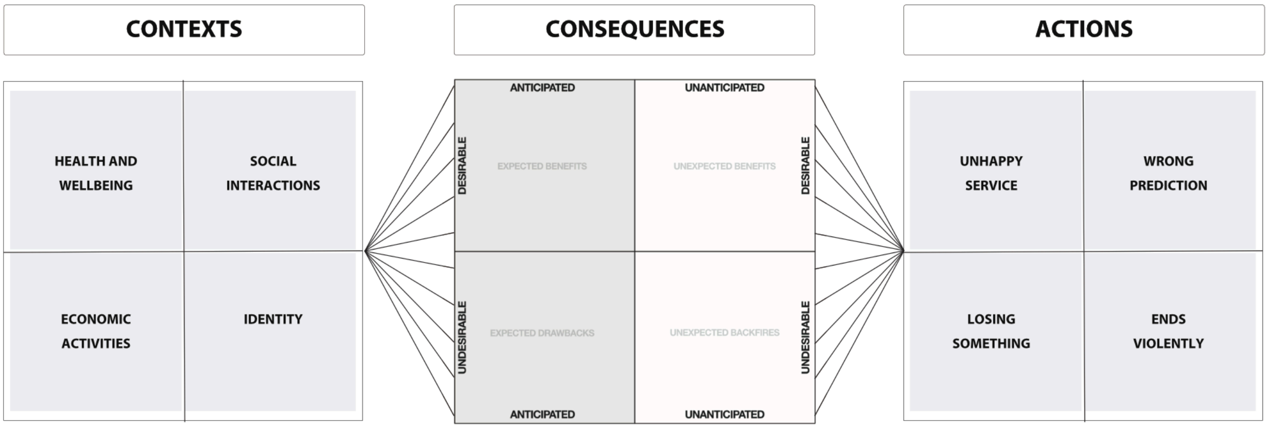


Fig. 25 - The workshop divided participants into four groups. Each of them addressed a particular highly sensitive area (CONTEXT). Each group developed the consequential quadrant proposed by Michaels (2019). From this activity four main outcomes emerged in terms of unintended actions from the system (ACTIONS).

3.3.4.1 Building the mechanism;

Automation

Once the context and outcomes were identified, in order to define **what is the future social and systematic structure/mechanism (systems of accountability) which could establish trust in HASs would be like**, I needed to understand the dynamics of the system. Current models that focus on designing trust in automation structure the process into three main interactive stages: expectations, experimentation, and reliability. (see Appendix B; Ch. 2 for expanded diagraming on automation)

Expectations depend on preliminary knowledge, recommendation by relatives, endorsement by celebrities, and anthropomorphic design attributes such as typological design, voice, 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 whether the system sounds comfortable and natural, define its communication style. Fluidity and automation, the use of recommendations, and low error rates define its ease of use and transparency, and the communication of intent defines its transparency and feedback level. Finally, reliability focuses on design strategies for reducing error rates within an automation system fundamentally based on stages and levels of automation (LoA) built around calibration systems (Fig. 26)

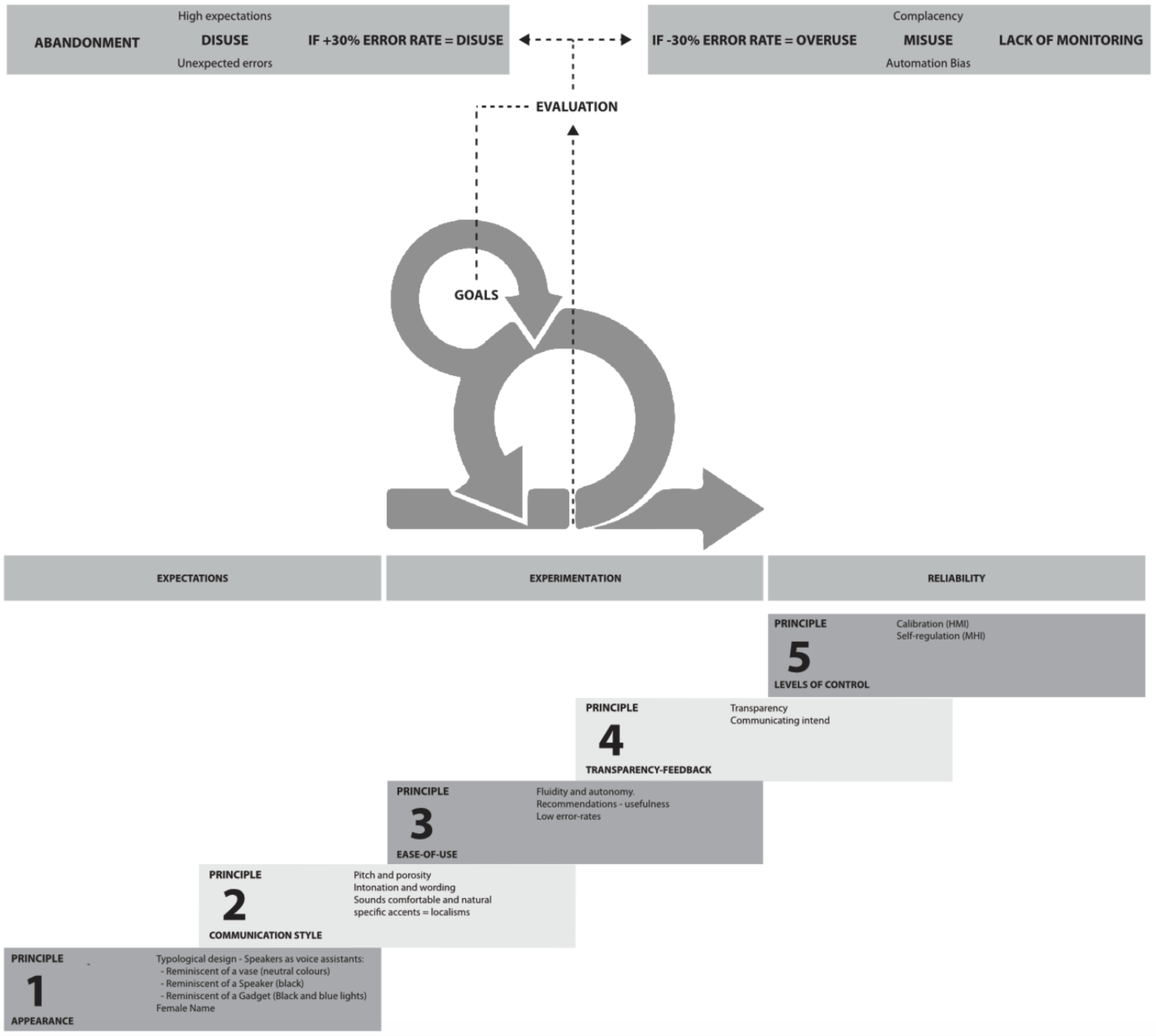


Fig. 26 - Framework of interactive phases on VAs. This includes design principles (presented by Hoff, K. A., & Bashir, M. (2015)), phases of interaction and interactive engagement rates. F. Galdon (2019a)..

As defined in the comparative study (See Fig. 15 in 3.3.1), the design intervention will focus on reliability. It will use levels of control to regulate/calibrate the integrity and intentionality within the system to address the rising concern around inferences. (See Mortier et al., 2014, p. 5-6)

Kaber (2018) points out that levels of automation (LoA) are a fundamental design characteristic that determines the ability of operators to provide adequate oversight and interaction with the automation of the system. Levels aim to improve reliability by simplifying interactions. In this context, reliability refers to the extent to which the actions of the automation are understandable and predictable (Endsley, 2017). Automated systems which clarify their reasoning are more likely to be trusted (Simpson, 1995; Lee, 2004).

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 and avoid unwanted results which may result in situations that cannot be changed (Drnec, 2016). System faults refer to specific system events, rather than the overall performance of the system. In general, system faults have a negative impact on trust in automation. When faults occur, trust levels are affected dramatically. Recovery after these events is much slower, even when the automation generally performs adequately (Moray & Inagaki, 1999; Parasuraman & Manzey, 2010; Parasuraman & Riley, 1997).

In this scenario, 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 (Lewis, 2018). Once reliability has been judged, the most important factor of trust in automation is the predictability of performance over time (Lee, 1992). Predictability is enhanced by implementing LoAs. The idea of gradient-based models of approximation with positive, negative, and neutral spectra has been embodied through the concept of scales, or levels of trust (LoT). Research in the area of human factors presents evidence that the more reliable the system, the more likely it is to be trusted (Parasuraman, 1997; Parasuraman, 2008; Parasuraman, 2010). This positions this area as the most relevant for building and establishing trust in automation.

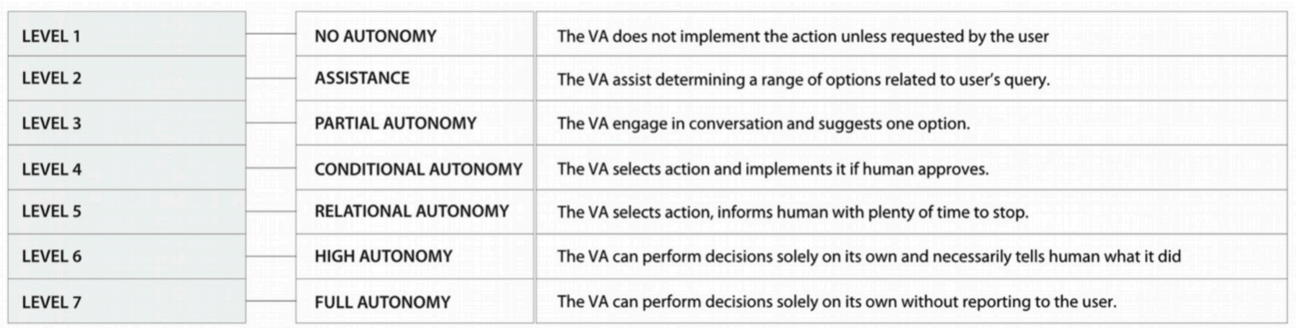
Endsley (2017) argues that the most crucial benefit of the levels approach is its communicative value to key stakeholders (e.g., system operators, designers, and programme managers) about the intrinsic notion that there are different ways and degrees of automation implementation. The fact that there is a whole range of options between fully manual and fully automated levels enhances the understanding of these systems by non-experts. This method has proven successful in providing a solid foundation to understand human-automation interactions (HAIs) at a deeper level. This is highly relevant when confronting an invisible entity making decisions while working in the background.

In this context, Sheridan (1978) introduced LoAs in a seminal work in 1978. This is the most commonly used and reliable model; however, other models exist. Other prominent frameworks in the area of LoA are, for instance, those suggested by Parasuraman, Sheridan, and Wickens (2000). These researchers present a framework that differs radically from earlier approaches. When structuring a scale, they propose a four-level structure outlining four classes or types of automation functions to account for human-machine-interaction. Wickens et al. (2010), in their degrees of automation approach, propose a similar approach to Parasuraman, with a small addition of the notion of degrees (high and low). An approach closer to Sheridan's is presented by Westin, Hilburn & Borst (2013). They present a seven-point scale ranging from total human control to total automation in the context of air traffic management. The multi-variable framework identified by Marinik, Bishop, Fitchett, Morgan, Trimble & Blanco (2014) integrates both approaches: stages and levels of control. This framework is widely used in current vehicle automation research. Finally, the most recent taxonomy is presented by Johnson, Miller, Rusnock, & Jacques (2017). This framework shifts the priority from LoAs to levels of control given a particular situation. It introduces flexibility and contextual awareness. In this context, Kaber (2018) points out that the decision about levels, and their design, must be made by the system's designer. In this scenario, designers should be involved in the derivation of LoA in a collaborative sense due to evidence of empathic misalignment. In this context, "Communities of practice" (Lohmann, 2017, p. 131: Rittel, 1969) — groups of individuals assembled from a range of different publics — will help cancel out blind spots (See Galdon, 2019a: Galdon, 2019b: Galdon, 2019c).

LoA should include adaptive automation, a granularity of control, and automation interface design. Levels are a fundamental design characteristic that determines the ability of operators to provide adequate oversight and interaction with system automation. In this context, levels remain a central design decision associated with the design of automated and autonomous systems that must be addressed in the system's design. The first design question I needed to answer was: how many levels of control should my scale have?

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 commonly used validated scale was developed by Jian, Bisantz, and Drury (2000). This is a seven-point scale articulated to measure global trust in automation. Recent studies using the scale presented excellent internal reliability (Buckley, 2018). Other scales include Mayer and Davis's propensity to trust scale (1999), Lee and Moray's subjective rating scale (Moray, 2000), the Human-Computer Trust questionnaire by Madsen & Gregor (2000), and a cross-cultural trust in automation scale by Chien et al. (2014). Their functionality for measuring trust ranges from particular types of automation, such as autonomous vehicles (Garcia, 2015), to robotics (Yagoda, 2012).

Building on these arguments, this study proposes the articulation of a seven-level odd scale. This type of scale proposes a neutral element and two extremes that allocate extreme perspectives: in this case, no autonomy and full autonomy. It uses Sheridan (1978) model as its foundation to adapt the scale for VAs and the increasing LoA that is expected to evolve in future developments (Table. 4).



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.

Table. 4 - Proposed Levels of Autonomy. F. Galdon (2019a)

Although several models address the nature and practice of automation systems, models in automation that lead to autonomy designed explicitly for VAs and focusing on trust remained unsolved. The model presented is a first step in building a system capable of building and maintaining trust in HASs. However, recent research in the area of robustness in HASs shows 0% adversarial accuracy when evaluating a deep network against stronger adversaries (Athalye, 2018; Uesato, 2018). As the researchers 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” (Kohli, 2019). In this emerging paradigm is the technology that takes the initiative for the interaction (Ortega, 2018). This approach places HASs at the centre and positions trust as the fundamental element to design. In this context, as we cannot fully guarantee the output of the interaction, we need to start talking about accountability and reparation as a posteriori elements to address trust in HASs. In this context, I asked two fundamental questions: due to the impossibility of thoroughly monitoring systems in real-time, due to its ever-increasing complexity, if something goes wrong, **who is accountable?** And, **is there any strategy to repair the trust of the user within the system?** This reflection developed the focus of research towards reparation and accountability and opened a design space for a multi-scalar model to accommodate multiple variables into a single mechanism. This multi-dimensional perspective presented a novelty, as single scales were the dominant paradigm. The following sections will focus on developing specific scales to deal with accountability and reparation.

Accountability

Building from the critical analysis presented above, this section presents a multi-level taxonomy of accountability levels specifically adapted to the future development of virtual assistants in the context of HASs from a Human-Human-Interaction (HHI) perspective to address the proposed sub-questions. (See Appendix B; Ch. 2 for expanded diagramming on automation)

With the rise of HASs, activity in the field of human factors has focused on designing appropriate tools to address this new class of technology (Hancock, 2017). Recent investigations, such as MIT’s research project The Moral Machine (MIT, 2020), on the ethical dilemmas of autonomous vehicle use, point to ethical decision-making in the context of HASs as a central area to address and design (Awad, 2018). Furthermore, due to their persuasive capabilities, concerns are also being raised in the area of VAs 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 medical advisor, diagnosing and providing treatment in the process (Jin, 2018). Further innovations are transforming VAs into legal or financial advisers, dating services and employment agencies. 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 in highly sensitive areas such as health and wellbeing, identity, social interactions and economically related activities. However, one fundamental question remains: if something goes wrong, who should be accountable for the action?

As we move into a Machine-Human-Interaction paradigm (MHI), questions of accountability remain unsolved. This fact makes HASs a focal area, and research must try to address the implications of trust from their perspective (Ortega, 2018). Traditionally, accountability in complex automated VAs has been little researched due to the non-autonomous nature of the interactions. They were based on one-off queries focused on non-dangerous outcomes, such as playing songs or providing the weather forecast. Nowadays, as systems become more automated and unsupervised, the potential outcomes of these interventions are probing capital for the successful development and implementation of these systems in society. In this context, these black boxes represent the essential obscurity within the system (Ashby, 1956) that requires a posteriori reparations to account for trust. Black boxes are widely acknowledged to be a central trust issue for AI.

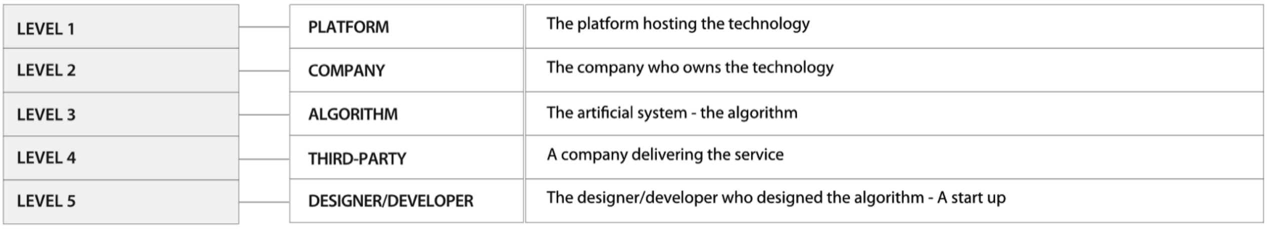
Recent research into the area of robustness in HASs shows 0% adversarial accuracy when evaluating a deep network against stronger adversaries (Athalye, 2018; Uesato, 2018). In order to address this problem, they are using interval bound propagation (Ehlers, 2017; Katz, 2017; Mirman, 2018) with some success. However, as presented earlier, “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” (Kohli, 2019). In this context of the continuous evolution of unsupervised HASs, we must change our approach. We have to start talking about trust and accountability and reparation as 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 context, I have designed a multi-level taxonomy of levels of accountability specifically designed to address the increasing autonomy of highly automated VAs. It integrates a gradient spectrum of levels ranging from the system to the user. Building on Blombaum (2014), it is structured in five distinctive levels: the platform hosting all the interactions (Level 1), the company developing the technology (Level 2), the algorithm (Level 3), a third-party delivering a service (Level 4), and finally, the developer/designer designing the actions/skills algorithm (Level 5). (Tab. 5)

In this scenario a legal debate may emerge regarding the accountability of algorithms, as they are not juridical entities and are designed by developers. The work of Rubel et al. (2019), for instance, points towards “agency laundering”: a moral wrong which consists in distancing oneself from morally suspect actions, regardless of whether those actions were intended or not, by blaming the algorithm (Rubel, Castro, and Pham 2019). However, in this area debates are taking place on whether we should tax robots; there is a case in Korea, as reported by the British Council, in which a robot named Sophia has been granted a passport. As acknowledged in the same article,

> While Saudi Arabia is the first country to grant citizenship to an AI-enabled android, it is not alone in pushing for more rights for robots. In 2017 the European Parliament proposed a set of regulations to govern the use and creation of artificial intelligence, including the granting of 'electronic personhood' to the most advanced machines to ensure their rights and responsibilities (British Council, 2021)

Furthermore, the increasing reality of an algorithm adjusting other algorithms due to interconnected and entangled systems complexity is a blurred area for determining accountability. These legal debates go beyond the scope of this PhD; however, as a prospective designer addressing future interactions, my duty is to infer potential interactions, and the accountability of algorithms seems a pertinent category to be integrated into the system.



LEVEL 1	PLATFORM	The platform hosting the technology
LEVEL 2	COMPANY	The company who owns the technology
LEVEL 3	ALGORITHM	The artificial system - the algorithm
LEVEL 4	THIRD-PARTY	A company delivering the service
LEVEL 5	DESIGNER/DEVELOPER	The designer/developer who designed the algorithm - A start up

Table. 7 - Proposed Levels of Accountability. F. Galdon (2019c)

Reparation

The literature in the area of automation calls for the development of reparation strategies (Bottom, 2002; Kim, 2004; Kohn, 2018). These strategies are becoming crucial 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, or the user will stop using them (Parasuraman, 2000; Wickens, 2007; Wang, 2009). As these systems become more autonomous, ubiquitous, and unsupervised, the development of reparation techniques becomes fundamental for the adequate development and integration of these systems in society. (See Appendix B; Ch. 2 for expanded diagramming on automation)

Traditionally, research on reparation focuses on different typologies such as apologies, or denials, and the timing of delivering them. Bansal and Zahedi (2015) investigated how trust may be rebuilt after it is violated by adverse events in data privacy, including the efficacy of the three most frequent response types – an apology, a 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 report that is critical to investigate the typology of violation events.

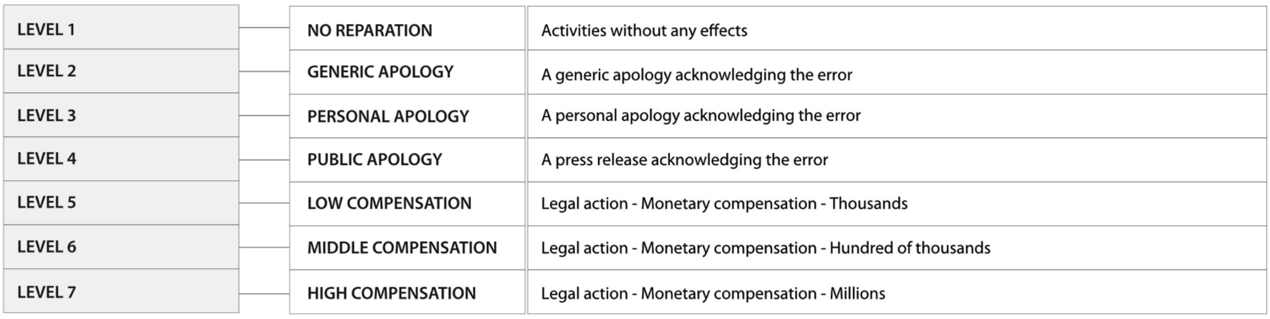
Their research was ground-breaking. However, it was based on current models of VAs, such as Alexa, which are equipped with the capacity of responding to one-off queries. However, with the emergence of HASs such as Duplex (capable of taking the initiative in interaction and of establishing and maintaining 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 study, I am mindful of 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.

A testament to this approach may be the BSI (UKRI) *Responsible Innovation Guide PAS 440:2020*, or the new *Liability for artificial intelligence and other emerging digital technologies* published by the EU in November 2019. According to the latest version;

> Their rollout must come with sufficient safeguards, to minimise the risk of harm these technologies may cause, such as bodily injury or other harm. In the EU, product safety regulations ensure this is the case. However, such regulations cannot completely exclude the possibility of damage resulting from the operation of these technologies. If this happens, victims will seek compensation. (EU, 2019b, p.3).

In this scenario, I have structured a scale and integrated a new gradation of compensation levels to complement the apology spectrum in order to test whether they are needed to account for the type of interactions emerging from highly automated VA patents, prototypes, and demos. In this context, and following the levels of autonomy strategy implemented in the previous section, I designed a multi-level taxonomy of levels of reparation specifically designed to address the increasing autonomy of highly automated VAs. It integrates a gradient spectrum ranging from no reparation to high compensation (Table 6). It is structured in seven distinctive levels, organised into 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)



LEVEL 1	NO REPARATION	Activities without any effects
LEVEL 2	GENERIC APOLOGY	A generic apology acknowledging the error
LEVEL 3	PERSONAL APOLOGY	A personal apology acknowledging the error
LEVEL 4	PUBLIC APOLOGY	A press release acknowledging the error
LEVEL 5	LOW COMPENSATION	Legal action - Monetary compensation - Thousands
LEVEL 6	MIDDLE COMPENSATION	Legal action - Monetary compensation - Hundred of thousands
LEVEL 7	HIGH COMPENSATION	Legal action - Monetary compensation - Millions

Table. 6 - Proposed levels of Reparation. F. Galdon (2019b)

3.3.4.2 Discussion

I now return to the second question, **What would a future social and systematic structure/mechanism which could establish trust in highly automated systems be like?** Based on the research conducted, I present autonomy, accountability, and reparation levels as fundamental variables to integrate into this mechanism. In this process, I recommend that design must combine a holistic and contextual perspective on trust in order to be able to integrate the impact of contexts on interactions. Trust formation is a dynamic process that starts before a user's first contact with the system and continues long thereafter. Thus, design interventions need to be able to adapt.

In this context, levels are a simplification of reality. However, they facilitate the understanding of HASs. This research provides a foundational LoA based on a generic scale with seven points to address a multitude of cases with varying contexts (Galdon, 2019a). However, as we transition towards Highly Autonomous Systems (HAuSs), this investigation acknowledges the moral imperative of design to address unintended consequences. In this context, I present a scale addressing issues of accountability in HAuSs (Galdon, 2019c), and a scale of levels of reparation (Galdon, 2019b). These variables will form a foundation for building a future structure/mechanism to build trust in HAuSs.

These scales are the first to be specifically designed to address the rising concern about HASs in VAs. No other publications exploring these possibilities were found at the time of this research.

In this process, Consequences emerged as an important method because it provided a systematic model to address asymmetries. It identified the three fundamental levels at which they were relevant; consequences (intended and unintended), contexts (highly sensitive areas: health, economy, identity, and social interactions), and unintended actions (unhappy actions, inaccurate predictions, the loss of something, and violence). This triangulation has the flexibility to address a multiplicity of contexts, cultures, and behaviours. In this process, workshops emerged as a reliable process through which to address positive and negative consequential potentialities.

These contributions have been made possible by implementing a prospective approach, which enables the designer to go beyond what already exists. The fundamental prospective approach of design, based on planning, solution-based problem solving, problem shaping, synthesis, preparedness, readiness, and appropriateness, provides a suitable framework to access these future spaces for knowledge. The possibility of inferring potential and preliminary knowledge makes workshops an ideal method to address the potential impact of prospective technological development. The outputs aim to provide guiding knowledge. In the process, they contribute to a contextualising Glanville's concept of "knowledge for transforming the future" as a probabilistic knowledge ontology (Glanville, 2005).

3.3.5 Counter-fictions

This section addresses counter-fictions to build a system/mechanism integrating all the variables addressed in this section – autonomy, accountability, reparation, contexts, and actions – to create trust in highly automated VAs via an applied case study.

Counter-fiction is an emerging experimental area in design practice. So far, only two publications were found during this research that explores its possibilities: a monographic journal issue (Multitudes, 2012), and a book (Belliot, 2018). This approach aims to address the relations of domination. Its primary approach, rather than being imposed or forced, is based on the co-production of emancipatory projects aimed to reduce repression and enhance individual freedom and responsibility by reversing power.

In this paradigm, trust is the main element to account for. This is understood as a mode of relationships between individuals. In this relational perspective, power is a dynamic and reciprocal force, addressed through asymmetric relations in which the one who is controlled sees their actions, cognition and possible effects reduced, although not determined, by the controller. Trust can be seen as a relation between, or as an influence on, the user and the system, and differs from the point of view of the spectrum of possibilities actually controlled by individuals. This approach positions trust as a fundamental variable to build and maintain the relationship.

In this context, the use of counter-fictional strategies emerged for the author as a method by which to address the asymmetric dynamics of the system, but also as an experimental approach to ground prospective interventions. These interventions can be placed a priori (before the interaction), meanwhile (during the interaction) or a posteriori (after the interaction). Its primary function is to reverse asymmetries through design. Outputs are expected to range from tools to frameworks.

This section presents an applied case study for an integrative multi-dimensional scalar system integrating the proposed taxonomies of levels of autonomy, accountability and reparation specifically adapted to VAs. In this context, the intersection between the critical issues of automation, and accountability acts as a focal point. It will do so by implementing an applied case in the context of energy management and consumption. An accompanying appendix documents the diagrammatic process. (See Appendix B).

The success of this implementation resides in generating a system/mechanism with the capability to reverse the asymmetries presented earlier during its interactive lifecycle. Its primary function is to reverse asymmetries of power and dominance through design. Interactive asymmetries are exploited via inferences (the knowledge and actions performed by the system based on the data gathered), which leads to tensions between integrity and intentionality. This system/mechanism must integrate pre-interaction, during-the-interaction, and post-interaction elements to deal with unintended actions and contextual variability to facilitate the design of trust in HAuS VAs from a consequential perspective.

3.3.5.1 Applied case

With around 120 million smart speakers circulating in the United States alone (NPR, 2019), VAs are expected to dominate interactions in the near future. They will play a fundamental role in energy management and consumption at home via specific applications such as Google's Nest or Amazon's Alexa Home. In this scenario, VAs are transitioning from automation to autonomy. A recent demo of a product called Duplex, presented by Google, offered an extraordinary level of fluency and autonomy never seen before. Therefore, research must focus attention on a new class of technology: HASs (Hancock, 2017). In this emerging paradigm is the technology that dictates who takes the initiative for the interaction (Ortega, 2018). This approach places HASs at the centre of the research, and positions trust as the fundamental element in design.

In this paradigm, the system will have the information and initiative to regulate human behaviour to optimise the impact of energy management and consumption. In this context, trust will be crucial for the adoption of new strategies in energy management and consumption by the user. However, as this persuasive approach will be fundamentally unsupervised, it may generate unintended consequences. Traditionally, complex automated systems required a human operator to appropriately calibrate their trust in the automation in order to achieve performance and safety goals. In this context, the literature has focused traditionally on the Human-Machine-Interaction (HMI) paradigm. However, recent evolutions in the nature and capabilities of automated systems are transitioning to a Human-Human-Interaction (HHI) paradigm to precisely define and calibrate trust in automation. In this case, the author proposes a relational approach in the context of HHI that directly aims to ensure that emerging HAS interactions remain focused on the user's needs and preferences. Its primary function is to reverse potential asymmetries of power and dominance through design.

- Model development

As presented in previous sections, the inquiry started by building a foundational scale of levels of autonomy. This technique has been widely used in the human factors field over the last 40 years. However, as a consequence of the impossibility of monitoring complex dynamic systems, due to their complexity, two fundamental questions emerge; if something goes wrong, who is responsible? And, is it possible to repair the trust of the user in the system? These questions led to the articulation of two complementary consequential scales: a five-level scale on accountability and a seven-level scale on reparation. As part of this inquiry, context and actions emerged as fundamental variables to incorporate in the framework, as they determined the right combination of levels. Then, building on the asymmetric analysis, two more variables were integrated: access (data points) and inferences (predicting capabilities), as they play a fundamental role in enabling asymmetries (Wachter, 2018)(Mortier et al., 2014, p. 5-6). (Fig. 28)

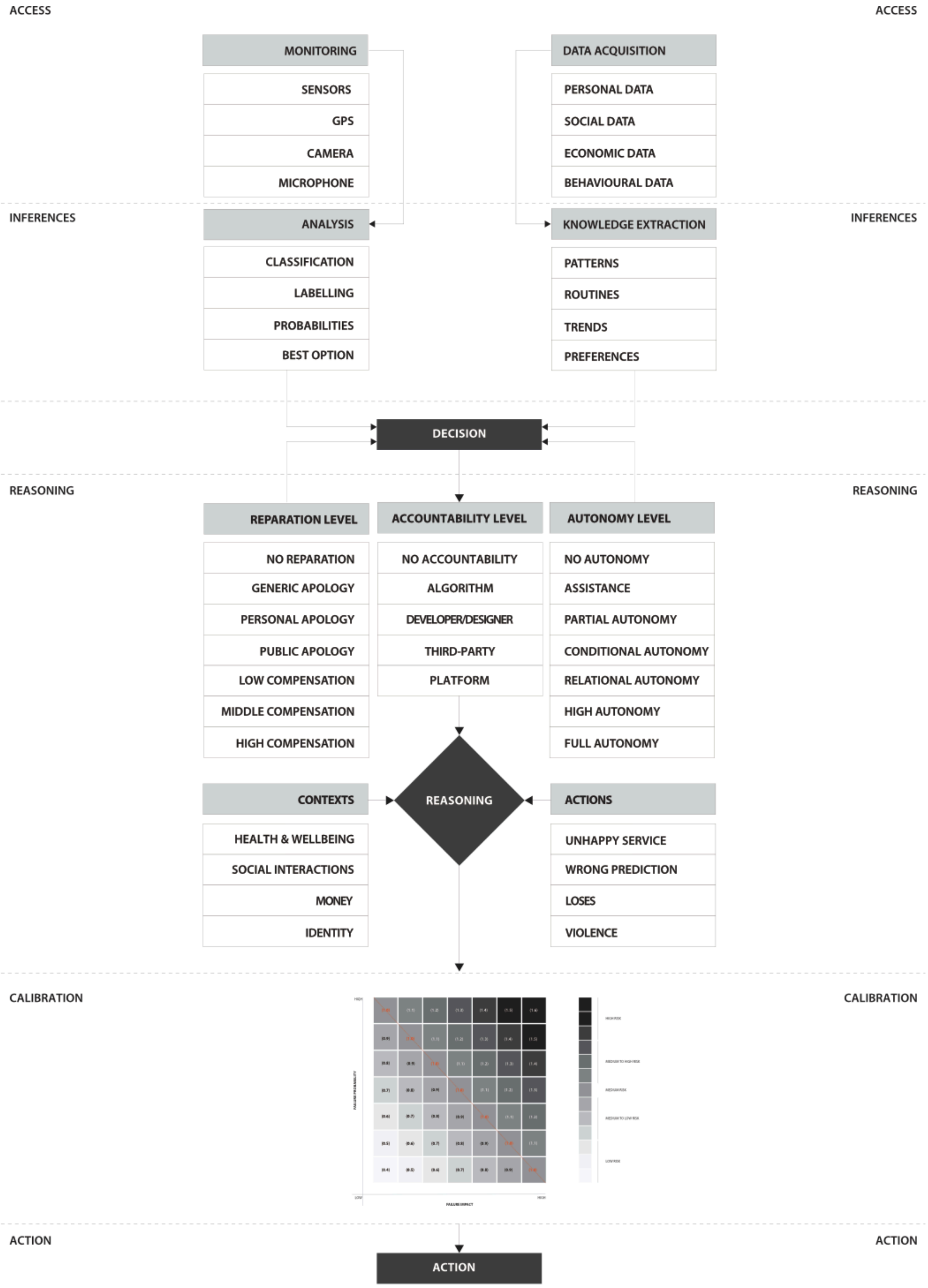


Fig. 28 - VA consequential reasoning system design. This system is structured in five distinctive levels. Access contains data acquisition points. Inferences contain knowledge extraction and analysis actions normally performed by ML systems. Reasoning contains five variables affecting the preliminary decision at different stages. Autonomy and reparation try to balance the initial decision, whereas accountability, contexts and actions operate as factors to affect the secondary decision. A calibration matrix allows us to observe the potential impact of the interaction on a trust scale before the system perfume the action.

This calibration system was structured in four levels;

- The area of access integrates a range of data points from two main perspectives; past profiling data (personal, social, biometric and environmental), and in vivo monitoring data (sensors, GPS, cameras and microphones).
- The area related to inferences integrates the variables of inferences; knowledge extraction capabilities (patterns, routines, trends, preferences), and analysis (classification, labelling, probabilities and best option).
- The area of synthetic consequential reasoning integrates the scales of autonomy, reparation and accountability, as well as contexts (health and wellbeing, social interactions, emancipation and identity) and unintended consequences (unsatisfactory services, inaccurate predictions, losses and unexpected violent endings).
- The integration of all these elements into a multi-dimensional framework led to the design of a calibration system.

From this point, I decided to develop a calculator[1] (follow link below for access) to work out how to translate highly conceptual and philosophical concepts into mathematical forms that the machine would understand. Finally, a calibration matrix was designed to map the intent of the system. It was structured and organised in five levels: low risk, medium to low risk, medium risk, medium to high risk and high risk. This system enables us to obtain a trust rating, and to map/infer the potential impact of an action/skill in context. This element was also integrated into the calculator, and a simulation tool emerged.

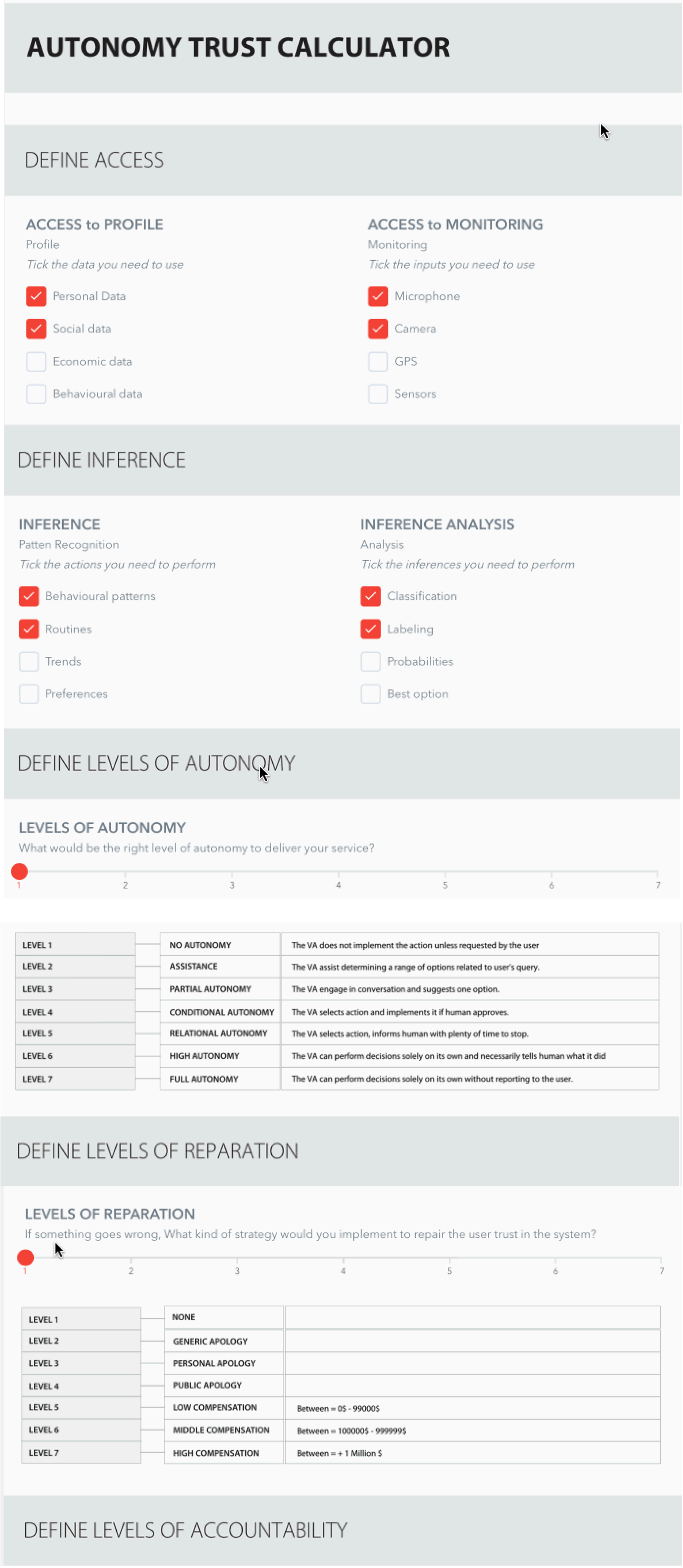
The weighting system in the calibration system was developed based on a straightforward premise: the higher the impact of an interaction in terms of reparation, the lower the autonomy a system can have. It operates like a weighing scale.

Accountability then adds a factor to this output. The access and inferences variables add a decimal to the rating resulting from this calculation. The more elements you use, the more you will be punished. (Fig. 29-30)

The idea here is to force companies/developers to optimise the data and inferences applied, thus optimising the impact of potential asymmetries. Finally, contexts and unintended actions insert a variable value. Each of the four possible options contains a different value, ranging from 0.1 to 0.25. Their values depend on the context at hand. This contextual operational value was obtained via workshops. These workshops provided preliminary knowledge about which contexts and actions are most sensitive to a particular area of interaction.

The possibility of inferring potential and preliminary knowledge makes co-design workshops an ideal method to address the potential impact of a prospective technological development. The outputs aim to provide guiding knowledge. In the process, they are contributing to a contextualization of Glanville's concept of "knowledge for transforming the future"as a probabilistic knowledge ontology (Glanville, 2005).

[1] https://fgedesign.wixsite.com/calibration



AUTONOMY TRUST CALCULATOR

DEFINE ACCESS

ACCESS to PROFILE
Profile
Tick the data you need to use

☑ Personal Data
☑ Social data
□ Economic data
□ Behavioural data

ACCESS to MONITORING
Monitoring
Tick the inputs you need to use

☑ Microphone
☑ Camera
□ GPS
□ Sensors

DEFINE INFERENCE

INFERENCE
Patten Recognition
Tick the actions you need to perform

☑ Behavioural patterns
☑ Routines
□ Trends
□ Preferences

INFERENCE ANALYSIS
Analysis
Tick the inferences you need to perform

☑ Classification
☑ Labeling
□ Probabilities
□ Best option

DEFINE LEVELS OF AUTONOMY

LEVELS OF AUTONOMY
What would be the right level of autonomy to deliver your service?

● 1 — 2 — 3 — 4 — 5 — 6 — 7

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.

DEFINE LEVELS OF REPARATION

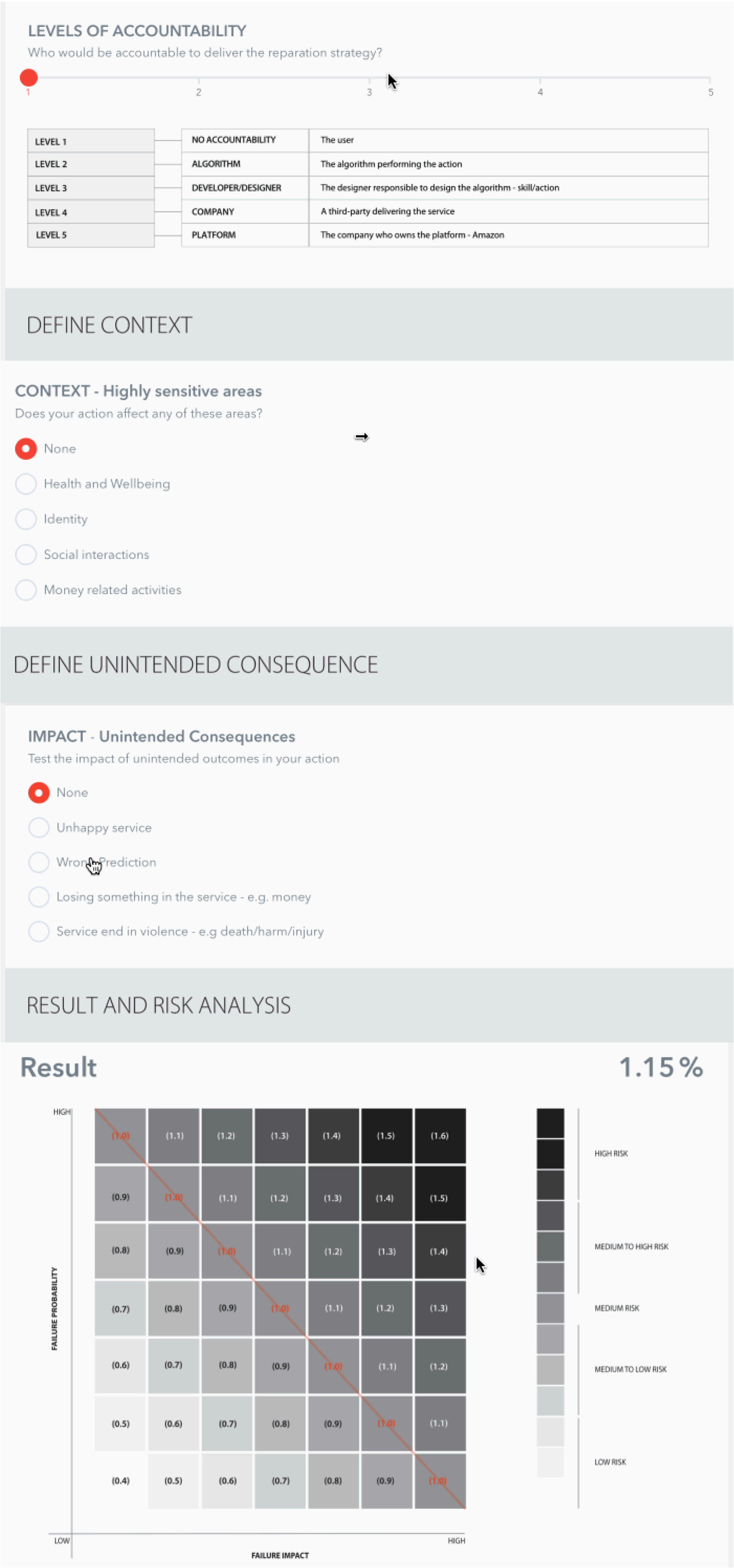
LEVELS OF REPARATION
If something goes wrong, What kind of strategy would you implement to repair the user trust in the system?

● 1 — 2 — 3 — 4 — 5 — 6 — 7

LEVEL 1	NONE	
LEVEL 2	GENERIC APOLOGY	
LEVEL 3	PERSONAL APOLOGY	
LEVEL 4	PUBLIC APOLOGY	
LEVEL 5	LOW COMPENSATION	Between = 0$ - 99000$
LEVEL 6	MIDDLE COMPENSATION	Between = 100000$ - 999999$
LEVEL 7	HIGH COMPENSATION	Between = + 1 Million $

DEFINE LEVELS OF ACCOUNTABILITY

Fig. 29 - Calculator design - continues below. F. Galdon



LEVELS OF ACCOUNTABILITY

Who would be accountable to deliver the reparation strategy?

LEVEL 1	NO ACCOUNTABILITY	The user
LEVEL 2	ALGORITHM	The algorithm performing the action
LEVEL 3	DEVELOPER/DESIGNER	The designer responsible to design the algorithm - skill/action
LEVEL 4	COMPANY	A third-party delivering the service
LEVEL 5	PLATFORM	The company who owns the platform - Amazon

DEFINE CONTEXT

CONTEXT - Highly sensitive areas

Does your action affect any of these areas?

- ● None
- ○ Health and Wellbeing
- ○ Identity
- ○ Social interactions
- ○ Money related activities

DEFINE UNINTENDED CONSEQUENCE

IMPACT - Unintended Consequences

Test the impact of unintended outcomes in your action

- ● None
- ○ Unhappy service
- ○ Wrong Prediction
- ○ Losing something in the service - e.g. money
- ○ Service end in violence - e.g death/harm/injury

RESULT AND RISK ANALYSIS

Result 1.15%

(1.0)	(1.1)	(1.2)	(1.3)	(1.4)	(1.5)	(1.6)
(0.9)	(1.0)	(1.1)	(1.2)	(1.3)	(1.4)	(1.5)
(0.8)	(0.9)	(1.0)	(1.1)	(1.2)	(1.3)	(1.4)
(0.7)	(0.8)	(0.9)	(1.0)	(1.1)	(1.2)	(1.3)
(0.6)	(0.7)	(0.8)	(0.9)	(1.0)	(1.1)	(1.2)
(0.5)	(0.6)	(0.7)	(0.8)	(0.9)	(1.0)	(1.1)
(0.4)	(0.5)	(0.6)	(0.7)	(0.8)	(0.9)	(1.0)

Fig. 30 - Calculator design. F. Galdon

- Calibration adaptation to energy management and consumption

In this context, a workshop was implemented to understand how contexts and their unintended derivative consequences affect trust in highly automated VAs in the area of energy management and consumption.

Particular attention was given to the relevance of contextual events and dynamism in enhancing trust by understanding that trust formation is a dynamic process that starts before the user's first contact with the system and continues long thereafter.

Furthermore, following the evolving nature of the system, factors affecting trust and the system itself change during user interactions over time; thus, systems need to be able to adapt and evolve.

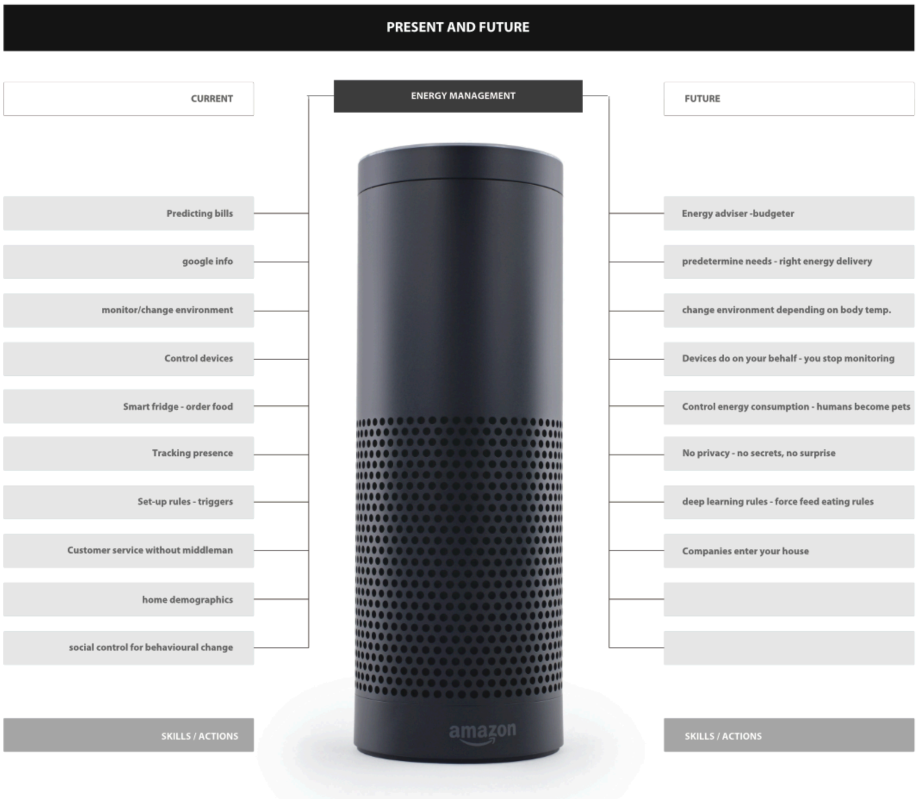
A workshop with ten postgraduate students from the School of Design at the RCA was devised to investigate energy management and consumption in the context of VAs to assess the weighing system.

The rationale of working with designers builds on Jasanoff's account of the ethics of invention. She positions 'benevolent' technologists, designers and businessmen' as the most qualified individuals to develop trusted systems. Consequently, the deontological position of myself and the participants, who are fundamentally designers/technologists/entrepreneurs, has been crucial for the integration of applied ethics and emancipatory directionalities in collective activities.

In this context, the diversity of the student body at the RCA in terms of background, programme of study, culture and nationality, and their diverse, critical and enabling capabilities, plus the unifying element of being designers, provided an ideal group of participants to develop the task at hand. (See Appendix 4 for detailed workshop material).

Participants were divided into two groups of five.

- First, participants analysed the current development of VAs in the context of energy management and consumption and mapped current skills/actions. Then they projected them into the future, using "what if...?" questions to understand the impact of the evolution of these systems (Fig. 31-32).
- After this task, participants were asked to conduct a consequential analysis, mapping desired and undesired consequences (Fig. 33-34).
- After this analysis, they mapped the prospective outcomes in terms of impact on contexts and impact on actions (Fig. 35-36).
- Finally, they were presented with two quadrants to map the outcomes in highly sensitive areas in terms of contexts and actions (Fig. 37-38). These analyses allowed the author to weight the system (Fig. 39).



PRESENT AND FUTURE

CURRENT

ENERGY MANAGEMENT

FUTURE

Predicting bills

google info

monitor/change environment

Control devices

Smart fridge - order food

Tracking presence

Set-up rules - triggers

Customer service without middleman

home demographics

social control for behavioural change

SKILLS / ACTIONS

Energy adviser -budgeter

predetermine needs - right energy delivery

change environment depending on body temp.

Devices do on your behalf - you stop monitoring

Control energy consumption - humans become pets

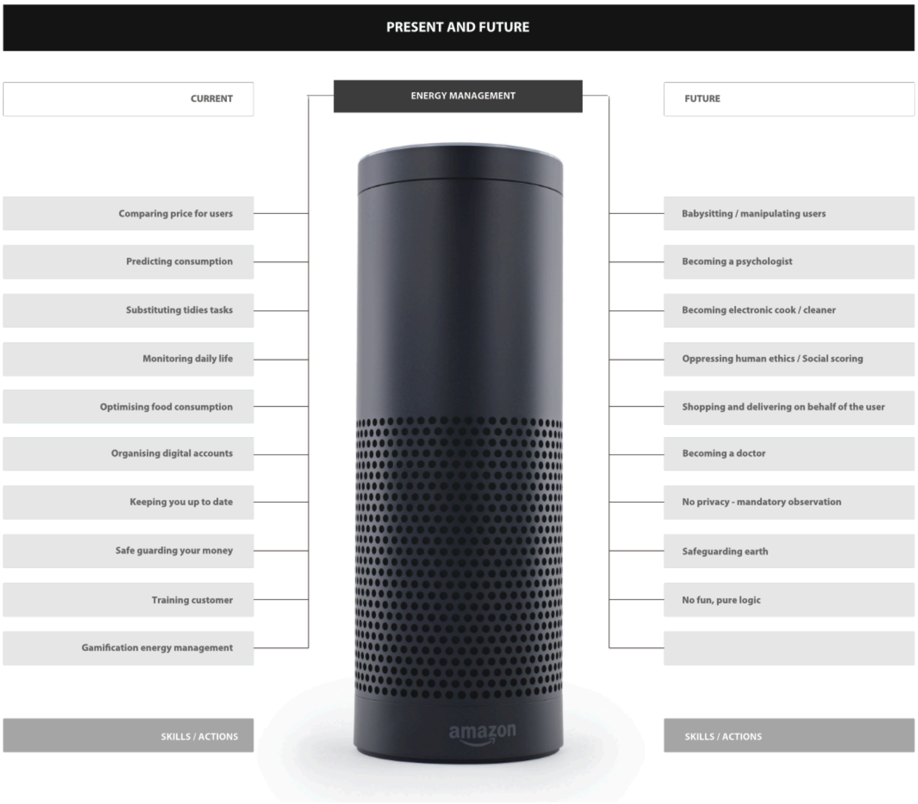
No privacy - no secrets, no surprise

deep learning rules - force feed eating rules

Companies enter your house

SKILLS / ACTIONS

Fig. 31 - Results mapping task - Group1.



PRESENT AND FUTURE

CURRENT

ENERGY MANAGEMENT

FUTURE

Comparing price for users

Predicting consumption

Substituting tidies tasks

Monitoring daily life

Optimising food consumption

Organising digital accounts

Keeping you up to date

Safe guarding your money

Training customer

Gamification energy management

SKILLS / ACTIONS

Babysitting / manipulating users

Becoming a psychologist

Becoming electronic cook / cleaner

Oppressing human ethics / Social scoring

Shopping and delivering on behalf of the user

Becoming a doctor

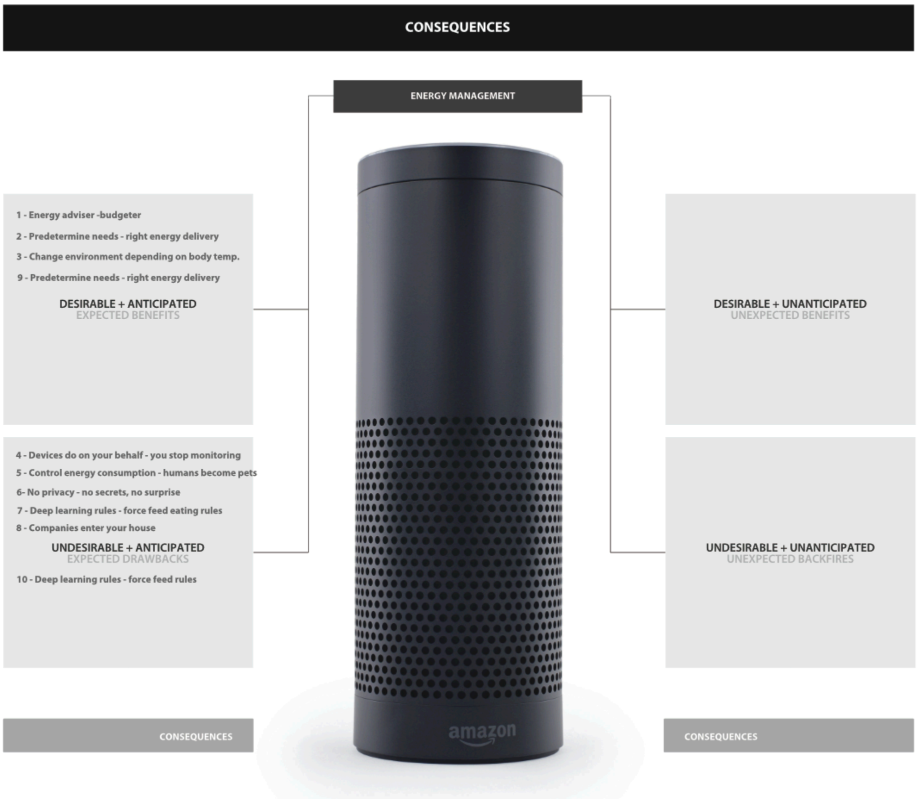
No privacy - mandatory observation

Safeguarding earth

No fun, pure logic

SKILLS / ACTIONS

Fig. 32 - Results mapping task - Group2.



CONSEQUENCES

ENERGY MANAGEMENT

1 - Energy adviser -budgeter

2 - Predetermine needs - right energy delivery

3 - Change environment depending on body temp.

9 - Predetermine needs - right energy delivery

DESIRABLE + ANTICIPATED
EXPECTED BENEFITS

4 - Devices do on your behalf - you stop monitoring

5 - Control energy consumption - humans become pets

6- No privacy - no secrets, no surprise

7 - Deep learning rules - force feed eating rules

8 - Companies enter your house

UNDESIRABLE + ANTICIPATED
EXPECTED DRAWBACKS

10 - Deep learning rules - force feed rules

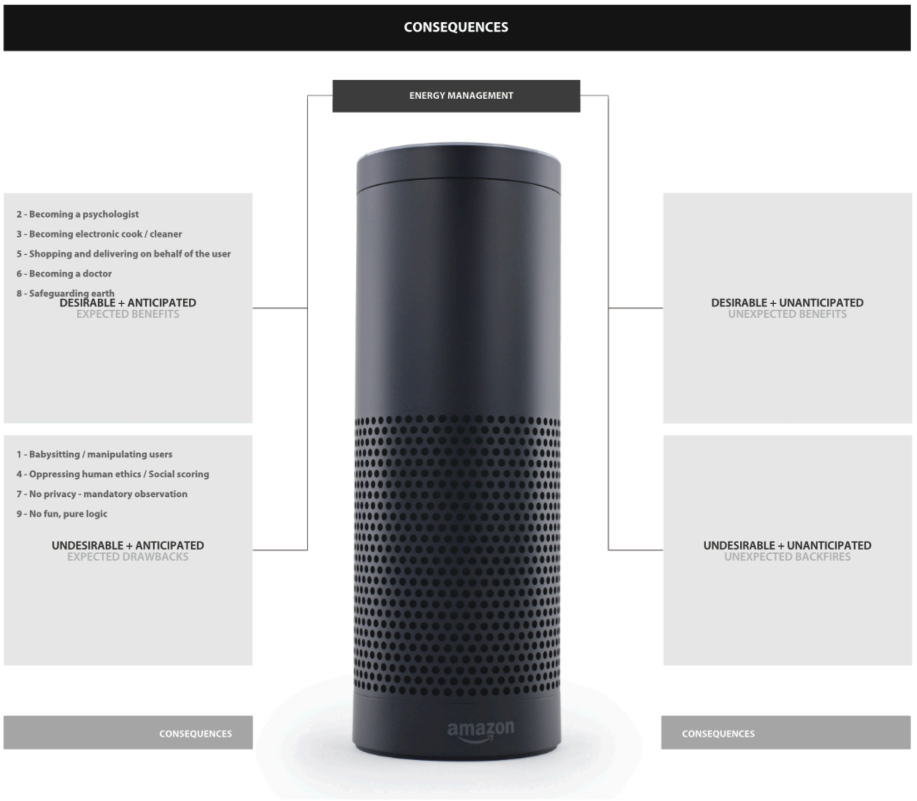
DESIRABLE + UNANTICIPATED
UNEXPECTED BENEFITS

UNDESIRABLE + UNANTICIPATED
UNEXPECTED BACKFIRES

CONSEQUENCES

CONSEQUENCES

Fig. 33 - Results consequential analysis task - Group1.



CONSEQUENCES

ENERGY MANAGEMENT

2 - Becoming a psychologist

3 - Becoming electronic cook / cleaner

5 - Shopping and delivering on behalf of the user

6 - Becoming a doctor

8 - Safeguarding earth

DESIRABLE + ANTICIPATED
EXPECTED BENEFITS

1 - Babysitting / manipulating users

4 - Oppressing human ethics / Social scoring

7 - No privacy - mandatory observation

9 - No fun, pure logic

UNDESIRABLE + ANTICIPATED
EXPECTED DRAWBACKS

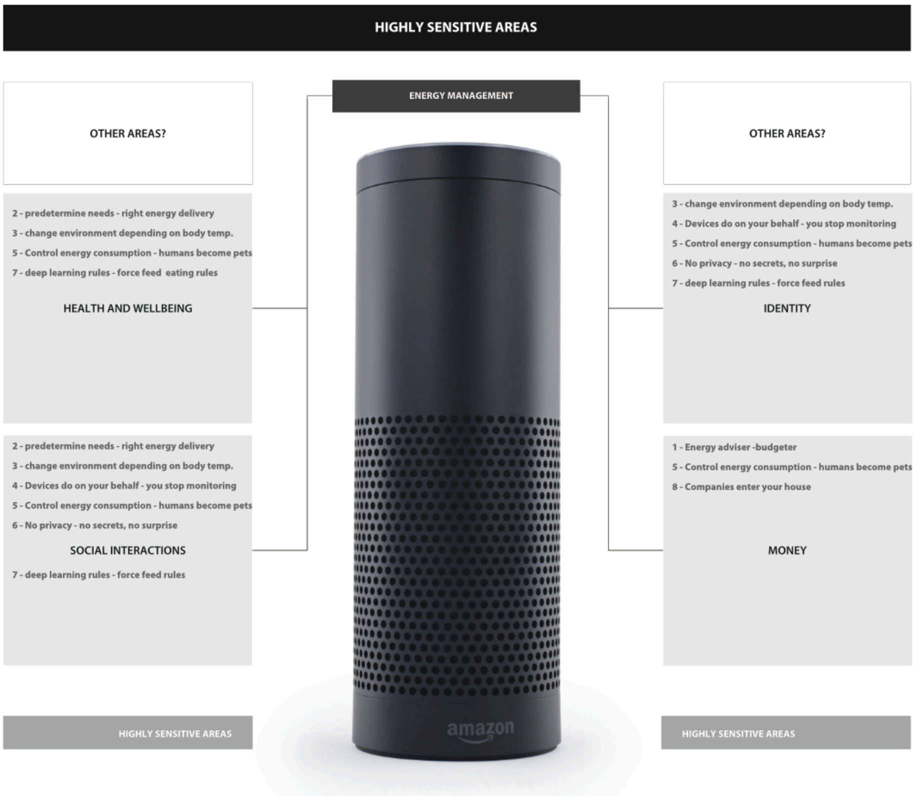
DESIRABLE + UNANTICIPATED
UNEXPECTED BENEFITS

UNDESIRABLE + UNANTICIPATED
UNEXPECTED BACKFIRES

CONSEQUENCES

CONSEQUENCES

Fig. 34 - Results consequential analysis task - Group2.



HIGHLY SENSITIVE AREAS

ENERGY MANAGEMENT

OTHER AREAS?

2 - predetermine needs - right energy delivery
3 - change environment depending on body temp.
5 - Control energy consumption - humans become pets
7 - deep learning rules - force feed eating rules

HEALTH AND WELLBEING

2 - predetermine needs - right energy delivery
3 - change environment depending on body temp.
4 - Devices do on your behalf - you stop monitoring
5 - Control energy consumption - humans become pets
6 - No privacy - no secrets, no surprise

SOCIAL INTERACTIONS

7 - deep learning rules - force feed rules

HIGHLY SENSITIVE AREAS

OTHER AREAS?

3 - change environment depending on body temp.
4 - Devices do on your behalf - you stop monitoring
5 - Control energy consumption - humans become pets
6 - No privacy - no secrets, no surprise
7 - deep learning rules - force feed rules

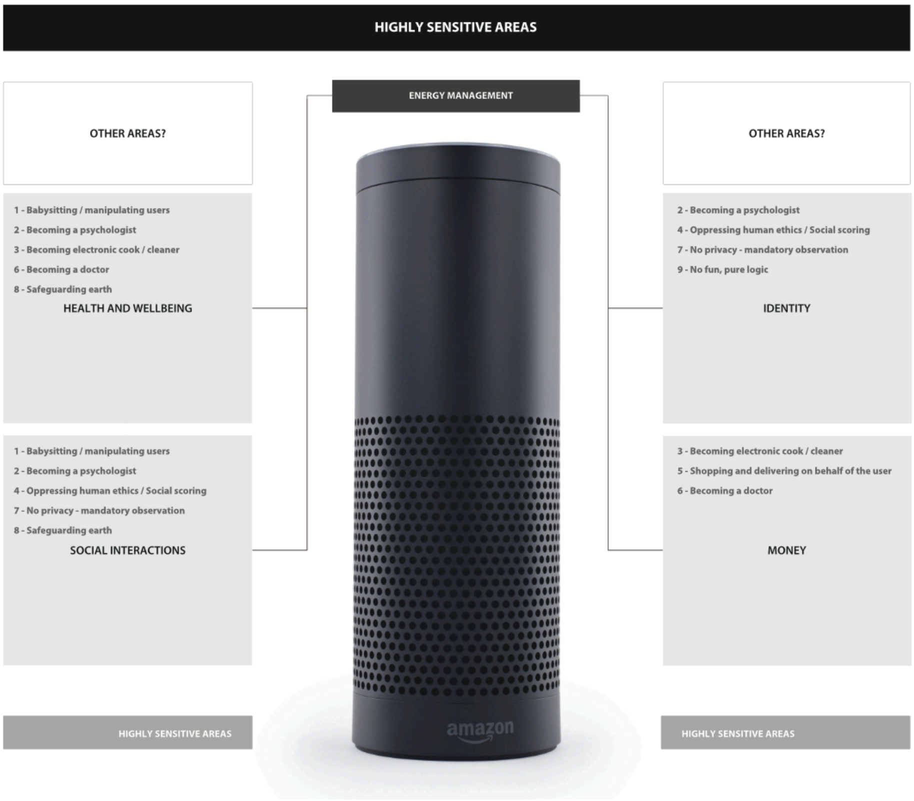
IDENTITY

1 - Energy adviser -budgeter
5 - Control energy consumption - humans become pets
8 - Companies enter your house

MONEY

HIGHLY SENSITIVE AREAS

Fig. 35 - Highly sensitive areas impact analysis task - Group1.



HIGHLY SENSITIVE AREAS

ENERGY MANAGEMENT

OTHER AREAS?

1 - Babysitting / manipulating users
2 - Becoming a psychologist
3 - Becoming electronic cook / cleaner
6 - Becoming a doctor
8 - Safeguarding earth

HEALTH AND WELLBEING

1 - Babysitting / manipulating users
2 - Becoming a psychologist
4 - Oppressing human ethics / Social scoring
7 - No privacy - mandatory observation
8 - Safeguarding earth

SOCIAL INTERACTIONS

HIGHLY SENSITIVE AREAS

OTHER AREAS?

2 - Becoming a psychologist
4 - Oppressing human ethics / Social scoring
7 - No privacy - mandatory observation
9 - No fun, pure logic

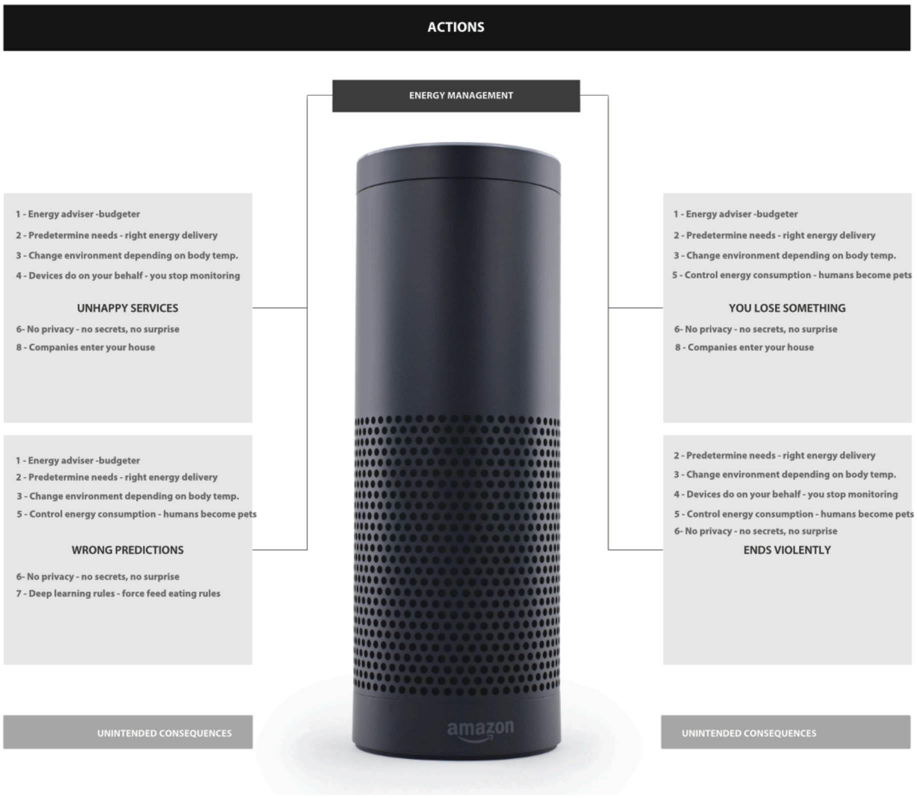
IDENTITY

3 - Becoming electronic cook / cleaner
5 - Shopping and delivering on behalf of the user
6 - Becoming a doctor

MONEY

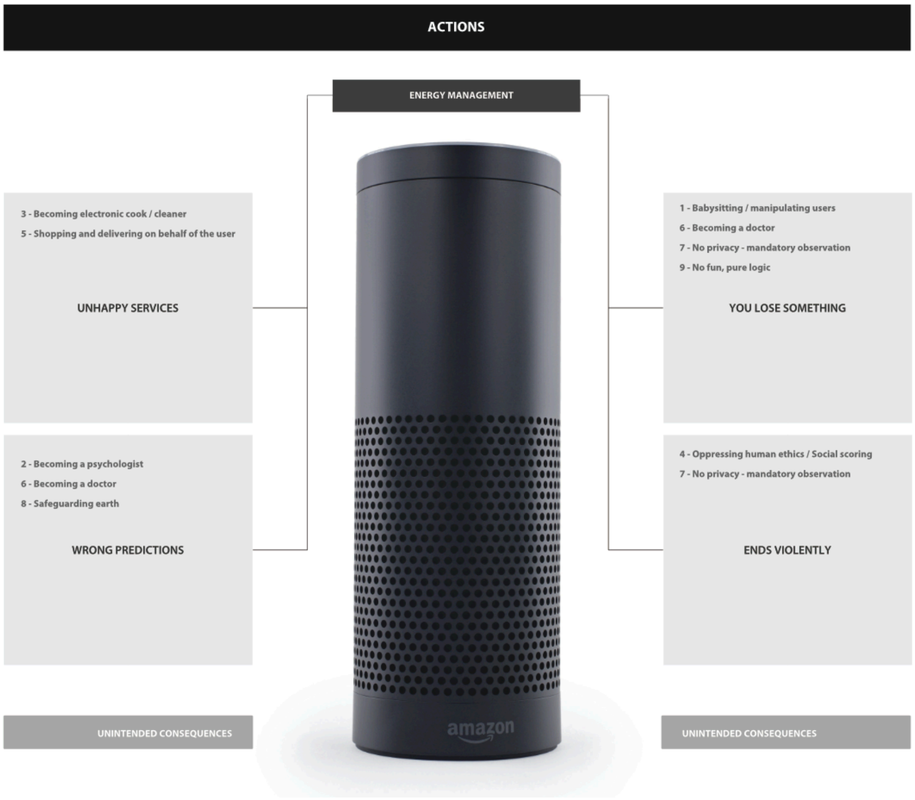
HIGHLY SENSITIVE AREAS

Fig. 36 - Highly sensitive areas impact analysis task - Group2.



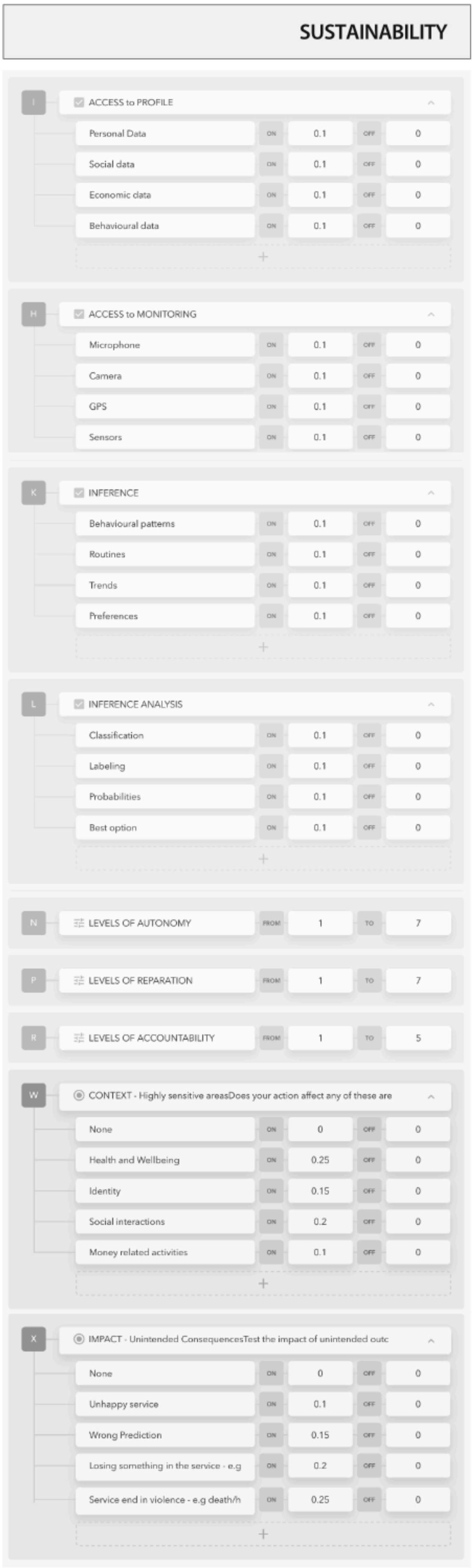
ACTIONS
ENERGY MANAGEMENT
1 - Energy adviser -budgeter
2 - Predetermine needs - right energy delivery
3 - Change environment depending on body temp.
4 - Devices do on your behalf - you stop monitoring
UNHAPPY SERVICES
6- No privacy - no secrets, no surprise
8 - Companies enter your house
1 - Energy adviser -budgeter
2 - Predetermine needs - right energy delivery
3 - Change environment depending on body temp.
5 - Control energy consumption - humans become pets
WRONG PREDICTIONS
6- No privacy - no secrets, no surprise
7 - Deep learning rules - force feed eating rules
1 - Energy adviser -budgeter
2 - Predetermine needs - right energy delivery
3 - Change environment depending on body temp.
5 - Control energy consumption - humans become pets
YOU LOSE SOMETHING
6- No privacy - no secrets, no surprise
8 - Companies enter your house
2 - Predetermine needs - right energy delivery
3 - Change environment depending on body temp.
4 - Devices do on your behalf - you stop monitoring
5 - Control energy consumption - humans become pets
6- No privacy - no secrets, no surprise
ENDS VIOLENTLY
UNINTENDED CONSEQUENCES
amazon
UNINTENDED CONSEQUENCES

Fig. 37 - Unintended consequences impact analysis task - Group1.



ACTIONS
ENERGY MANAGEMENT
3 - Becoming electronic cook / cleaner
5 - Shopping and delivering on behalf of the user
UNHAPPY SERVICES
2 - Becoming a psychologist
6 - Becoming a doctor
8 - Safeguarding earth
WRONG PREDICTIONS
1 - Babysitting / manipulating users
6 - Becoming a doctor
7 - No privacy - mandatory observation
9 - No fun, pure logic
YOU LOSE SOMETHING
4 - Oppressing human ethics / Social scoring
7 - No privacy - mandatory observation
ENDS VIOLENTLY
UNINTENDED CONSEQUENCES
amazon
UNINTENDED CONSEQUENCES

Fig. 38 - Unintended consequences impact analysis task - Group2.



SUSTAINABILITY

I ☑ ACCESS to PROFILE

Personal Data	ON	0.1	OFF	0
Social data	ON	0.1	OFF	0
Economic data	ON	0.1	OFF	0
Behavioural data	ON	0.1	OFF	0

H ☑ ACCESS to MONITORING

Microphone	ON	0.1	OFF	0
Camera	ON	0.1	OFF	0
GPS	ON	0.1	OFF	0
Sensors	ON	0.1	OFF	0

K ☑ INFERENCE

Behavioural patterns	ON	0.1	OFF	0
Routines	ON	0.1	OFF	0
Trends	ON	0.1	OFF	0
Preferences	ON	0.1	OFF	0

L ☑ INFERENCE ANALYSIS

Classification	ON	0.1	OFF	0
Labeling	ON	0.1	OFF	0
Probabilities	ON	0.1	OFF	0
Best option	ON	0.1	OFF	0

N	LEVELS OF AUTONOMY	FROM	1	TO	7
P	LEVELS OF REPARATION	FROM	1	TO	7
R	LEVELS OF ACCOUNTABILITY	FROM	1	TO	5

W ◉ CONTEXT - Highly sensitive areasDoes your action affect any of these are

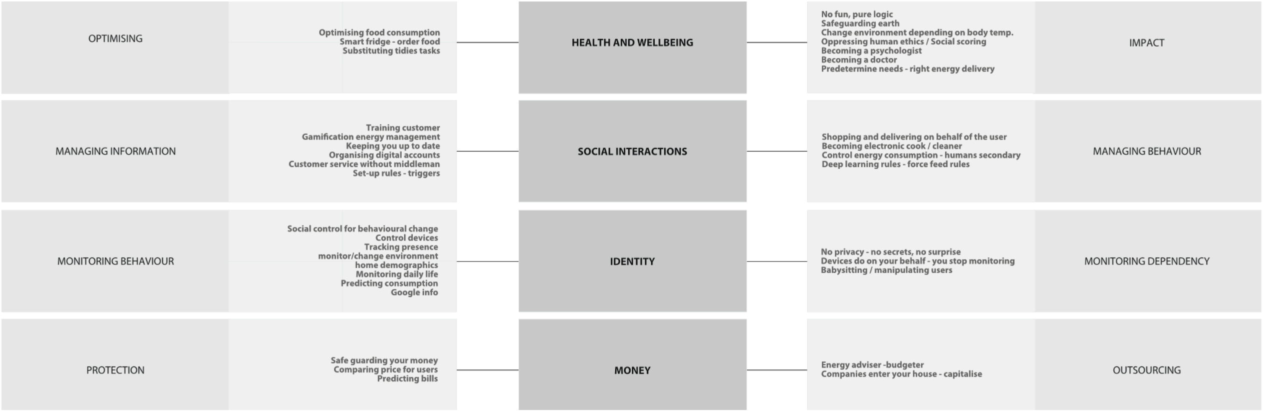
None	ON	0	OFF	0
Health and Wellbeing	ON	0.25	OFF	0
Identity	ON	0.15	OFF	0
Social interactions	ON	0.2	OFF	0
Money related activities	ON	0.1	OFF	0

X ◉ IMPACT - Unintended ConsequencesTest the impact of unintended outc

None	ON	0	OFF	0
Unhappy service	ON	0.1	OFF	0
Wrong Prediction	ON	0.15	OFF	0
Losing something in the service - e.g	ON	0.2	OFF	0
Service end in violence - e.g death/h	ON	0.25	OFF	0

Fig. 39 - Calibration weighting system is the context of sustainability. F. Galdon

The prospective workshop presented two main transitions of VAs in the context of energy management and consumption: from information management to behaviour management, and from concerns around privacy to concerns around the impact on health and wellbeing. In this context, the prospective element informed the weighting system in future scenarios: Health and wellbeing emerged as the most concerning area for users. Social interactions and identity followed it. Finally, economically related activities are the least concerning highly sensitive areas (Tab. 7). This investigation has been published and was presented at a leading conference on applied energy at MIT in May 2019 (AEAB2019). And contributes to other studies in the area of trust and virtual assistants in the context of conversational agents in sustainability (see for instance Gnewuch et al., 2018)



OPTIMISING	Optimising food consumption Smart fridge - order food Substituting tidies tasks	**HEALTH AND WELLBEING**	No fun, pure logic Safeguarding earth Change environment depending on body temp. Oppressing human ethics / Social scoring Becoming a psychologist Becoming a doctor Predetermine needs - right energy delivery	IMPACT
MANAGING INFORMATION	Training customer Gamification energy management Keeping you up to date Organising digital accounts Customer service without middleman Set-up rules - triggers	**SOCIAL INTERACTIONS**	Shopping and delivering on behalf of the user Becoming electronic cook / cleaner Control energy consumption - humans secondary Deep learning rules - force feed rules	MANAGING BEHAVIOUR
MONITORING BEHAVIOUR	Social control for behavioural change Control devices Tracking presence monitor/change environment home demographics Monitoring daily life Predicting consumption Google info	**IDENTITY**	No privacy - no secrets, no surprise Devices do on your behalf - you stop monitoring Babysitting / manipulating users	MONITORING DEPENDENCY
PROTECTION	Safe guarding your money Comparing price for users Predicting bills	**MONEY**	Energy adviser -budgeter Companies enter your house - capitalise	OUTSOURCING

Tab. 7 - Comparative analysis between present and future technological impact.

Discussion

This study presents **a future systematic mechanism that could establish trust in highly automated systems**. In order to do so, it integrates access, inferences, consequential reasoning, contexts, and actions to obtain a trust rating illustrating the potential impact of an action/skill in context. This approach provides a mechanism to simulate and/or reverse asymmetries in the system. In this process, prospective co-design workshops emerged as a defining method to address the potential impact of prospective technological development. This method enables the designer to go beyond what already exists and provides a suitable framework to access these future spaces for knowledge. The possibility of inferring potential and preliminary knowledge provides guiding knowledge for transforming the future as a space for applied ethics in design.

Following the presentation at MIT, I was invited to present the investigation at IDEO in Boston. As part of the presentation, the directors pointed out the potential of the calculator to be used as a simulation tool to facilitate the design of trusted AI. In order to understand to what extent this calibration mechanism could be used for this purpose, I implemented two evaluative exercises.

3.3.5.2 Tool Evaluation

- Tool evaluation study

First, I investigated whether the tool and the rating developed affected the design of digital systems. In order to evaluate the proposed tool, a workshop with twelve participants from the Masters in Research (MRes) programme at the Royal College of Art was implemented. The students represented a mix of backgrounds in fashion, textiles, architecture, computer science, industrial design, and engineering. The selection criteria for the participants were the same as those outlined earlier. (See Appendix 5 for detailed workshop material).

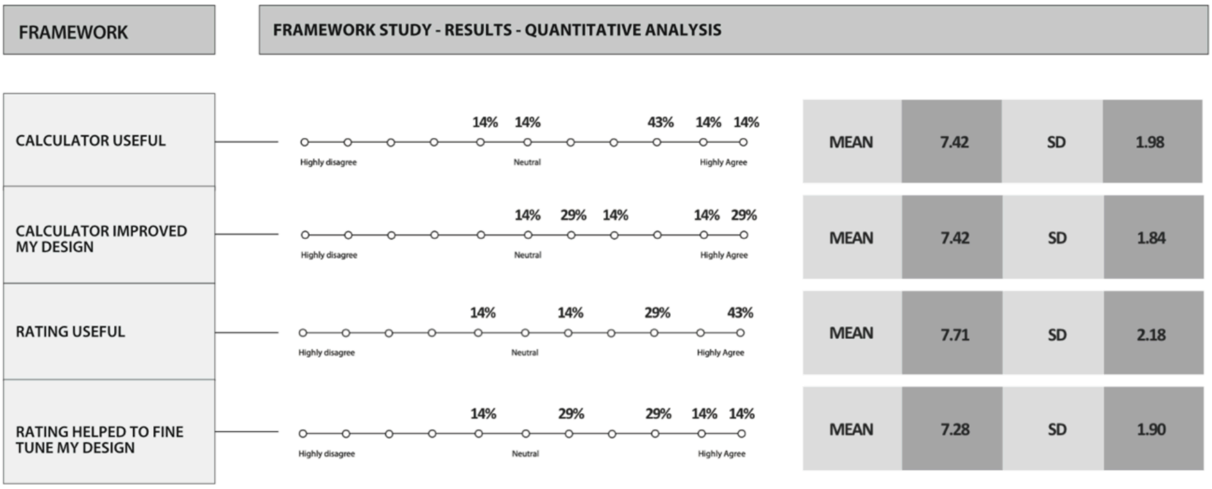
The author defined the central area of intervention; health and wellbeing. This area was selected specifically for its moral and ethical impact. Then a design task around a highly automated VA capable of diagnosing and providing treatment in the area of depression was structured. As part of the workshop, the author introduced a demo of Google's Duplex to illustrate the prospective nature of the system, and a small analysis underlined the critical characteristics of emerging VAs. Students had 50 minutes to complete this task. They were provided with the aforementioned tool in the form of a calculator with all the variables. This tool provided a trust rating to calibrate interactions beforehand.

In order to understand the validity of the tool, a comparative analysis was implemented to identify whether new elements that had not been considered in the proposed tool would emerge. Once the task was completed, the author designed a semi-structured questionnaire to understand four elements: the usefulness of the calculator, whether the calculator helped them to improve their design, the specific usefulness of the rating, and whether the rating helped them to fine-tune their decisions. The questionnaire consisted of two areas: a quantitative section asked participants to rate these elements by using an eleven-point Likert scale and a qualitative section asked participants to expand on why and how these elements had affected their design.

Discussion

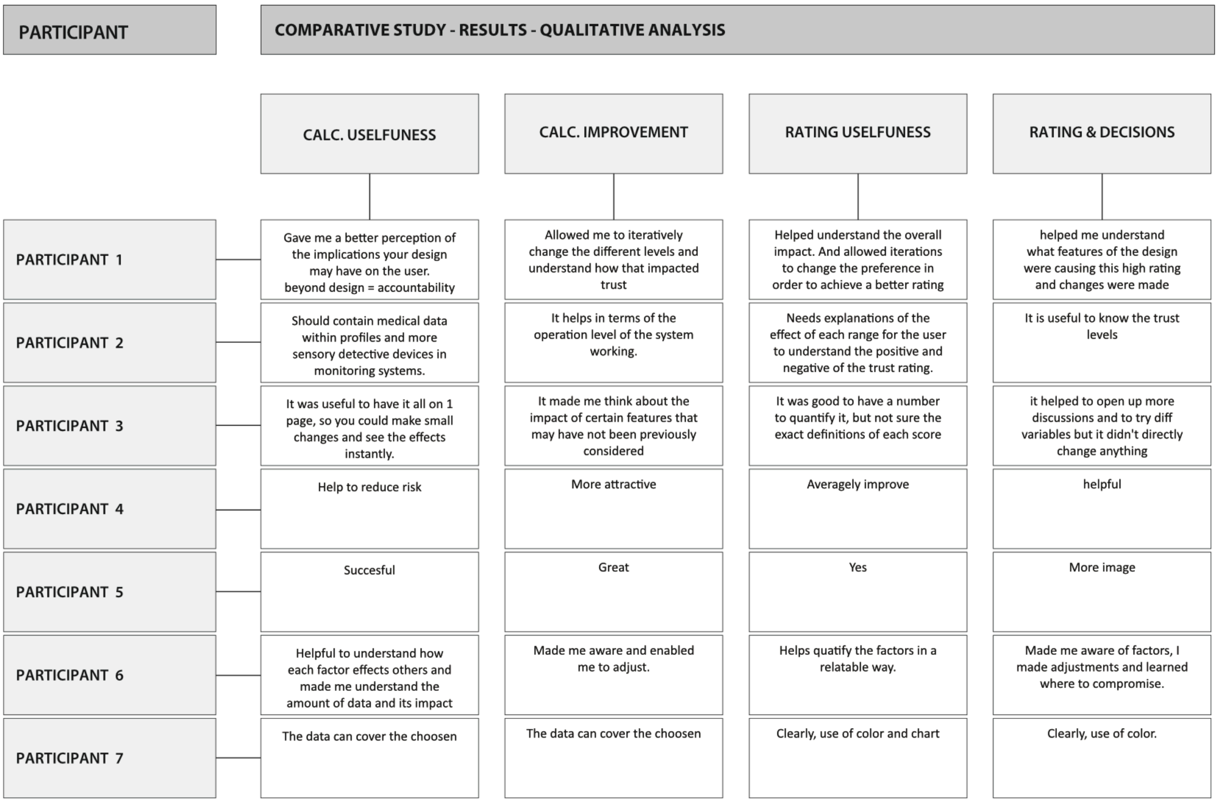
In terms of the usefulness of the tool in the form of a calculator, participants rated it with a 7.42 mean value. In terms of product improvement, participants also rated the usefulness of the tool with a 7.42 mean value. In terms of the rating usefulness, participants rated this element with a 7.71 mean value. Finally, in terms of the effect of the rating to fine-tune decisions, participants rated it with a 7.28 mean value (Table 8).

In qualitative terms, participants described how these elements affected their decisions by understanding the impact of the interaction beforehand. This exercise led to participants reducing risks by having a better perception of the implications their design may have on the user's trust. From the results presented we can establish that the framework and its mode of calculation are useful to facilitate the design of trusted systems (Table 9).



FRAMEWORK	FRAMEWORK STUDY - RESULTS - QUANTITATIVE ANALYSIS				
CALCULATOR USEFUL	14% 14% 43% 14% 14% Highly disagree Neutral Highly Agree	MEAN	7.42	SD	1.98
CALCULATOR IMPROVED MY DESIGN	14% 29% 14% 14% 29% Highly disagree Neutral Highly Agree	MEAN	7.42	SD	1.84
RATING USEFUL	14% 14% 29% 43% Highly disagree Neutral Highly Agree	MEAN	7.71	SD	2.18
RATING HELPED TO FINE TUNE MY DESIGN	14% 29% 29% 14% 14% Highly disagree Neutral Highly Agree	MEAN	7.28	SD	1.90

Table. 8 - Results - Quantitative analysis. F. Galdon



PARTICIPANT	COMPARATIVE STUDY - RESULTS - QUALITATIVE ANALYSIS			
	CALC. USELFUNESS	CALC. IMPROVEMENT	RATING USELFUNESS	RATING & DECISIONS
PARTICIPANT 1	Gave me a better perception of the implications your design may have on the user. beyond design = accountability	Allowed me to iteratively change the different levels and understand how that impacted trust	Helped understand the overall impact. And allowed iterations to change the preference in order to achieve a better rating	helped me understand what features of the design were causing this high rating and changes were made
PARTICIPANT 2	Should contain medical data within profiles and more sensory detective devices in monitoring systems.	It helps in terms of the operation level of the system working.	Needs explanations of the effect of each range for the user to understand the positive and negative of the trust rating.	It is useful to know the trust levels
PARTICIPANT 3	It was useful to have it all on 1 page, so you could make small changes and see the effects instantly.	It made me think about the impact of certain features that may have not been previously considered	It was good to have a number to quantify it, but not sure the exact definitions of each score	it helped to open up more discussions and to try diff variables but it didn't directly change anything
PARTICIPANT 4	Help to reduce risk	More attractive	Averagely improve	helpful
PARTICIPANT 5	Succesful	Great	Yes	More image
PARTICIPANT 6	Helpful to understand how each factor effects others and made me understand the amount of data and its impact	Made me aware and enabled me to adjust.	Helps quatify the factors in a relatable way.	Made me aware of factors, I made adjustments and learned where to compromise.
PARTICIPANT 7	The data can cover the choosen	The data can cover the choosen	Clearly, use of color and chart	Clearly, use of color.

Table. 9 - Results - Qualitative analysis. F. Galdon

This evaluation exercise investigated an innovative multi-dimensional scalar tool integrating post-interaction element such as accountability and reparation, and integrating unintended actions, contexts, access, and inferences as fundamental variables to facilitate the design of trust from a consequential perspective on unsupervised highly automated computational systems. As part of this process, a form of calculation emerges to facilitate the calibration of trust in the context of HAS. From the results presented, participants support the usefulness of the tool and the mode of calculation to design trusted systems.

This investigation has been peer-reviewed, published, and was presented at an international conference in technology at CHUV Lausanne in April 2020 (IHIET2020).

- Tool evaluation Comparative study

Then I investigated how this tool, which fundamentally revolves around levels of control, compared to other frameworks used to design trust (specifications and principles) via a comparative study. For the last forty years, human factors approached the design of automated systems by articulating Levels of control as a design strategy to appropriately calibrate trust in order to achieve performance and safety goals (Sheridan, 1978). However, Principles have recently been proposed as a design strategy from social and ethical perspectives to address trust (Floridi, 2019). Finally, Specifications are being proposed from a computational perspective as a design strategy to address the rising concerns about highly automated systems (Ortega, 2018).

This section presents a comparative study of these frameworks to understand which of the three frameworks is best suited to design trust in the context of HASs. It will do so by addressing trust design in four case studies specifically designed to address the rising concern about these systems in the area of health and wellbeing.

In this regard, I organized a workshop with 12 participants from the Masters in Research (MRes) program at the Royal College of Art. The selection criteria for participants are the same as outlined earlier.

In order to evaluate the validity of the frameworks presented, I implemented a comparative study. According to Bukhari (2011), comparative studies analyse and compare two or more objects or ideas to examine, compare, and contrast them in order to show how two or more subjects are similar or different. Building on this perspective, the author built a comparative study of the three main frameworks acknowledged to design trust in AI: specifications, principles, and levels of control, in order to identify which one is best prepared to address the rising concern about highly automated systems. In this context the author aimed for a mixed methodology, combining constructive approaches in the form of a design workshop, experimental design to control some variables, a semi-structured questionnaire, and a post-activity debate synthesis to evaluate the outputs.

In order to address the task at hand, the author defined the main area of intervention: health and wellbeing. Then four exercises were structured around systems capable of diagnosing and providing treatment in the areas of anxiety, obesity, depression, and addiction. The author introduced a video demonstration of Duplex to illustrate the prospective nature of emerging VAs and a small analysis that underlined the key characteristics of the system. The participants had 50 minutes to complete each task, which consisted of four parts:

- A mapping exercise to identify potential interventions
- An introduction to a design framework.
- An 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.
- An interaction task consisting of a user journey and a potential design intervention. This part was structured into three areas; before the interaction, during the interaction, and after the interaction. (See Appendix 5 for detailed workshop material).

The first exercise introduced no framework. It operated as a control mechanism to understand what the participants 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 final exercise, a multi-dimensional framework was presented in collaboration with a trust calculator to facilitate participants' output by inserting a mode of calculation by which a trust rating could be obtained.

Once all the exercises were completed, the author introduced a semi-structured questionnaire to understand which framework was best suited to design trust in HASs. The questionnaire consisted of two areas: a quantitative area asked participants to rate the four frameworks proposed – with no framework, principles, specifications or levels, by using an eleven-point Likert scale – and a qualitative area, asking participants to define the pros and cons of each framework.

Discussion

In the quantitative area, Specifications emerged as the most favoured framework by the participants, who rated it as 7.57 in mean value. It was followed by Levels of control, rated as 6.71, and no framework, as 6.57. The least favoured framework was Principles, with a 5.71 mean value (Table 10).

When reviewing the qualitative data obtained by asking participants to describe the pros and cons of each framework, they praise Specifications for their semi-structured nature, which provides them with a flexible, yet 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 pointed to the limitations of Specifications to address trust in ever-evolving systems, as it is a one-time a priori intervention that does not allow for a posteriori rectification. It is described as a powerful tool to understand user needs but is limited in terms of designing trusted systems, especially in the context of HASs, with unsupervised and ever-evolving capabilities (Table 11). In this context, Levels are described as a tool to implement quick adjustments; they are beneficial and enhance distributed self-optimisation to maintain control over the system. Furthermore, when integrating the

FRAMEWORK	COMPARATIVE STUDY - RESULTS - QUANTITATIVE ANALYSIS														
	Highly disagree					Neutral					Highly Agree				
LEVELS OF CONTROL				14%	14%			29%	29%		14%	MEAN	6.71	SD	2.24
PRINCIPLES					29%		14%	29%	14%		14%	MEAN	6.57	SD	1.98
SPECIFICATIONS					14%		14%	14%	14%	29%	14%	MEAN	7.57	SD	1.91
NO FRAMEWORK			14%	14%	29%			14%			29%	MEAN	5.71	SD	3.05

Table. 10 - Results - Quantitative analysis. F. Galdon

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

Table. 11 - Results - Qualitative analysis. F. Galdon

calculator into the levels and providing a form of calculation, participants described this combination as useful in reducing risks, integrating a critical dimension into product development, and enhancing explainability in the design process. Principles, though, are seen as a philosophical element that can prompt relevant debates. Finally, no framework is described as open, yet too loose in focus and too abstract to address the rising concern presented (Table. 11).

These outputs are significant because they correlate with a paper published in *Nature* in November 2019 by the Oxford Internet Institute, claiming that Principles are not enough to design trusted AI systems (Mittelstadt, 2019). In this context, instead of providing a categorical excluding output, I propose to build an integrative multi-dimensional design framework by acknowledging the critical beneficial elements of the three main frameworks by distributing these paradigms over time. Based on these results, Levels of control emerge as the most reliable option to design trust in HASs, 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.

This evaluation exercise presents leading insights by providing a comparative study of proposed frameworks to design trust in AI. From the results presented, participants support the usefulness of Levels of control to design trusted systems. Although limited in scale, the results provide a highly relevant contribution to knowledge, as no other study we identified has compared all these elements simultaneously. In the process, it provides knowledge for future actions via the categorisation of existing frameworks to address the rising concerns about trust in AI. This investigation has been published and was presented at an international conference at CHUV, Lausanne, in April 2020 (IHIET2020). These two studies contribute to other papers in the area of trust and virtual assistants in the context of conversational agents in healthcare (See for instance Laranjo et al., 2018).

In this chapter, I have implemented the proposed methodology of Prospective Design. As a result, seven publications have been generated to evaluate different aspects of **a future systematic mechanism that could establish trust in highly automated systems**. All the methods and their underlying techniques aim to make fully operational Prospective Design to contextualise and make fully operational Glanville's concept of knowledge "for transforming the future" as a probabilistic knowledge ontology (Glanville, 2005).

With the calculator as a simulation tool, I try to establish a system of control in a second-order cybernetic sense in which the concept of control exists 'between things' (the designer and the system).

The next section will evaluate the proposed methodology.

3.4 MODEL EVALUATION

In order to evaluate the proposed methodology, I used a Q-experimental design methods perspective. According to the UNICEF Methodological Briefs Impact Evaluation No. 8,

> "Quasi-experimental design by definition lacks random assignment, however. Assignment to conditions (treatment versus no treatment or comparison) is by means of self-selection (by which participants choose treatment for themselves) or administrator selection (e.g., by officials, teachers, policymakers and so on) or both of these routes" (Shadish et al., 2012 on White, H., & S. Sabarwal, 2014)

There is a debate in academic circles about whether quasi-experiments are natural experiments or not. The fundamental difference is the criterion for assignment. In quasi-experiments, this criterion is selected by the researcher, whereas in natural experiments the assignment occurs "naturally", without the researcher's intervention.

As stated in 2.8 Q-Experimental are ideal when, for instance, "a new programme is implemented" (Leatherdale, 2019, p.19). This flexibility and approach to the new is crucial for design research in which the notions of full control of the variables and repeatability are impossible. These aspects of natural experiments make them ideal for evaluating a research methodology in the context of design and the consequences of our choices. As a testament to their value, the 2021 Nobel prize of economics has been awarded to Natural experiments.

According to Leatherdale (2019), the methodological tool used to evaluate the impact of the intervention is called a "natural experimental study". And the particular design used by a researcher to evaluate a natural experiment "will largely depend on the type of data that are available when the natural experiment occurs" (Leatherdale, 2019, p. 19).

On one hand, quasi-experiments are subject to concerns regarding internal validity, because the treatment and control groups may not be comparable at the baseline. Furthermore, the knowledge produced by this type of experiment revolves around approximations rather than conclusions, due to a variety of extraneous and confounding variables. Therefore, conclusions of causal relationships are difficult to determine. The knowledge generated is probabilistic and preliminary in nature, as social and personal factors may affect the outputs. In order to address these issues, I have conducted a contextual evaluation against the most reliable model in the field; The EU Commission's latest paper on AI. On the other hand, since quasi-experiments are natural experiments, findings may be applied to other subjects and settings, allowing for some generalisations to be made about the inquiry. This aspect enhances transferability, an attribute that aligns better with design practice, as repeatability is an impossible condition for a practice concerned with the new, yet-to-be or not-fully-formed. Furthermore, this method is efficient in longitudinal research that involves longer time periods that can be followed up in different environments. This aspect is relevant, as design is dependent on time and context.

3.4.1 Design

In quasi-experiments we have several possible variations. The first major distinction is whether there are one or two groups of participants. The second is how often measurements are taken (Campbell & Stanley, 1963). In this context, we have a range of designs:

> - The (single-multiple group) post-test-only design: a design in which participants are given a treatment and then tested.
> - The (single/multiple group) pre-test/post-test design: A design in which participants take a pre-test, then receive some treatment, and then take a post-test measure.
> - The (single/multiple group) time-series design: a design in which participants are measured repeatedly before and after a treatment. (Jackson, 2009, p. 321)

In this case, I implemented an adaptation of the multiple Group post-test-only design, also known as the Non-equivalent Control Group Post-Test-Only Design. In this type of design, the control group is non-equivalent, meaning that "participants are not assigned to either the experimental or the control group in a random manner" (Jackson, 2009, p. 323). They are members of each group because they have decided to participate in a specific workshop call. The pre-test was unnecessary to establish equivalence between groups because all participants were design students at the Royal College of Art and the workshops were both about the future technological development of VAs.

The treatment was the main variable (a simplified version of the methodology versus a complete version of the methodology). And the post-test analysed differences in outputs. In this context, the experimental group tests/assess the model as it is intended, and the control group is presented with a simplified version of the model. Therefore, the design, or in this case, the methodology, can be said to have caused some difference in outcomes between the experimental and control groups. In order to evaluate the final model, two workshops have been used to test critical aspects of the methodology proposed.

3.4.2 Workshops

The first workshop invited 20 participants from the School of Design at the RCA to test, on the one hand, differences between the group and individual work, and on the other the simplified systematic analysis of unintended consequences presented by Mark Michaels (Michaels, 2019) (Fig. 40). The participants were divided into four groups of five members.

Given a potential technological development, the framework presented by Michaels asked participants to analyse four elements: anticipated desired, anticipated undesired, unanticipated desired, and unanticipated undesired potential outputs. As a result, the anticipated quadrants were better developed, with 61 proposals, whereas the unanticipated aspects of product development presented 54 proposals in total from the participants.

Unanticipated undesired outcomes presented a distinct challenge for participants, as they referenced known issues. Answers were logical, rational, and expected. There was a lack of originality and a reluctance to go "beyond"'. The author had to instigate debate by introducing some examples. However, instead of widening the scope of outputs, these examples become replicated by variation or integration. Occasionally some participants proposed exciting ideas, but the group dynamics demanded consensus and prevented them from going 'beyond' what they already knew, thus limiting abductive thinking and jeopardising prospective strategies to build trust. In anticipatory contexts, it is fundamental to go 'beyond' what already exists. Only if you can imagine contentious developments can you develop strategies to mitigate prospective consequences and design trusted systems.

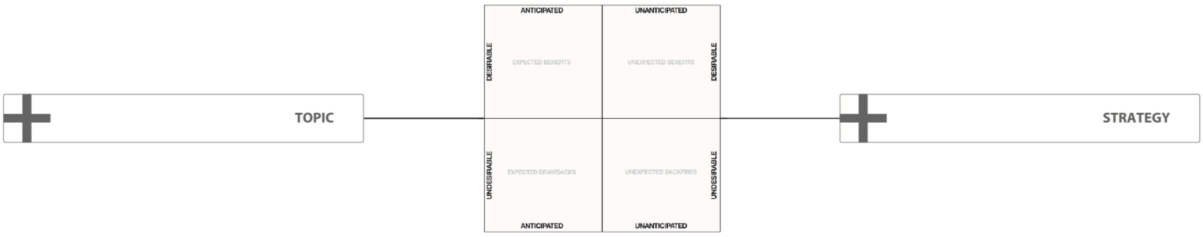


Fig. 40 - Mark Michaels (2019) Consequential analysis toolkit.

The second hour of the first workshop aimed to carry out the same task again from an individual perspective. A booklet for individual development was distributed among participants. The engagement was articulated around the idea that they could re-appropriate the method by integrating their own individual research into the process. Half an hour into the task, half of the participants left the workshop. It seems that they need constant engagement, and when requested to conduct individual work and reflect within themselves, they tend to disengage and abandon the task. The other half engaged as expected, with 20% of participants engaging vigorously, to the extent of asking whether they could carry on with the task at home after the workshop. However, outcomes were built from the previous task. Again, a lack of "going beyond" what is already known or proposed was present (Appendix Workshop 1).

The second workshop invited 10 participants from the School of Design and Architecture at the RCA to test and improve a multi-layered approach to systematically analysing consequences by addressing contexts and actions to propose mitigating strategies. Participants were divided into two groups. The workshop was structured completely to operate as a group task in order to maintain engagement. All the participants completed the two-hour workshop, and they engaged consistently through all the stages (Appendix Workshop 2).

The second workshop aimed to investigate anticipatory analytical skills further. As a result, the author introduced a range of variations. First, students mapped the current state of development (what a VA can do today). Then, in order to address the challenge of originality and the lack of "going beyond", it introduced a “what if …?” approach to allow participants to break away from their logical and rational thinking and project possible or potential developments for the technology (Fig. 41). This task was successful, and unexpected outcomes emerged, allowing participants to go 'beyond’ what already exists. This approach included positive and negative outcomes. As an example, the outputs presented food-related issues and the smart fridge as highly relevant in the context of energy consumption and management for future developments of VAs in this area. This result was highly unexpected, and when presenting this particular outcome at MIT and IDEO it was received with surprise, yet made total sense in relation to the future impact of the smart fridge. These processes enable participants to develop strategies to mitigate prospective consequences and design trusted systems.

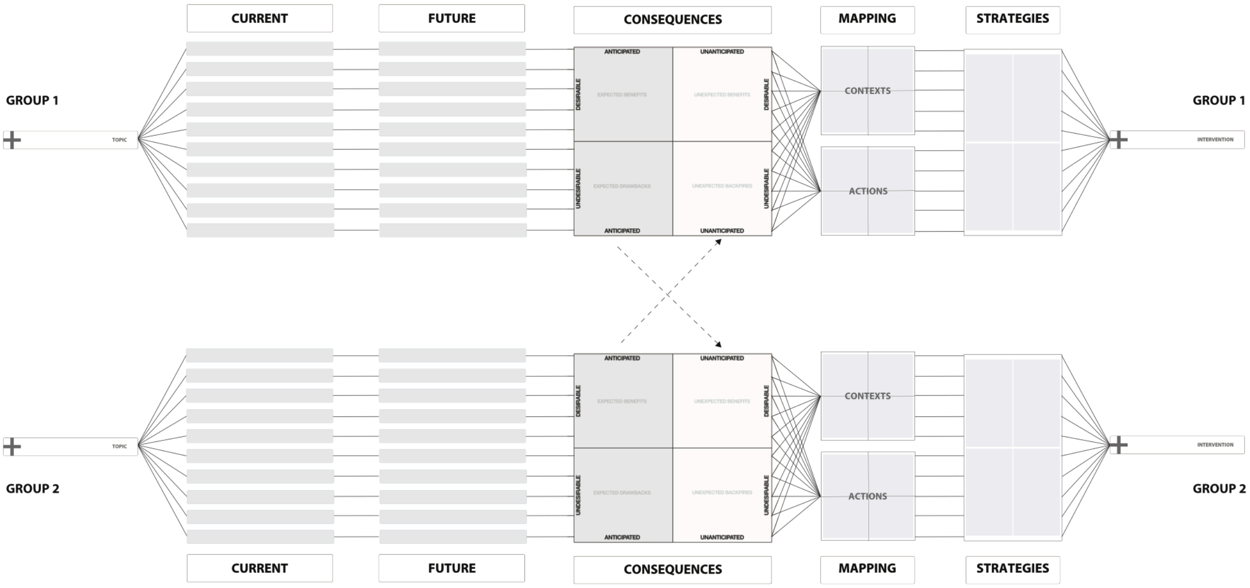


Fig. 41 - Consequential analysis for Prospective Design. F. Galdon.

- Discussion

In terms of outputs, the workshop aimed to evaluate critical aspects of the methodology proposed. The workshop aimed to understand whether prospective insights could be transformed into applied ethical interventions and grounded in the real world by applying a systematic analysis between the insight and the design activity to design trusted systems. The system analysis consisted of a three-level analytical process of the system at hand. First, the participants were asked to conduct the consequences quadrant used in Workshop 1, then each group mapped the anticipated desired and undesired, and by confronting both groups, the unanticipated emerged for each group. This element presented participants with their own limitations and enhanced self-criticality. They then mapped the prospective outcomes in terms of the impact on contexts and the impact of actions. This analytical step allowed them to understand the impact of contexts and actions on users. Finally, participants were asked to complete a design activity consisting of developing preventive strategies for the potentially harmful and power-asymmetric interactions they had mapped. They were requested to use counter-fictional principles to transform the dystopic into real-world strategies that could be applied. The results presented strategies aiming to ground prospective insights in potential real-case interventions that aimed to reverse asymmetries to build trust. These outputs support the main question; **what kind of design methodology would enable us to establish and maintain trust in highly automated and unsupervised systems?**

In this study, the author proposes Prospective Design as a methodology to address unintended consequences to enhance trust. Trust plays a fundamental role as a mechanism to deal with uncertainty and risk. Trust formation is a dynamic process, starting before the user's first contact with the system and continuing long thereafter. In this context, understanding how contexts and actions and the unintended consequences that derive from them, affect trust in HASs is fundamental for the adequate design of these systems. The proposed methodology combines systems analysis with extrapolations and constructivist perspectives to address the rising concerns of exponential technological developments, providing an applied ethical model for designing future(s).

In the results presented, I suggest a need to include prospective ethical frameworks in design to involve students in ethical issues: to go beyond what already exists, as well as beyond the positive impact of technology and design strategies to address and/or mitigate unintended consequences, as they are fundamental for the optimum development of society; to propose that things can be otherwise.

In the process, it challenges and develops current notions in design research based on technological progress that revolves around product development or speculations to a model based on ethical responsibility, which places equal value on the process of design and the impact of the system on society. In this context, abductive thinking becomes the primary design mindset in driving the transition from current to potential states, leading to the mediation of anticipated and non-anticipated consequences. Success, therefore, resides in generating prospective real-world strategies/products/interventions aimed at mitigating unintended consequences to enhance trust. The Prospective Design framework introduces a process to deal with the increasing complexity of wicked problems, black-box technologies, and AI/ML technology acceleration, enhancing social values and ethical principles in the process.

3.5 CONTEXTUALISATION

In February 2020 the European Commission (EC) released a White Paper on AI. They created a group of the 53 key AI experts in Europe, led by Professor Luciano Floridi, Director of the Oxford Internet Institute, and regarded as one of the most eminent researchers in the field of ethics in AI (Fig. 42). Their backgrounds range from ethics and artificial intelligence to the philosophy of law. It includes all the major companies in the field, such as Apple, Microsoft, Twitter, and Facebook.

Fig. 42. Group of 53 experts in Artificial Intelligence. European Commission

The EC group's paper states that the design of artificial intelligence is about building trust. This process is structured in a lifecycle. It starts before the interaction and then follows beyond the interaction.

In this context, the EC group identifies two main areas which are highly sensitive; democracy, which they embody around the idea of rights, and the environment, which they embody around the idea of energy and sustainability. If I map my publications against this framework, we can see how my papers cover the breadth of this spectrum. I have produced two papers on energy consumption and management, one paper proposing a new digital right, two papers proposing reparation strategies after the interaction, two more papers proposing calibration strategies during the interaction, and one paper proposing a simulation tool to address elements before the interaction (Fig. 43). Furthermore, I have published two papers proposing a new ontology and epistemology to enable a new methodology: Prospective Design. This fundamental aspect of how to approach these problems is missing in the EC White Paper.

If we deconstruct the document page by page, we can implement a detailed comparative evaluation. For example, on page 9 is a discussion of trust, which the EC group defines as a problem of asymmetries leading to unintended effects. As I have documented in this Ph.D., this was the main topic and problem I have identified to resolve during my inquiry. As can be seen below, my paper 'Prospective Design' addresses this. Other examples are, for instance on page 12. The White Paper talks about the quintessential problems involving AI: uncertainty, which revolves around black boxes, complexity, predictability, and autonomous behaviour. In these areas I have published; one paper addressing black-boxes, another paper addressing complexity, two papers addressing predictability, and one paper addressing automated behaviour. On page 14, the EC group considers problems in exponentiality. The transition from products to services is identified, from launch to updates, and the role of third parties. All these elements were identified in my inquiry, thus validating the process. This was very challenging because, as mentioned above, there was no handbook on how to design trust in AI. In this area, I have published three papers addressing all these problems.

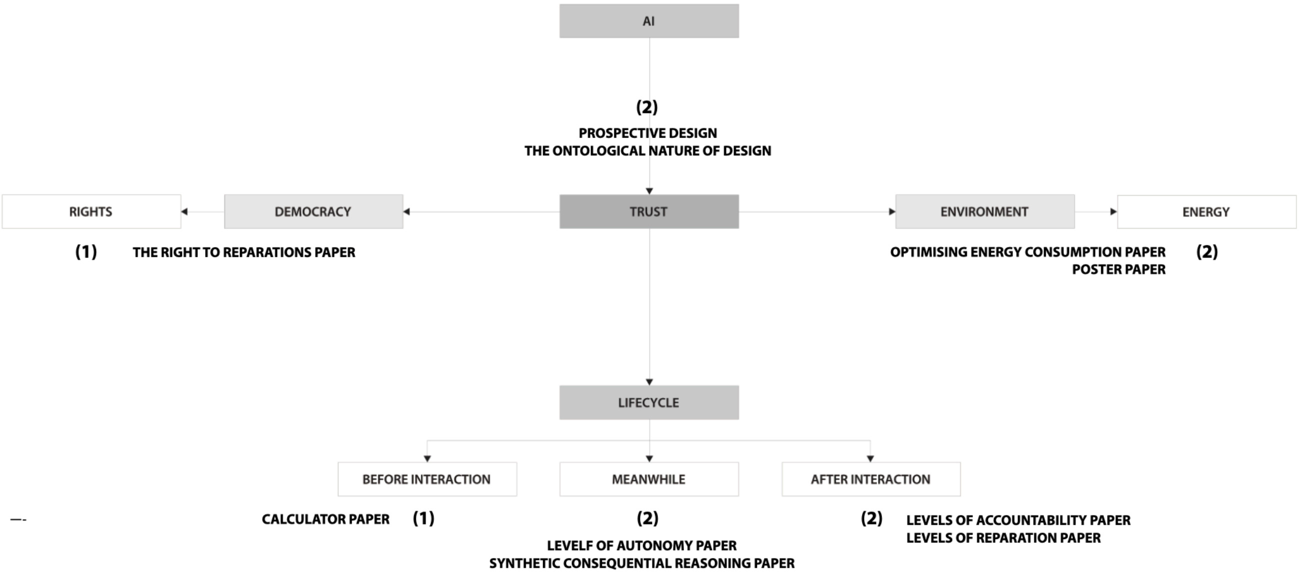

AI
(2)
PROSPECTIVE DESIGN
THE ONTOLOGICAL NATURE OF DESIGN
RIGHTS
DEMOCRACY
TRUST
ENVIRONMENT
ENERGY
(1)
THE RIGHT TO REPARATIONS PAPER
OPTIMISING ENERGY CONSUMPTION PAPER
POSTER PAPER
(2)
LIFECYCLE
BEFORE INTERACTION
MEANWHILE
AFTER INTERACTION
CALCULATOR PAPER
(1)
(2)
LEVELF OF AUTONOMY PAPER
SYNTHETIC CONSEQUENTIAL REASONING PAPER
(2)
LEVELS OF ACCOUNTABILITY PAPER
LEVELS OF REPARATION PAPER

Fig. 43. This diagram embodies the fundamental argument of The European Commission's White Paper on AI and how my publications address them. It explains that designing AI is about building trust. It recognises two areas with high impact; democracy (digital rights) and the environment (energy and sustainability), and trust needs to be implemented through its lifecycle. My papers present insights in all the main areas discussed.

As can be seen, my papers cover every aspect of that document except how to build datasets, and infrastructures (see for example Mortier et al., 2014 work on Human Data Interaction). These are the only areas that I have not covered because they go beyond the scope of my research, which was focused on algorithms. This entire body of research has been submitted to the EU Commission's participatory process for their consideration.

By implementing Prospective Design, I could deal with notions of preparedness, readiness, and appropriateness — which demand to go one step further. This framework enables the researcher to go beyond what exists and investigate the potential unintended consequences emerging from technological developments. And as we have seen in this contextual analysis, the results have been promising. What is meaningful about this comparative analysis is that in the same period of time that this group of fifty-three experts has taken to identify the problems, I have been able to identify the problems and provide solutions. Implementing Prospective Design enabled me to implement prospective thinking. Which in turn allowed me to develop prospective research.

02/2020	EUROPEAN COMISSION WHITE PAPER - ON ARTIFICIAL INTELLIGENCE : A EU		
	ISSUE	**PROBLEM**	**DEFINI**
PAGE 9	**TRUST**	ASYMETRIES	UNINTENDED EFFECT
PAGE 10	**HARM**	MATERIAL INMATERIAL	DEATH LOSS
PAGE 12	**PROBLEMS**	UNCERTAINTY	BLACK-BOX COMPLEXITY UNPREDICTABILITY AUTONOMOUS BEHA
PAGE 13	**CONTEXTS**	THEY ARE VERY RELEVANT	

PAGE 14	**LIMITATIONS**	EXPONENTIALITY	FROM PRODUCTS TO
			FROM LAUNCH TO U
			FROM MAKER TO 3rd

PAGE 15	**LIABILITY**	PROTECTION	
PAGE 17	**DEVELOPMENT**	RISK	HIGH/LOW
PAGE 19	**DESIGN**	DATA TRANSPARENCY ALGORITHM	DIVERSITY AND PRIV INTENTIONS ROBUSTNESS/ACCUR
PAGE 21	**OUTPUTS**	HUMAN CONSENT	
PAGE 22	**ADDRESSES**	ACTORS	
PAGE 24	**COMPLIACE**	INTERVENTIONS	EX ANTE EX POST

UROPEAN APPROACH TO EXCELLENCE AND TRUST

TION	PUBLICATIONS ADDRESSING THE ISSUE
TS	PROSPECTIVE DESIGN (1 PAPER)
	CASES/SURVEYS (3 PAPERS) CASES/SURVEYS (3 PAPERS)
VIOUR	PROSPECTIVE DESIGN (1 PAPER) SYNTHETIC CONSEQUENTIAL REASONING (1 PAPER) LEVELS OF ACCOUNTABILITY/REPARATION (2 PAPERS) DIGITAL RIGHT (1 PAPER)
	PROSPECTIVE DESIGN (1 PAPER)
SERVICES PDATES PARTIES	LEVELS OF ACCOUNTABILITY/REPARATION (2 PAPERS) LEVELS OF ACCOUNTABILITY/REPARATION (2 PAPERS) LEVELS OF ACCOUNTABILITY/REPARATION (2 PAPERS)
	LEVELS OF ACCOUNTABILITY/REPARATION (2 PAPERS)
	RISK MATRIX AND CALCULATION TOOL (4 PAPERS)
ACY RACY/RELIABILITY	LEVELS OF AUTONOMY/ACCOUNTABILITY/REPARATION (3 PAPERS) LEVELS OF AUTONOMY/ACCOUNTABILITY/REPARATION (3 PAPERS)
	LEVELS (3 PAPERS)
	LEVELS OF ACCOUNTABILITY (1 PAPERS)
	SIMULATION (2 PAPER) CALIBRATION (1 PAPER) REPARATION (2 PAPERS)

4 CHAPTER

MODEL DISCUSSION

4.1 INTRODUCTION

This inquiry raised a large number of questions, many relating to empiricism and knowledge, prospectivity and dissemination: should design practice be reactive or proactive? How can we access the future in a reliable way? How do we identify the appropriate areas for intervention? Do we have to involve other people? If so, who, how, when, and why? How can we go beyond fiction and ground prospective insights in real-world interventions affecting change? Where do we implement the intervention?

In this chapter, I will discuss key insights emerging from this Ph.D. I will then use the case study to outline and critically analyse a possible method for prospecting futures. In its development, I have considered positions towards design from the fields of Critical and Speculative Design (Dunne), Co-Speculation and Transition Design (Lohmann) and connected them with insights from Trust (Botsman), Action and Prospectivity (Glanville), Relationality (Blauvelt), and diagrammatic de-materialisation.

My aim is to demonstrate how this method functions and to explain its underlying ethically based principles, leading to the concluding chapter, in which I will give an outlook on its potential components and dimensions. First, however, I will outline a number of key insights from the inquiry, grouped under the following headings: Preliminary ideas evolution; The object of inquiry; Diagramming; Participation and design; Proposed methodology; Futuring; Process and limitations, and Contributions.

4.2 PRELIMINARY IDEAS EVOLUTION

When I began my research project, I expected the outputs to lead to new knowledge of the kind that I now characterise as "low level” insights.

The original proposal for my Ph.D. presupposed that, for example, a designer interested in designing trust would be able to use a range of principles to design an object/service. The outputs produced in this research via comparative studies have scaled back my initial expectations that principles will broaden designers’ understanding of their designs. Contrary to this, it has been by implementing a lifecycle systems approach with an ethical perspective that I have extended my knowledge and understanding of the nature of trust and the type of knowledge it can contain at high and low levels. This approach has proved to be more reliable than the initial proposition.

Furthermore, while the multi-dimensional scalar system presented is indeed able to display the dynamics of the system in the context of unexpected events, as we have seen, there is an extended context of every person pointing towards cultural differences which demands properly designed tools capable of addressing extended contexts and actions. The integration of these extended variables has provided critical accuracy to build an even more reliable model.

Crucial to enabling these insights are not only the tools or methods generated but also the process and knowledge of creating them in collaborative efforts with participants, scholars, and interested researchers. By participating in the design process, designers were more aware of the characteristics of trust and the potential for designing it to prevent unintended outcomes. These aspects raised participants' deontological understanding, an outcome which I did not anticipate. Collaborating and involving them in the process made me, and them, more informed and critical about tools, methods, and every aspect of the system.

One of the defining issues of this investigation was the restricted access to the current technological state of the art, as these developments are conducted by private companies and are protected by IP legislation. Concerns emerged when these limitations prevented an adequate assessment of what was happening. In this context, I approached this limitation by implementing multiple temporal strategies through prospective research. This approach led to a range of new methods such as probabilistic extrapolations and asymmetries to address access and exponential technological development. These matters need to be considered when developing research into exponential technological tools and systems.

Finally, the limitations of current models of research grounded in the present generated a struggle which transformed the research and provided a reflective space which led to impactful insights to reconfigure current models of research design. These processes provoked investigations into the chronology and origin of design research (Archer, 1978), prospective practices (Glanville, 2005), the origin of knowledge, and its categorisation (Aristotle, 1984/1998/2000), which led to a fundamental contribution into its original forms and intentions.

4.3 THE OBJECT OF INQUIRY

Although VAs are still in their infancy, the preliminary insight was that investigating the prospective developments of this type of interaction device would reveal the particular challenges of highly automated interactions for scholarly research.

This hypothesis has been successful. It has been so because these devices present a dense cluster of domestic interactions, attracting a vast number of users in the process, but primarily because they are developing interactions into highly sensitive areas such as health and wellbeing, economy-related activities, social interactions, and identity. These contexts proved a fertile source of prospectivity and feedback through workshops and co-design activities. Examples of this were the relevance of cultural contexts from co-design workshops, as well as the many insights provided by practitioners and researchers in response to presentations. The domesticity of this device enabled relationality and understanding.

VAs are established nodes in our social fabric. These systems are capable of attracting a wide and expanding range of publics primed for services and knowledge transfer in large commercial environment. They offer a wide range of actions, services, information, and activities, increasingly interactive and co-created with independent designers/developers and large organisations. They are an invaluable object of inquiry into HASs because their interactions are acquired, archived, contextualised, interpreted, curated, enabled, and instructed by artificial systems, and have no system of accountability embedded to prevent or mitigate unintended consequences. Potential interactions between systems and users are a rich source of knowing and knowledge, from theoretical to practical, as each interaction is an embodied expression of a vision. These visions represent a concept or a form of technical, social, commercial, or intellectual power relationship contextualised within its time, environment, culture, and socio-economic infrastructure. Access to potential VA interactions as an archive of potential futures enables design research not only to engage in future-focused prospectivities based on impact in action but also to draw conclusions about potential technological developments.

4.4 DIAGRAMMING

According to Buckley, diagrams "serve several purposes, for example, to act as a direct and indirect means of analysis, representation and catalyst for discussion" (Buckley, 2013. p. 149). They serve a multiplicity of roles and "can thus be an extremely powerful tool in that they can have relevance at all stages of the research process" (Buckley, 2013. p. 149).

In the context of AI, diagrams have been traditionally used in computer science as schematic tools to explain the internal functioning of a system (circuit boards). This approach was translated in this thesis to explain the interactive elements of the system functioning in the context of AI – schematics of interaction. This technique facilitated the understanding and communication of de-materialised systems. In this process, diagrams also become reflective tools. They helped me to structure knowledge in a manageable way to implement critical analysis via comparative or relational studies.

As a synthetic tool, they represent a reduction of reality, but this reduction facilitated understanding. Furthermore, this tool was particularly helpful to facilitate cross-disciplinary inquiry, and this element allowed me to find relationships between disciplines and fields.

In this thesis, diagrams become ideal tools for conducting research and embodying knowledge in the context of de-materialised design. To support this argument, in 2019, Kate Crawford and Vladan Joler's "Anatomy of an AI System" infographic of a VA won the Design Museum's Beazley Design of The Year Award.

4.5 PARTICIPATION AND DESIGN

As design research has moved from the industrial and scientific to social and humanistic questions, the role of the "other" has transitioned from passive user/consumer to active citizen/participant. One of the fundamental questions, as we transition into the second evolution in this third wave of relational design, is, what is the role of the "other" and how do the design researchers and their designs relate to them?

The fundamental paradigm at the time I started this Ph.D. was the idea of designing "with" (Lohmann, 2017) (Anderson, 2017), instead of designing "for", and that the designer was a facilitator of this process (Lohmann, 2017). However, as I was immersing myself in the literature of designing trust and applied ethics in the context of exponential technological development, it became evident that this model was limited in scope for these contexts/systems. Citizens with no understanding of how these systems operate have a reduced capacity to contribute meaningfully to the ethical development of these systems. And political processes are too slow to deal with the exponential nature of the time we are facing at this moment. The average time for the legislative process in the UK is 2 years. High-level research by Sheila Jasanoff (2016), was instead pointing to technologists and designers/developers as the fundamental enablers to address the rising concerns of these systems.

These insights led me to collaborate with these actors. Consequently, my deontological position and that of the participants, who were fundamentally designers/technologists/ entrepreneurs, has been crucial for the integration of applied ethics and emancipatory directionalities to collective activities. In this context, the cultural diversity of the student body at the RCA in terms of background, programmes of study, nationalities, and their diverse, critical, and enabling capabilities, plus their unifying element as designers, provided an ideal group of participants to develop the task at hand.

As a result, my work repositions the role of the designer from facilitator to expert, and his/her practice from consumption to care. In this context, we move from designing "with" the user to designing "on behalf of" the user. This repositioning gives us significant power and responsibility and demands an ethical and deontological perspective and education to enable this process effectively. Therefore, as we are moving into this second evolution of the third wave of design, as I have argued several times in this Ph.D., design education needs to place ethics at the centre of everything it does – to develop ethical frameworks to address the main task of design in the digital and exponentially technological age within which we live: preparedness, readiness, and appropriateness. However, when working with designers in co-design workshops, I identified that the temporal nature of the inquiry and the reasoning perspective affected the output tremendously. What my work demonstrates is that when a prospective approach is integrated, and ethics is placed at the beginning, the design (which is always a projection of the analysis), evolves in the right direction. What, with whom, and how you analyse something has a direct effect on what you project, and how you project it. This insight enabled me to develop the framework appropriately.

4.6 PROPOSED METHODOLOGY

In this section I will outline the stages leading towards its development and illustrate the actual method, introducing its different levels of complexity. I will be using my inquiry into VAs as a case study to explain the process to build the prospective methodology to design trust in HAuS.

4.6.1 Trajectories

I identified my matter of concern by conducting a range of time-based literature reviews. I investigated this in a multi-dimensional way, i.e. by cross-relational and comparative analyses. In this way, I entered into a systemic dialogue with the object and its relationships, trying to elicit its consequential implications. Coming to VAs with an ethical design background in the ethics of accessing biometric data, shaped by my studies at Goldsmiths and my interest in prospective interactions, I primarily thought of its potential evolutions, its consequences, and social implications. When I discussed this with other designers and researchers, we engaged in dialogues about how the object/system was understood, valued, and used in domestic settings. This helped me map the current state of the art, which mainly revolves around superficial domestic interactions. Based on these insights, I began to reflect on and speculate about the potential interactions of VAs beyond this context.

This process enables a proactive and contextual analytical approach to identify the object of inquiry departing from reactive practices around questions such as “what if...?” visions, trends, drivers, and signals. In this process, the object of inquiry emerges and dictates the trajectory, rather than being imposed. The elements that are chosen to conduct a comparative and relational analysis have a substantial impact in relation to the development and the final output, as this analysis builds the foundation and sets a trajectory for the development of the project. This process is experimental, and careful attention must be paid to the selection of elements. The relationship presented in this research identified three elements; technology, philosophy/sociology, and design practice. The first variable gave me an understanding of what is out there, how it has evolved, and where are we now. The second variable, philosophy, provided me with an intellectual and sociological framework to understand why. And the third variable, design, gave me a framework to figure out how I could approach it, and whether the existing models were capable of addressing it. The second timeline provided an ontological understanding of the technology. This process underpinned the relevance of AI breakthroughs in VAs developments.

In principle, technology, philosophy/sociology and design practice can all be transferred to any other technological development to allow researchers to identify trajectories. However, we may find other elements that could enhance or complement the enquiry. In these cases, the main challenge to overcome will be the capacity to identify an equivalent for any or all of the elements. This aspect opens up a space for further investigation.

4.6.2 Probabilistic extrapolations

As we are projecting the interaction into the future, questions of evidence regarding the prospective development and impact of emerging technology from a research perspective were raised by my supervisors. In this context, due to the limited access to emerging technologies available to researchers, I explored what kind of existing elements in the technological sphere I could use to build a triangulation that would identify potential developments. In this context three elements were identified:

- Demos: Demos are introduced by tech companies to illustrate the potentialities of new technologies. They can be used by researchers to understand the potential development of emerging technologies.
- Prototypes; Prototypes also present a case for potential technological developments. Prototypes may raise ethical questions and illustrate how technology may impact our lives, either positively or negatively.
- Patents; Patents illustrate a potential concrete development of a given technology.

These elements enabled me to map and triangulate potential technological developments and get a sense of their potential impact. This triangulation can be transferred to any other technological development to allow researchers to prospect for potential positive and negative interactions. This process presents a significant departure from Speculative Design, which does not include a factual triangulation to map potential development. In principle, demos, prototypes, and patents operate in all potential technological developments. However, we may find another context or technology without them, or without some of them. In these cases, the main challenge to overcome will be the capacity to identify an equivalent for any or all of the elements. This aspect opens up a space for further investigation. Also, the use of these elements limits the research to near-future contexts.

Finally, much reflection was implemented to find the right terminology for the method to be understood. According to the Oxford Dictionary, “understanding” is defined as “the particular way in which somebody understands something”. In this context, I decided to name it probabilistic extrapolations. Special attention was placed on enabling the term to be perfectly understood in the technological sphere, as this field is, and will be, developing these potential interactions. This field revolves around statistics; therefore, the signifier of probabilities fits perfectly the context. Furthermore, this is a concept easily understood by anybody, regardless of whether they come from the sciences or humanities. Language is fundamental to enhance adoption and transferability in the most impactful context.

4.6.3 Asymmetries

Asymmetries represent a fundamental addition to design as they allow us to identify where the problems are going to be. This process aims to uncover potential areas of conflict, exploitation, and injustice, which may have a tremendous impact on society.

This process is important because if you identify the asymmetries within the system, your design, which is a projection of your analysis, will be transformational and beneficial for society. This method goes beyond current models based on identifying problems or opportunities.

The identification of problems does not imply *per se* that this particular directionality is beneficial for society. It may happen, or it may not. This process provides a more focused and technology-driven method than Causal Layered Analysis. Systems analysis and understanding demands that you need to identify the dynamics of the systems, the actors, and the weigh of those actors within the system. You need to identify the agent/s and actions of trouble to identify and ethically weight your intervention's directionality.

I believe that this process can be employed as a useful active strategy for practice-based research, especially where dematerialised social interactions revolving around trust are concerned. This aspect opens up challenging spaces which demand a thorough analysis. The integration of case studies enabled a grounded process to address these interpretations, which enhanced design opportunities.

4.6.4 Consequences

This process aims to integrate ethical analysis into the development of new products and services. Ethics focuses on how a person should behave. It is a philosophy applicable to daily life or existence. It integrates two areas in order to determine rules or codes of conduct: philosophy, the art of asking questions, and morality, what is good or bad. Its main objective is to determine the right thing to do. Its ontology is based on creating social constructs for the optimum functioning of society. Its epistemology decodes these constructs while its output aims to set standards of behaviour for daily life. This process has been structured in three levels of consequential analysis addressing unintended consequences, contexts, and unintended actions.

In this area, I have used workshops to map, analyse and weigh the potential consequences of a given technological development. Building on Jasanoff's analysis, designers have been identified as the ideal partners/participants to map, develop, and implement these interventions.

The consequential analysis complements the research I conducted through structured levels of analysis. The levels have the potential to sharpen the analysis, as well as expand it, and even change its overall trajectory.

4.6.5 Counter-fictions

Up to this point in the process, I have been undertaking primary and secondary research, and I am now starting to envision future developments of VAs interactions on trust. Here, I am predominantly studying and experimenting with the system as a design researcher interested in the potential ethical uses of design. In this process, rather than conducting an individual, prophetic, value-based judgement about what a desired future might consist of, I am mapping and testing unthinkable, unprovable, and undesired futures against potential technological developments via workshops and co-design activities. This is illustrated in my diagram in which a section cut of the 'Cone of Futures' (Bezold and Hancock, 1994) is transformed into a matrix. In this process, the possible/plausible/probable/preferable is substituted by desired/undesired intended/unintended consequences.

On the cone proposed by Voros (2001), a multiplicity of futures extends from its starting point along a timeline, showing the scope of possible and probable futures. In my model, however, we infer that future by triangulating demos, prototypes, and patents. Once the trajectory is defined, we identify the asymmetries within the system and conduct a consequential analysis. Then we flip the projections (represented as a cone) to reverse the asymmetries by implementing counter-fictions. From this point, I am engaging with the design of a system/mechanism(s) that could reverse the asymmetries of the system to generate an emancipatory project. The challenge is to transform a negative potentiality into a preventive/protective applicable outcome: to ***counter*** the fiction into a real-world intervention (control, repression, and dependencies focus the intervention). In this context, the fiction becomes an object of inquiry, rather than an end. Here, I use them to create tools rather than to generate debates.

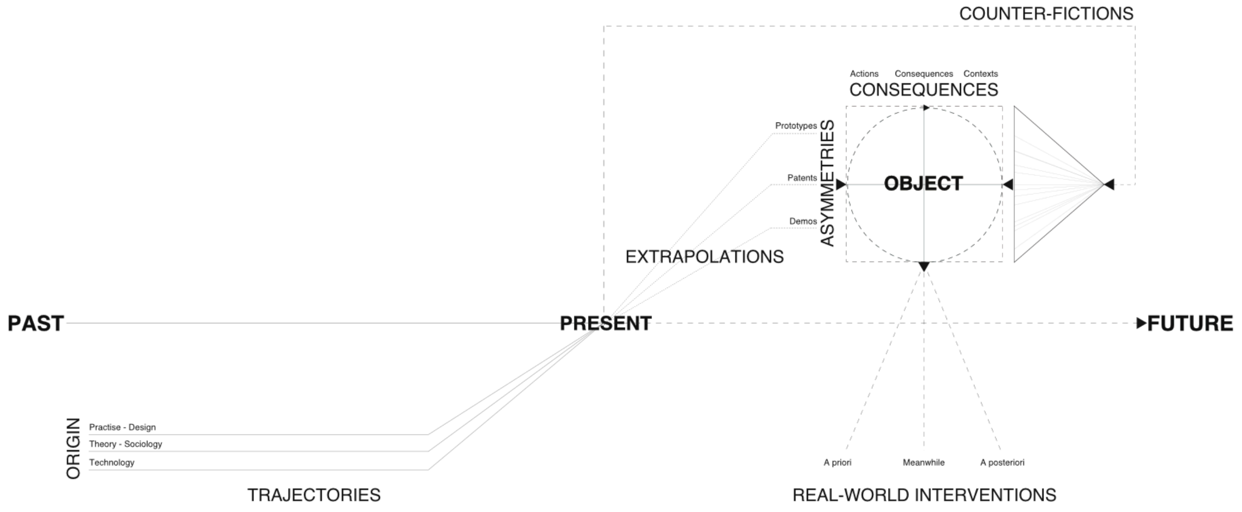
Although diagrams have been used throughout the stages, this is the space where they become fundamental design-practice embodiment tools to operate dematerialised interactions. Finally, to generate reliable feedback loops for my inquiry, I decided to operate this space both as a designer and as a presenter/disseminator. In this context, conferences and professional bodies were targeted to test the proposed strategies/mechanisms. This dual vision enabled me to reflect on my actions and act on my decisions.

This is where the method for practice-based co-design takes shape. Working with designers enabled me to map, understand and address potential areas of conflict. These "communities of practice" (Lohmann, 2017) aim to cancel out blind spots, as described by Rittel (1969) and Lohmann (2017). In this case, they relied on designers and their diverse, critical and enabling capabilities, rather than "any public". Building from research in the area (see Jasanoff, 2016), this process aims to select participants that could affect the development of the project beyond debate.

From the data collected, this structured analysis performed better when the co-design activity was implemented in a prospective context. Analysing technologies as they are currently configured only replicated current arguments, which prevented the process from going beyond what exists to propose that things could be otherwise.

4.6.6 Interventions

In this context, the use of counter-fictional strategies emerged as a strategy with which to address the dynamics of the system, but also as an experimental method to ground speculations. This process led the embodiment and final typology. Interventions can be placed a priori (before the interaction), meanwhile (during the interaction) or a posteriori (after the interaction). This *a posteriori* positionality represents an innovation in trust design. Prior to this Ph.D., interventions were implemented around simulation (before the interaction) or monitoring (during the interaction). By positioning the inquiry at the intersection of society and HASs, I was able to consider *a posteriori* interventions. This space generated the most relevant work in theory and practice.



1	TRAJECTORIES	DEFINE TRAJECTORIES Timelines - Designing Cross-Dimensional literature review + Comparative studies - From Origin to now *Archives*
2	PROB. EXTRAPOLATIONS	ANALYSE PROSPECTIVE DEVELOPMENTS Demos+Patent+Prototypes *Desk Research*
3	ASYMMETRIES	DEFINE AND ASSESS ASYMMETRIES IN THE RELATIONSHIP SYSTEM/OBJECT AND USER Data+Inferences+Dependencies *Case studies*
4	CONSEQUENCES	SYSTEMATICALLY ANALYSE CONSEQUENCES AND IMPACT Contexts > Unintended consequences = Unintended Actions *Workshops > Matrixes + Surveys > impact*
5	COUNTER-FICTIONS	DESIGN POTENTIAL INTERVENTION TO REVERT ASYMMETRIES Countering Control/Repression/Dependencies *Co-Design > Cones*
6	REAL-WORLD INTERVENTIONS	DESIGN INTERVENTION, TEMPORAL POSITION AND TYPOLOGY A priori/Meanwhile/A posteriori *Design embodiment*

Fig. 44. This diagram presents the final embodiment of the proposed methodology. It contains the methods, approach, variables to address, processes, and research techniques used.

4.7 FUTURING

Prospective Design aims to “**affect**” change, rather than “influencing” or “criticizing” it. It differs from other forms of future design studies. For instance, in *The department of seaweed* (2017), Julia Lohmann positions Co-Speculation (CoS) beyond Critical and Speculative Design (CSD). Building on John Wood’s Meta-design, her process is based on generating grassroots local activism to *influence* policy. I find this notion of *influencing* interesting and evolutive in relation to CSD’s provocations but limited in scope. When you "affect" something, it means that you have made it change. Conversely, when you "influence" something, it means that you have altered its behaviour, but not necessarily changed it. Influence is personal and emotional, whereas affect is systematic and relational. This perspective implies moving the process towards a **systematic process of ideation,** rather than a conceptual (Dunne) or materialistic (Lohmann) process of ideation. It aligns more with Transition Design (Irwin).

As we are placing the intervention in the context of dematerialised interactions, the output cannot be observed or graspable but can be **dissolved**. If CSD and CoS deal with materialism from a conceptual and experiential perspective, Prospective Design approaches the design process from a **consequential** perspective to insert an ethical directionality. In terms of participation, Prospective Design (PrD) also repositions Lohmann’s focus on ‘involving the user’, Dunne’s focus on ‘directing the user’, and Irwin’s focus on ‘connecting the user’, to outputs focused on designing **‘on behalf of the user’**. In the process, PrD aims to design **trust**, rather than engagement or comprehension.

In this process, PrD repositions the role of the designer from that of an author (Dunne) or facilitator (Lohmann; Irwin) to that of an **expert** in prospective future-led technological potentialities aimed at mitigating unintended consequences and reducing risks. The main intention of this approach is to protect users. It aims to **shape frameworks** rather than challenge them (Dunne), reframe them (Irwin), or provide a method to deal with them (Lohmann). The success of the output will be determined by the **potential** to **affect** change, as the decision to affect it does not rely on the designer but on somebody else. This position departs from grassroots activism (Irwin) that aims for a bottom-up process. Instead, PrD positions change in a relational context where this ‘*other*’ becomes capital. This process demands the identification of the actors involved in the system, and the *weight* of those actors within the system, (because it is this aspect that determines who is capable of enabling change). Finally, the output should be embodied in the appropriate typology. The agent/s of change need to be identified, and the output translated in a typology that they understand.

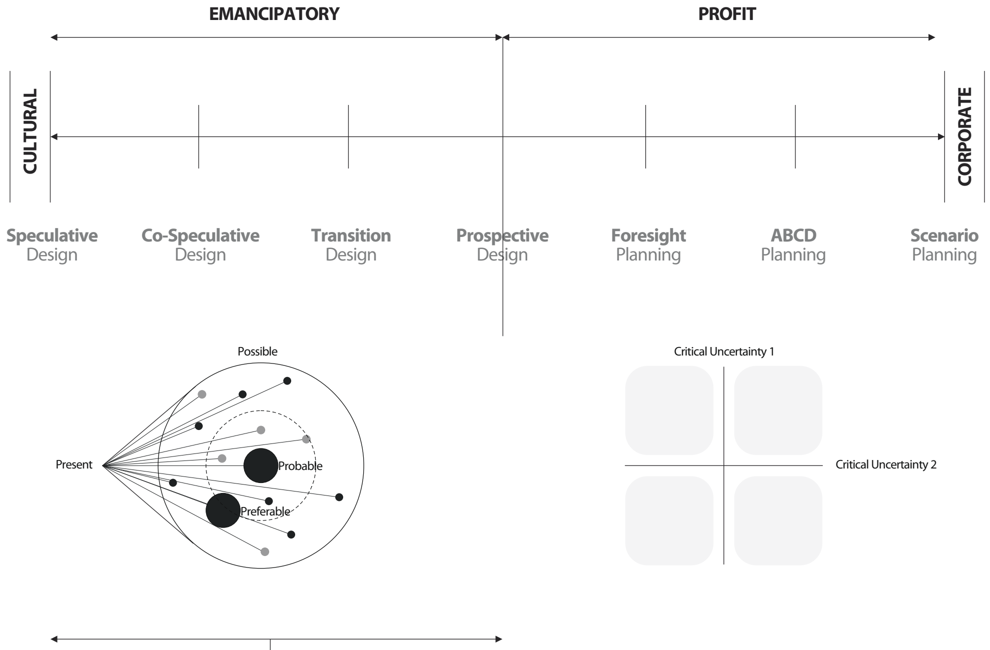
In this Ph.D., for instance, I submitted my research to the National Data Strategy board in the UK to affect the development of AI. All submissions were accepted by the board and included in an evidence bank. I have also proposed a new digital right

(Galdon, 2020a), which has been submitted to the EU Commission for their consideration (this typology is the kind of embodiment that politicians fully understand). I also proposed Synthetic Consequential Reasoning as a system to articulate synthetic morality, which aims to operate ethically from within the system instead of from outside as the political Right does. This aspect was debated with a top designer from one of the big four technological companies in the digital landscape. This was very relevant for him, as they have a Virtual Assistant in their catalogue. Whether this company, the National Data Strategy board or the EU Commission decides to implement these strategies is beyond my control. My duty as a PrD researcher is to prospect the future to propose that things can be otherwise by providing guiding knowledge for transforming the future in an applied and ethical manner.

With PrD, I have investigated ways of designing trust in the context of digital systems, black-box technologies, uncertainty, unpredictability, and automated behaviour, based on exponential technological developments. The framework I have presented in this research provides a focused and systematic approach to ways of addressing these issues. It presents a model that is significantly more substantial and reliable than Humanness Design (HD) or Transparent Design (TrD), which rely on anthropomorphism/deception or explainability/predictability.

In the process, PrD questions models in design futures such as Speculative Design, Foresight Planning, ABCD Planning or Scenario Planning, which rely on reactive practices around "what if …?" questions, visions, trends, signs or drivers, rather than grounded projections supported by background research to justify, focus and guide the projection. PrD extends recent models such as Transitional Design or Co-speculation by identifying key attributes in systems dynamics such as probabilistic extrapolations elements (demos, prototypes, and patents), asymmetric elements (data, inferences, and dependencies), and consequential elements (contexts + unintended consequences = unintended actions), and focuses the intervention by countering control, repression, and/or dependencies. In the process, it changes orthodoxies of participation and design operationality. Finally, PrD can access the future via probabilistic knowledge. This aspect allows this practice to operate in the future.

PrD engages with design processes that might not result in immediate interventions, and with designers looking at these systems to build and implement ethical and emancipatory projects from the short to the long term. This approach moves design's temporal frame towards the future and shifts the sharing of knowledge from the "known"' to the "partially know", from the "factual" to the "potential" and from the "intended" to the "unintended"'. In this context, design research becomes an orthogonal node for grounded transformational directionality and emancipatory led practices, leading to a space for affecting change (Table 12).



	SPECULATIVE DESIGN	CoSPECULATIVE DESIGN	TRANSITION DESIGN	PROSPECTIVE DESIGN
FUNCTION	PROVOKING	INFLUENCING	AMPLYFING	EFFECTING
DESIGNING	ENGAGEMENT	COMPREHENSION	ACTION	TRUST
PROCESS	CONCEPTUAL	MATERIAL	SYSTEMIC	SYSTEMIC
	(PROCESS OF IDEATION)	(PROCESS OF IDEATION)	(PROCESS OF IDEATION)	(PROCESS OF IDEATION)
REASONING	THOUGHT EXPERIMENT	EXPERIENTIAL	SOCIAL	CONSEQUENTIAL
SYSTEMS	CRITIQUE	VALUES	VALUES	ASYMMETRIES
DESIGN IS ...	OBSERVED	GRASPABLE	ENACTED	DISSOLVED
OUTPUT	DIRECTIONAL	ITERATIVE	CONNECTIVE	PROTECTIVE
	(DIRECTING THE USER)	(INVOLVING THE USER)	(CONNECTING THE USER)	(ON BEHALF OF THE USER)
RESULT	CHALLENGING	DEALING WITH	RE-FRAMING	SHAPING
	(FRAMEWORKS)	(FRAMEWORKS)	(FRAMEWORKS)	(FRAMEWORKS)
ROLE	AUTHOR	FACILITATOR	FACILITATOR	EXPERT

Table 12. This table presents a comparative analysis between future design methodologies in the emancipatory area. It contrasts PrD with CoS, TD and SD to illustrate the differences between them.

4.8 PROCESS AND LIMITATIONS

The prospective and probabilistic ontological nature of the knowledge generated presented a preliminary limitation for this research. In order to address this conundrum, I have argued for a repositioning of the origin of design research within an Aristotelian rationale of productive knowledge. This positioning implies that design research has no end, as it is always implicated and will remain in exchange. This exchange always redefines the subjects involved by effecting a shift in power and status through its transformational nature. It cannot transcend time, like mathematics, and depends on time, contexts, and circumstances. Therefore, it assumes past, present and future timeframes and the impact of a changing environment, and future social and economic factors. It is instrumental and situated, and its value is social, economic, and environmental.

In this context, building from an Aristotelian perspective of productive knowledge, design research is concerned with establishing competing standards of value rather than securing boundaries of knowledge, and its practice is based on the capacity to make new futures involving abductive reasoning. It is concerned with something coming into being indicating that things can be otherwise and beyond themselves, as currently configured. It is concerned with the indeterminate and the possible within alternative possibilities: from passive intellect (contemplation becoming its object) to active intellect (an object being defined) to prospective intellect (an object being transformational by exchange) (Galdon, 2019g).

In the prospective framework, I have proposed that design research can access the future. However, current models of research are limited by the present, either by observation or measurement. In order to address this fundamental aspect, I have presented the concept of probabilistic knowledge by building on new approaches in design and economics. Probabilistic knowledge in the context of design could be defined as the potential impact of transformational initiatives. The value of design research as presented here is social, therefore aiming for mixed methodologies to implement strategies building informed interventions to support planning, solution-based problem-solving, problem-shaping, synthesis, preparedness, and appropriateness in the built environment. These aspects are fundamental for the optimum development of society in an ever-evolving world based on exponential technological developments. This approach has so far been inaccessible due to the present limited frameworks of sociology and science that can only analyse what already exists. In this process, I propose that I am making a contribution to knowledge by contextualising Glanville's concept of knowledge for transforming the future as a probabilistic knowledge ontology (Glanville, 2005).

The traditional paradigm positions design as a method within research, which creates tensions that arise between the prospective nature of design and the factual requirements of working in the present. There is an ontological problem between the nature of design as future led and prospective and the nature of research which is present based and factual. I argue that the core nature of design is probabilistic research, not empirically-driven research. We trade a degree of accuracy for access to areas yet-to-be or not-fully-formed. Therefore, our output is probabilistic, and research is always preliminary in its nature. Moreover, in exchange, we provide guiding knowledge for prospective technological developments: "knowledge for" instead of "knowledge of". We are concerned with how things "ought to be"(Simon, 1996, pp.111-167) instead of how things are. These elements reposition design research as directional and transformational.

In this scenario, as the life of the intervention is placed into the future, time to assess the impact of the design is extended during its lifetime. Validation is always a posteriori, and the proposed output becomes the main element to be assessed. The validity of the output generated, whether in a commercial or research context, will be judged by its ethical appropriateness and the potential transformational impact.

In this context, the implementation of a progressive and cross-disciplinary publishing strategy allowed me to mitigate assumptions in the process by contextualising and confirming my outputs progressively. This strategy enabled me to build robustness in the context of abductive research in a design research context. This framework allows the researcher to go beyond what exists and investigate the potentialities emerging from technological developments (Galdon, 2021a). As a result, I have published twelve papers in a wide range of fields, from Industry 4.0, Human Factors and design research to applied science and Design Futures. This approach to practice aims to enhance the impact of the research in terms of outputs and scrutiny by diverse audiences to maximise its transversality, and therefore its robustness.

4.9 CONTRIBUTIONS

The development of trust design aims to contribute significantly to the safe and ethical development of technological solutions by understanding;

- The actions taken to ensure that an item, system, system of systems or network is free from adverse impacts by considering threats at the early stages of developing a new product.
- The conceptualisation, development, and implementation of Prospective Design as a framework to address preventive design.
- The conceptualisation, development, and implementation of a probabilistic knowledge ontology to enable prospective and preventive design to operate in the context of research.
- The conceptualisation, development, and implementation of design consequentialism as a method to address unintended consequences in order to build ethics in decision making in the context of exponential technology developments.
- The actions taken to enable researchers, designers, and developers to go beyond what exists and ask the kind of questions that would allow them to mitigate potential unintended consequences through applied ethics in design.

4.9.1 Enabling design research to operate in the future

From a historical perspective on design research, Archer proposed design as a third culture, distinguishing it from the sciences and humanities (Archer, 1978), and Jones, Glanville, and Auger suggested the prospective nature of design by introducing the future as a place for inquiry (Jones, 1970; Glanville, 2005; Auger, 2012). However, none of them have resolved the implications of accessing these areas and the distinctiveness of design in terms of knowledge output. In their cases, they redirected the output to established models. In my model, I have enabled design research practice to operate the future in keeping its true ontological nature by bringing back Aristotle's productive knowledge and connecting it to probabilistic knowledge (Galdon, 2019g). This thesis advances knowledge by making a fundamental contribution to the contextualisation of Glanville's concept of *"knowledge for"*transforming the future as a probabilistic knowledge ontology.

4.9.2 Towards Relational Design 2.0

In his article "Towards relational design", Andrew Blauvelt (2008) proposed that we are moving towards a relationally based, contextually specific design. In his account, he structures the evolution of design into three main epochs: Modern design, from 1900-1950, focused on form, disseminated rationally and potentially universally. Post-modernist design, ranging from 1960-2000, focused on design's meaning-making potential, symbolic value, semantic dimension, and narrative potential. And relational design, ranging from 2000 to the present, focuses on effects on users, pragmatic and programmatic constraints, rhetorical impact, and the ability to facilitate social interactions.

In this context, Blauvelt presents IDEO and Dunne and Raby as primary practitioners in this new evolution. In his account, he describes relational design as including performative, pragmatic, programmatic, process-oriented, open-ended, experiential and participatory elements, suggesting that it moves away from designing discrete objects "to the creation of systems and more open-ended frameworks for engagement: designs for making designs" (Blauvelt, 2008). He presents the Roomba as a later embodiment of his proposition.

I would agree with the key idea of a third major wave in design that focuses on designing relationships. However, what Blauvelt missed in his account, as this thesis demonstrates, is that the nature of the system of interaction demands a different kind of design and timely intervention. In reactive systems you design engagement, and you can be reactive because you have control. The developer hard-codes all the possible interactions. In proactive systems, such as highly automated virtual assistants, you design trust because you are designing a set of rules in systems that are unsupervised and that keep evolving. In this context, you need to prospect potential interactions to envision unintended consequences. In a sense, we could characterise this design era as Relational Design 2.0.

If the first wave of design offered us a multiplicity of forms, and the second a multiplicity of meanings and interpretations, the first part of the third wave presented a multiplicity of contingent, boundaries and/or conditional solutions: open-ended rather than closed systems; real-world constraints and contexts over idealised utopias; relational connections instead of reflexive imbrication; the end of discrete objects, hermetic meanings and the beginning of connected ecologies".

In this scenario, the second part of this wave presents trust as a fundamental element to design: unsupervised versus supervised systems; unintended consequences versus control; prospective versus reactive; emancipation versus manipulation; uncertainty versus transparency; not-fully-knowing versus knowing; reparation and accountability versus impunity, and the ubiquity of fluid cyber-blended and hyper-connected ecologies. In this context, this thesis advances knowledge by making a fundamental contribution to contextualising this second part of the third wave of design in what I characterise as **the consequential turn**; The transition from conceptual to pragmatic future interactions.

4.9.3 Prospective Design beyond design

For a cross-disciplinary research project like this Ph.D., it is common for the researcher to fulfil a multiplicity of roles in the research process. This implication demanded acknowledgment of the position I needed to occupy in different parts of the inquiry: not only as a researcher, or designer, but increasingly as an ethicist, theorist, or humanist. This progression of my position into the field of applied ethics by collaborating with other researchers enabled me to complement my own knowledge and skills acquired during my studies at Goldsmiths at the intersection of science, technology, design, and ethics. Here, I argue that the exposure to ideas between disciplines "pollinates" participants, prompting them to make contributions outside their "own" field. This led me to publish six papers in non-design-related fields to test the validity of the propositions. Which in turn, enhanced robustness in the process. However, the prospective and probabilistic ontological nature of the knowledge generated in this cross-pollination remained a fundamental issue in enhancing cross-domain collaboration. This led to publishing a specific paper on the ontological nature of design.

In addition to the contributions to design outlined above, this investigation has also made contributions to other fields. In the field of Human Factors, it has contributed to a scale of levels of autonomy in Highly Automated VAs (see Galdon, 2019a). It also proposes the first comprehensive scale of levels of reparation for HAS (see Galdon, 2019c). It expands the emerging area of apology by including a gradation area integrating compensation (see Galdon, 2019b). This is a radical new addition to address unexpected outputs in HASs operating in highly sensitive areas such as health and wellbeing, social interaction, economy-related activities, and identity. This research concluded with the creation of a scale of levels of accountability (see Galdon, 2019b). These propositions advance knowledge by making a contribution to contextualising trust design at the intersection of AI, ethics, and society.

This research contributes to the current debate in the field of human factors between the simplification of levels of automation (Kaber, 2018) and its relationship to contexts and actions (Bradshaw, Hoffman, Johnson, and Woods, 2013). As a result, the papers published demonstrate that both are right. A multi-scale system can account for a wide range of possible scenarios in extended lifecycles. However, contexts and actions determine the most appropriate levels (see Galdon, 2019d).

In this process, this research has identified relevant highly sensitive areas where trust is important to users: health and wellbeing, identity, social interactions, and economic activities – and unexpected reactions to interactions such as unhappiness, inaccurate predictions, the loss of something, or violent endings (see Galdon, 2019a). It, therefore, provides relevant insights for potential developments in the area, fulfilling in the process Glanville's concept of knowledge "for transforming the future" as a probabilistic knowledge ontology (Glanville, 2005).

In addition, this Ph.D. makes contributions to ethical computing. In this area, building on the outlined contribution to the Human Factors community, this thesis proposes a multi-dimensional scalar system integrating post-interaction elements such as accountability and reparation, as well as unintended actions, contexts, access, and inferences as fundamental variables to address Synthetic Consequential Reasoning. This area aims to facilitate the design of ethical systems by inserting a sense of consequence as part of computational reasoning, thus contributing to the emerging area of synthetic morality (see Galdon, 2020c).

Building on this research, a form of calculation was created to facilitate the translation of ethical and philosophical concepts into computational reasoning. This method could be used to optimise and calibrate a system's decision-making process (see Galdon, 2020c). Furthermore, this process was embodied in a tool (calculator), that aims to facilitate the design of trusted systems (see Galdon 2020b). This tool represents an addition to the emerging area of Ethical Tech.

Finally, building from the research conducted in human factors, this thesis argues that a new digital right, the "right to reparation" (see Galdon, 2020a), is needed to address the accountability gap presented by highly automated complex systems incapable of thoroughly monitoring its actions in real-time (Kohli, 2019). The right to reparation follows the articulation of the "right to be forgotten" (Weber, 2011), the "right of explanation" (EU, 2019a) or more recently the "right to reasonable inference" (Wachter, 2018), and aims to ensure that emerging HAS interactions remain accountable while the development of highly automated technologies cannot fully guarantee their behaviour. The right to reparation was presented and published in the proceedings of IHIET'20 at CHUV Lausanne (Galdon, 2020a). This contribution advances knowledge by making a contribution to contextualising trust design at the intersection of design, law studies, and AI.

Ph.D

CONCLUSIONS

Humans have been reflecting on technological developments for millennia. We can find early discussion of these in the *Protagoras* of Plato, and, in a more contemporary context, in the writing of Walter Benjamin and the philosophers of the Frankfurt school. Like Jasanoff, they saw technological developments as creatures that are not neutral, but which are instruments of progress.

As we are transitioning from the predictability to the prescriptibility of behaviours, an urgent need emerges to control this type of technological development, due to its unlimited potential. We need to understand, as Lassalle suggests, that the Faustian expansion of these developments is vocationally expansive and abrasive. The entire algorithmic culture that we are generating is producing anthropological changes in our personal and collective unconscious. Algorithms are challenging the Kantian discourse of "*coming of age*" and making our freedoms more like assisted freedoms than real freedoms. We need to articulate and implement ethical limits, otherwise, these technological developments may become tools to enforce power. This, I would say, is the greatest challenge that we have ahead of us. And it is in this context that design has a fundamental role to play.

In this context, the paradigm of Highly Automated Systems (HAS) enabled me to explore more evolved automated systems than those that are currently configured, while avoiding a fully autonomous General Artificial Intelligence (GAI) interpretation, which would have drawn the enquiry into more philosophical debates around consciousness and speculative perspectives. This characterisation grounded the research while helping me to address notions of "towards", therefore opening a critical space for investigating VAs in the context of future evolutions. This space enabled Prospective Design to emerge.

The proposal of Prospective Design (PrD), in answering my hypothesis has explored how trust could be designed in the context of Highly Automated Virtual Assistants. In this process, preliminary studies provided early insights, gained by investigating trust design in the context of AI, VAs, and news media. These studies enabled me to identify the crucial prospective approach needed to address exponential systems that are continuously evolving. Understanding this fundamental idea has been crucial for the success of this research. This baseline was structured around a set of methods, which emerged through reflection and preliminary insights and embodiments, and were consolidated by continuous peer-review evaluations.

Trajectories were very important because they inserted a proactive method to implement high-order systems analysis via relational and comparative perspectives. This process allows a literature review and background research to enlighten prospective development by providing a space for a case to emerge. This is the process that allowed me to identify the VA as an object of inquiry: its characteristics, qualities, and the need for a different type of design methodology to address its nature.

Probabilistic extrapolations were extremely important to address notions of prospectivity in the context of research. No other method existed at the time in future design studies that would allow a design researcher to access the future and triangulate a potential development from a factual perspective. This method inserted rigour and robustness into the process.

Identifying and reversing asymmetries is the most effective methodological contribution I have made because it inserts an ethical directionality to product/system development. This process aims to uncover potential areas of conflict, exploitation, and injustice. This is extremely important, because if the asymmetries are identified within the system, the resulting design, which is always a projection of the analysis, will be transformational and beneficial for society.

Consequences were an important method because they provided a systematic model to address asymmetries. It identified the three fundamental levels where they are relevant; consequences (intended and unintended), contexts (highly sensitive areas: health, economy, identity, and social interactions), and unintended actions (unhappy actions, inaccurate predictions, the loss of something, and violence). This triangulation has the flexibility to address a multiplicity of contexts, cultures, and behaviours.

Counter-fictions was a radical addition because it challenged established orthodoxies on design futures, which are confronted by a positivist pro-consumer option versus a critical pro-citizen option. Instead, this method allows us to integrate both options into an emancipatory projection revolving around applied ethics. This perspective enables the designer to develop real-world interventions with an ethical and social component at their heart. It also challenges the idea that design futures need to be either ‘bright’ and utopic or ‘noir’ and uncanny. My project is embodied in a functional calculator.

This research has archived its intentions of building a methodology to design prospectively trust in the context of AI. The comparative study between the proposed methodology and the European Commission (EC) released White Paper on AI and the acceptance of several publications by the National Data Strategy board probes that I have been able to identify the key elements and design tools and frameworks to address the rising concerns and effect change. In the process, trust has become a ubiquitous force in this Ph.D. impregnating every aspect of the process. How do you design trust in the area of levels of control in AI? by designing reparation and accountability [systems]. Articulated in this Ph.D. in the form of a multi-scalar system, and embodied in a calculator. How do you design trust in design futures? by designing integrity [systems]. Articulated in this Ph.D. in the form of methods; trajectories inserted background research, probabilistic extrapolations inserted a factual model to triangulate potentialities, asymmetries focused the intervention, consequences inserted an ethical analysis to underpin unintended outcomes, and counter-fictions a method to reverse and ground them into real-world interventions. Finally, how do you design trust in design research? by designing robustness [systems]. Articulated in this Ph.D. in the form of a progressive cross-dimensional publishing process. In terms of what is left or uncertain from the proposed Prospective Design methodology, tailored workshops have tested key specific aspects of the methodology. Further research is needed to test the full extension of the methodology proposed. Another element that came from the workshops was the numerical nature of the calculator. In this context, some participants pointed to a different kind of output that could be more sympathetic to narrative. Therefore, the embodiment of language to reach different audiences opens up a space for further investigation.

CHALLENGES

Investigating the prospective developments of this type of interaction device revealed the particular challenges of highly automated interactions for scholarly research. The process of grounding these methods was very challenging, as there was no handbook specifying how to design trust in highly automated VAs. In this process, implementing a cross-disciplinary and progressive confirmation model to remove assumptions *in-the-process* and consolidate knowledge from an external perspective proved crucial. The objects created in the process (tools and frameworks) reflect a dialogue between design, futures, AI, and ethics, culminating in the Trust Calculator as an embodiment of a system/mechanism to facilitate the design of trust in HAuSs. This tool leads to a method of calculation to generate a trust rating: the score can be used to optimise, simulate or calibrate the system's decision-making process, and is based on a multi-dimensional scalar perspective. This approach has emerged as a more reliable strategy to build a systematic mechanism to establish trust in HAuSs as the published comparative study suggests. This innovative mechanism integrates post-interaction elements such as accountability and reparation, as well as unintended actions, contexts, access, and inferences, as fundamental variables to facilitate the design of trust from a consequential perspective on unsupervised highly automated computational systems through their extended lifecycle. In this context, the intersection between the critical issues of automation and accountability acted as a focal point.

In the process, this thesis has challenged and developed current notions in design research, based on technological progress and revolving around product development or speculations to a model based on ethical responsibility which places equal value on the process of design and the impact of the system on society. It positions trust as the main element in addressing the main task of design in the digital and exponentially technological age in which we are living: preparedness, readiness, and appropriateness. These issues demand that we need to go beyond what already exists, and beyond the positive impact of technology to design strategies to address and or mitigate unintended consequences, as they are fundamental for the optimal development of society; To propose that things can be otherwise.

The main design problem I have dealt with in this Ph.D. is the limits of predictability to define balance within the system. This idea was addressed, for instance, by early cyberneticians such as Forrester, with the idea of 'equilibrium'. However, the incapability of fully monitoring AI/ML systems due to their increasing and exponential complexity undermined the fundamental idea of systems returning to their "initial state" and prompted the problematics of designing for systems continuously performing in a "new state". This means that we are dealing with complex dynamic systems that are continuously evolving. This aspect is extremely important because is precisely this complexity that brings the idea of risk to the forefront. The difficulty of fully monitoring and predicting demands the generation of prospective and reparative spaces for inquiry that focus on accountability and reparation as strategies to mitigate unintended consequences to build trust. It is in this context that PrD thrives.

In this investigation, I have considered design's unique relationship to the future and how concepts of anticipation, probabilism, and prospectivity underpin a new understanding of design's relationship to uncertainty and trust. In effect, I have discussed how design cares for the future of transformations in an era where rapidly advancing technologies via exponential technological developments are challenging human-machine interactions. Probabilistic knowledge emerges as an ontological reality to address the intrinsically abductive nature of future design research by building on new approaches in design and economics. Probabilistic knowledge in the context of design could be defined as the potential impact of transformational initiatives. Ultimately this approach implies a different form of knowing and aims to position design research as the discipline best prepared for addressing the future.

PrD engages with design processes that might not result in immediate interventions, and with designers who look at these systems not only as a place to implement long-term emancipatory projects but also to build ethical businesses. This shifts design's temporal frame towards the future and shifts the sharing of knowledge from the "known" to the "partially-known", from the "factual" to the "potential" and from the "intended" to the "unintended". In this context, design research becomes an orthogonal node for grounded transformational directionality and emancipatory forming practice and a space in which to effect change. Prospective designers are not problem-solvers in the traditional design sense; we are solution-seekers for problems that do not fully exist yet, but which can be detrimental for society. Our wondering and experimentation is directional and contains a strong sociological baseline from which to apply techniques ethically to build emancipatory projects in relational contexts.

One of the fundamental debates during my research was the prospective nature of the research. Critics have presented reactive practices as more reliable approaches to designing trust. My research suggests otherwise. It has been able to articulate the right outputs by identifying the key methods to prospect the future, as the comparative study of my outputs against the latest White Paper from the EU Commission demonstrates.

However, it has been during the Covid-19 pandemic that this idea has clearly crystallised with peers. This event has proven how detrimental reactive decisions are in the context of exponential events. This crisis highlights the problems of not having processes and strategies in place. The speed and impact of the virus demanded more agile decisions than the ones which have been taken, and the results have been catastrophic. These are the implications of not prospecting and being able to think before doing. The UK wanted to produce a million vaccines without having a proven model. We are finding that respirators are making patients worse. The fundamental problem here is how difficult it is to deal with 'the NEW' when we have nothing in place. Prospective design cannot provide all the answers, because the future is unpredictable, but it can assist with its methods significantly by identifying the key elements needed to mitigate complex and exponential events and technologies. The knowledge is probabilistic, but this probability can be crucial to infer preliminary knowledge to protect people. This space of inquiry is as legitimate and important as science, sociology, or philosophy, and designers are the best prepared to deal with it because it aligns with our own ontological nature.

LIMITATIONS

The utilisation of patents, demos, and prototypes limits the scope of this methodology to near-future scenarios. These elements provide essential information to triangulate and ground the trajectory of a technological development to prospect potential interactions. Their removal would provide a larger scope, but the practitioner would also lose control in the projection. The output would be more related to speculation than prospection.

Another limit is related to the technology-led approach emerging from the research conducted. In this case, research into alternative models/elements to patents, demos, and prototypes, or prospective impact analysis strategies will be required. Further research into data projections, hybrid data, and other qualitative means of scoping prospective futures may expand the reach of this methodology beyond its current configuration.

FUTURE DIRECTIONS

Continuing research

A fundamental question at this point is: what does this research supports in the ai+trust+ethics landscape, and how do we move forward?

Going back to the fundamental issue addressed by this Ph.D; the problematics of designing for systems continuously performing in a “new state”, the EU framework has identified the main element to design; trust, however, it fails to address it adequately. The framework proposes the establishment and implementation of standards as the main strategy to deal with these systems; however, as we are dealing with complex dynamic systems that are continuously evolving, what we need to create are calibration systems. The problem with standards is that they are static, and need time to develop. What we have seen in recent years is that by the time we complete this process, they are obsolete. The legislative process is very limited in terms of dealing with exponential technological developments. This aspect is extremely important because it is precisely this complexity that brings the idea of risk to the forefront. The increasing difficulty of fully monitoring and predicting large systems demands the generation of a new set of techniques and strategies to monitor and calibrate “evolving systems”, as well as, prospective strategies to mitigate unintended consequences. This Ph.D. provides a basic model in the form of Prospective Design and synthetic consequential reasoning to build upon.

At the same time, we need to complement this strategy with prospective and reparative spaces for inquiry that focus on accountability and reparation as strategies to mitigate unintended consequences in order to build and maintain trust. In the areas of accountability and reparation, cultural contexts will play a fundamental role in establishing the right approach. Research into cultural signifiers and sensitivity, inclusion, and respect for minorities will help to build capacity and accuracy in determining the model which is most adequate.

Recommendation 1 - Set up a new Prospective Design public-private institute focusing on data, algorithms, and infrastructure to combine efforts, and ensure the coordination of research and innovation in future AI developments. This recommendation addresses the lack of preparedness and prospectivity in the policy sector that organisations like the EU currently deploy to address and regulate people's interactions with HASs.

Recommendation 2 - Establish a new calibration centre structured around two main areas; first, a research lab to foster the development of calibration systems/strategies, and second, a new lab to test and monitor the decision-making process of algorithms. This recommendation addresses the lack of understanding about systems that are continuously evolving in the policy sector that organisations like the EC, the UK, the UN, and the U.S. federal government currently deploy to regulate people's interactions with HASs.

Recommendation 3 - Develop an observatory for the integration of contextual cultural analysis to repair trust via participatory social programmes. Existing research currently focuses on preventive strategies. But what happens when the harm is done?

This recommendation addresses the lack of understanding about social contextuality and the impact on minorities in the policy sector that organisations like the EC, the UN, or the U.S. federal government currently deploy to regulate people's interactions with HASs.

Recommendation 4 - Establish a public-private agora to enforce an index for the development and integration of accountability strategies to maintain trust. This recommendation addresses the lack of suitable strategies for systems impact in the policy sector that organisations like the EC, the UN, or the U.S. federal government currently deploy to regulate people's interactions with HASs.

Recommendation 5 - Establish and support an Advanced Ethical Skills Programme via networks of leading universities and higher education institutes to upskill designers and technologists. This recommendation addresses the lack of retroactive ethical programmes about the ethical impact of systems in the policy sector that organisations like the the EC, UK, or the U.S. federal government currently deploy to regulate people's interactions with HASs.

FURTHER DIRECTIONS

Transferability

One question remaining is whether this methodology can be transferred to other domains. Transferability is defined as “The ability to apply the results of research in one context to another similar context. Also, the extent to which a study invites readers to make connections between elements of the study and their own experiences” (Barnes et al., 2020). Based on the work conducted, the number of papers published in different fields, and the key insights outlined, this methodology is potentially transferable to other fields or events of practice-based inquiry in the context of exponential developments revolving around uncertainty and risk, where trust and ethics play a fundamental role. Socio-political events and processes where preparedness is fundamental (e.g pandemics) or fields with outputs operating autonomously, such as synthetic biology, have the potential to benefit from this methodology.

Exology

In this prospective process, a final question is “what kind of research paradigm may enable us to operate this space of not-fully-knowing?” A review of current models positions all the paradigms in the context of knowing: by means of opinion, by means of observation, by means of sensing, or by means of values.

However, the case I am proposing acknowledges that we do-not-fully-know. What, then, is the research paradigm of not-fully-knowing? This issue is also being explored in Object-Oriented-Ontology, in particular in the context of Timothy Morton's 'hyperobjects': he defines elements such as the internet, or nature as "entities with such vast temporal and spatial dimensions that defeat traditional ideas about what a thing is." (Morton, 2017). These elements affect what these objects are, their impact, and how we think, coexist and experience our politics and ethics. In this context, I propose ***Exology*** as a new research paradigm to deal with processes in which it is not possible to fully know by any other means (Exo- means 'outside', 'outer', 'external'). This element adds a new dimension to Ontology (the nature of reality or being), Epistemology (what constitutes acceptable knowledge), Axiology (the role of values), and Doxology (the role of public opinion). In this context, ***Exology*** may be considered as a new dimension to address the role of the unknown and uncertainty.

Fernando Galdon 05/09/2020

Ph.D

REFERENCES

A

- Alvesson, M. and Sko¨ldberg, K. (1994). Tolkning och Reflektion. Vetenskapsfilosofi och Kvalitativ Metod, Studentlitteratur, Lund.
- Anderson, P. et al. (2018). Foresight Review on Design for Safety. London: Lloyds Register Foundation.
- Andreewsky, E. and Bourcier, D. (2000). "Abduction in language interpretation and law making", Kybernetes, Vol. 29 No. 7/8, pp. 836-45.
- Archer, L. B. (1978) Time for a revolution in art and design education. RCA Papers No. 6. London: Royal College of Art, London.
- Aristotle (2000). Book VI. In R. Crisp (ed.), Aristotle: Nicomachean ethics (Cambridge Texts in the History of Philosophy, pp. 103-118). Cambridge: Cambridge University Press. Doi:10.1017/CBO9780511802058.010
- Aristotle (1998) Metaphysics. Hugh Lawson-Tancred (tr.), London: Penguin, 1998. ISBN 0140446192.
- Aristotle (1984) Rhetoric. W. Rhys Roberts (tr.). in The complete works of Aristotle. Vol. II. Jonathon Barnes (ed.). Princeton, NJ: Princeton University Press.
- Arlbjørn, J.S. and Halldorsson, A. (2002). "Logistics knowledge creation: reflections on content, context and processes", International Journal of Physical Distribution & Logistics Management, Vol. 32 No. 1, pp. 22-40.
- Ashby, W. R. (1956). An Introduction to Cybernetics. London: Chapman & Hall Ltd. ISBN 9781614277651.
- Athalye, A., Carlini, N. and Wagner, D. (2018). Obfuscated gradients give a false sense of security: circumventing defenses to adversarial examples. In: Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

- Atoyan, H., Duquet, J. R., and Robert, J. M. (2006) Trust in new decision aid systems. In: Proceedings of the 18th International Conference of the Association Francophone d'Interaction Homme-Machine (pp. 115–122). New York, NY: ACM.
- Atwill, J. (1998) Rhetoric reclaimed: Aristotle and the liberal arts tradition. Ithaca, NY: Cornell University Press.
- Auger, J. (2012) Why robot? Speculative design, the domestication of technology and the considered future. Ph.D thesis, Royal College of Art.
- Awad, E. et al. (2018). The Moral Machine experiment. Nature, 24 October. Doi:10.1038/s41586-018-0637-6

B

- Bachmann, R., & Zaheer, A. (2006) Handbook of trust research. Cheltenham: Edward Elgar Publishing.
- Bachmann, R., & Zaheer, A. (eds.) (2013) Handbook of advances in trust research. Cheltenham: Edward Elgar Publishing.
- Bagheri, N., & Jamieson, G. A. (2004). The impact of context- related reliability on automation failure detection and scanning behaviour. In: 2004 IEEE International Conference on Systems, Man and Cybernetics, 1, 212–217.
- Balog, M., Gaunt, A. L., Brockschmidt, M., Nowozin, S., & Tarlow, D. (2017). DeepCoder: learning how to write programs. In: ICLR 2017. arXiv:1611.01989
- Ball, T., (1977). Plato and Aristotle: the unity versus the autonomy of theory and practice. In: Political theory and praxis: new perspectives. pp. 55-70. Minneapolis: University of Minnesota Press.
- Bansal G. & Zahedi, F. M. (2015). Trust violation and repair: The information privacy per- spective. Decision Support Systems 71 (2015), 62–77
- Barnes, J., Conrad, K., Demont-Heinrich, C., Graziano, M., Kowalski, D., Neufeld, J., Zamora, J., & Palmquist, M. (1994-2020). Generalizability and Transferability. The WAC Clearinghouse. Colorado State University. Available at https://wac.colostate.edu/resources/writing/guides/.
- Bass, E. J., Baumgart, L. A., & Shepley, K. K. (2013). The effect of information analysis automation display content on human judgment performance in noisy environments. Journal of Cognitive Engineering and Decision Making, 7, 49–65.
- Bean, N. H., Rice, S. C., & Keller, M. D. (2011). The effect of gestalt psychology on the system-wide trust strategy in automation. In: Proceedings of the Human Factors and Ergonomics Society 55th Annual Meeting (pp. 1417–1421). Santa Monica, CA: Human Factors and Ergonomics Society.
- Belliot, E. (2018). Counter-fictional design. Critique d'art. Accessed 01/11/2018. Available from: http://jour- nals.openedition.org/critiquedart/19220

- Bezold, C. and Hancock, T. (1994) An overview of the health futures field. Report of an international consultation convened by the World Health Organization, Geneva, July 19-23 1994. Geneva: WHO
- Blauvert, A. (2008) Towards relational design. Design observer 11.03.08. Accessed: http://art.yale.edu/file_columns/0000/0076/blauvelt.pdf
- Blobaum, B. (2014). Trust and journalism in a digital environment. Reuters Institute for the Study of Journalism. Accessed 30/04/2015. Available from: https://reutersinstitute.politics.ox.ac.uk/sites/default/files/Trust%20and%20Journalism%20in%20a%20Digital%20Environment.pdf
- Blobaum, B. (2016). Trust and communication in a digitised world: models and concepts of trust research. Cham: Springer.
- Botsman, R. (2017). Who can you trust? How technology brought us together and why it might drive us apart. London: Penguin.
- Bottom, W. P., Gibson, K., Daniels, S. E., & Murnighan, J. K. (2002). When talk is not cheap: Substantive penance and expressions of intent in rebuilding cooperation. Organisation Science, 13(5), 497-513.
- Boyd Davis, S., Gristwood, S. (2016). The structure of design processes: ideal and reality in Bruce Archer's 1968 doctoral thesis. In: Proceedings of DRS 2016, Design Research Society 50th Anniversary Conference. Brighton, UK, 27–30 June 2016.
- Bradshaw, J. M., Hoffman, R. R., Woods, D. D., & Johnson, M. (2013). The seven deadly myths of autonomous systems. IEEE Intelligent Systems, 28(3), 54–61.
- British Council (2021). Should robots be citizens? Accessed 01/02/2021. Available from: https://www.britishcouncil.org/anyone-anywhere/explore/digital-identities/robots-citizens
- Brooks, F.P. (2010). The design of design: essays from a computer scientist. Hoboken, NJ: Addison-Wesley Professional
- BSI (2016). BS 8611:2016: Robots and robotic devices. Guide to the ethical design and application of robots and robotic systems. London: British standards Institution.
- BSI (2020). Responsible Innovation Guide PAS 440. Available at: https://pages.bsigroup.com/l/35972/2020-03-17/2cgcnc1?utm_source=pardot&utm_medium=email&utm_campaign=SM-STAN-LAU-PAS-PAS440-2003
- Buchanan, R., (1992) Wicked problems in design thinking, Design Issues, 8(2) (Spring), 5-21.
- Buchanan, T. W. (2007). Retrieval of emotional memories. Psychological Bulletin, Vol 133(5), 761- 779.
- Buckley, C. A., & Waring, M. J. (2013). Using diagrams to support the research process: examples from grounded theory. Qualitative Research, 13(2), 148–172. Doi:10.1177/1468794112472280
- Buckley, L., Kaye, S. A., & Pradhan, A. K. (2018). Psychosocial factors associated with intended use of automated vehicles: a simulated driving study. Accident Analysis & Prevention, 115, 202–208. Available from: https://doi.org/10.1016/j.aap.2018.03.021

- Bukhari, S. A. H. (2011). What is comparative study SSRN, November 20. Available from: https://ssrn.com/abstract=1962328 or http://dx.doi.org/10.2139/ssrn.1962328

C

- Campbell,D.T., & Stanley, J.C.(1963).Experimental and quasi-experimental designs for research. Boston: Houghton Mifflin.
- Castaldo, S. (2007). Trust in market relationships. Cheltenham, UK and Northampton, MA: Edward Elgar.
- Chen, D.H.C.; Dahlman, C. J., (2006). The knowledge economy, the KAM methodology and World Bank operations (English). World Bank. Available from: http://documents.worldbank.org/curated/en/695211468153873436/The-knowledge-economy-the-KAM-methodology-and-World-Bank-operations Accessed 20/08/2018.
- Chen, J. Y. C., Procci, K., Boyce, M., Wright, J., & Garcia, A. (2014). Situation awareness-based agent transparency. Defense Technical Information Center. Available from http://www.dtic.mil/docs/citations/ADA600351,
- Chien, S. Y., Lewis, M., Semnani-Azad, Z., & Sycara, K. (2014). An empirical model of cultural factors on trust in automation. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 58, 859–863. Available from: https://doi.org/10.1177%2F1541931214581181
- Chokshi, N. (2018). Is Alexa listening? Amazon Echo sent out recording of couple’s conversation. New York Times, 25 May. Available from: https://www.nytimes.com/2018/05/25/business/amazon-alexa-conversation-shared-echo.html. [Accessed 12 September 2018].
- Claisse, F. (2012). Contr(ôl)e-fiction: de l'empire à l'interzone. Multitudes, 48(1), 106-117. doi:10.3917/mult.048.0106. Available from: https://doi.org/10.3917/mult.048.0106.
- Conklin, J., (2006) Dialogue mapping: building shared understanding of wicked problems. Chichester: John Wiley & Sons.
- Colorado State University (2013). Transferability: Glossary of Key Terms. Accessed 02/02/2018. Available from: http://writing.colostate.edu/guides/guide.cfm?guideid=90, [accessed 3 May 2020].
- Corritore, C. L., Kracher, B., & Wiedenbeck, S. (2003). On-line trust: concepts, evolving themes, a model. International Journal of Human-Computer Studies, 58(6), 737–758
- Cox, G. (2005). Cox review of creativity in business: building on the UK’s strengths. London: HM Treasury Accessed 20/03/2019. Available from: https://webarchive.nationalarchives.gov.uk/20120704143146/http://www.hm-treasury.gov.uk/d/Cox_review-foreword-definition-terms-exec-summary.pdf
- Craig, P., Cooper, C., Gunnell, D., Haw, S., Lawson, K., Macintyre, S., Thompson, S. (2012). Using natural experiments to evaluate population health interventions: New

medical research council guidance. Journal of Epidemiology and Community Health, 66, 1182–1186

- Craig, P., Cooper, C., Gunnell, D., Haw, S., Lawson, K., Macintyre, S., Thompson, S. (2011). Using natural experiments to evaluate population health interventions: Guidance for producers and users of evidence. London, UK: Medical Research Council.
- Crain M (2018) The limits of transparency: data brokers and commodification. New Media Soc 20(1):88–104. https://doi.org/10.1177/1461444816657096
- Cramer-Petersen, C. L., Christensen, B. T., & Ahmed-Kristensen, S. (2019). Empirically analysing design reasoning patterns: abductive-deductive reasoning patterns dominate design idea generation. Design Studies, 60, 39-70. DOI: 10.1016/j.destud.2018.10.001
- Cross, N., (1982). Designerly ways of knowing, Design Studies 3(4), 221-227.

D

- Danermark, B. (2001). Explaining Society: An Introduction to Critical Realism in the Social Sciences, Routledge, Florence, KY.
- Deutsch, M. (1973). The resolution of conflict: constructive and destructive processes. New Haven, CT: Yale University Press.
- de Visser, E. J., Cohen, M., Freedy, A., & Parasuraman, R. (2014). A design methodology for trust cue calibration in cognitive agents. In: Lecture notes in computer science, pp. 251–262. Cham: Springer.
- de Visser, E. J., Krueger, F., McKnight, P., Scheid, S., Smith, M., Chalk, S., & Parasuraman, R. (2012). The world is not enough: trust in cognitive agents. In: Proceedings of the Human Factors and Ergonomics Society 56th Annual Meeting (pp. 263–267). Santa Monica, CA: Human Factors and Ergonomics Society
- de Visser, E. J., Pak, R., & Shaw, T. H. (2018). From ‘automation’ to ‘autonomy’: the importance of trust repair in human–machine interaction. Ergonomics, 61(10),1409-1427. Available from: https://doi.org/10.1080/00140139.2018.1457725
- Dewey, J. (1927) The public and its problems. Chicago: Swallow Press.
- Dewey, J. (1933) How we think: a restatement of the relation of reflective thinking to the educative process. Boston: Houghton Mifflin.
- DiNardo, J. (2008). Natural experiments and quasi-natural experiments. In Durlauf, Steven N.; Blume, Lawrence E (eds.). The new Palgrave dictionary of economics (Second ed.). London: Palgrave Macmillan, pp. 856–864
- Dorst, K., (2011) The core of “Design Thinking” and its application. Design Studies, 32, 521-532. Available from: https://doi.org/10.1016/j.destud.2011.07.006
- Dorst, K. (2010). The nature of design thinking. In: DTRS8 Interpreting Design Thinking: Design Thinking Research Symposium Proceedings, 2010, pp. 131 - 139
- Douven, I. (2011). Abduction. The Stanford encyclopedia of philosophy (Spring 2011 Edition), Ed- ward N. Zalta ed. Accessed 21/04/2018. Available on: plato.stanford.edu/archives/spr2011/entries/abduction

- Drnec, K., Marathe, A. R., Lukos, J. R., & 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. Available from: 10.3389/fnhum.2016.00290
- Drucker, P. F. (1969). The age of discontinuity: Guidelines to our changing society. New York: Harper & Row.
- Dubois, A. and Gadde, L.-E. (2002). "Systematic combining: an abductive approach to case research", Journal of Business Research, Vol. 55, pp. 553-60.
- Dunning, T. (2012). Natural experiments in the social sciences: a design-based approach. Cambridge: Cambridge University Press.

E

- Ehlers, R. (2017). Formal verification of piece-wise linear feed-forward neural networks. BT - automated technology for verification and analysis. In: 15th International Symposium, ATVA 2017, Pune, India, October 3-6, 2017, Proceedings.
- Eichstaedt, J. C., Smith, R. J., Merchant, R. M., Ungar, L. H., Crutchley, P., Preoţiuc-Pietro, D., Asch, D. A., Schwartz, H. D. (2018). Facebook language predicts depression in medical records. Proceedings of the National Academy of Sciences 115 (44), 11203-11208. Available from: DOI: 10.1073/pnas.1802331115
- Elkins, A.C., Derrick, D.C. The Sound of Trust: Voice as a Measurement of Trust During Interactions with Embodied Conversational Agents. Group Decis Negot 22, 897–913 (2013). https://doi.org/10.1007/s10726-012-9339-x
- 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. Available from: https://doi.org/10.1177/0018720816681350
- Engeler, B. (2017) Towards prospective design. The Design Journal, 20: sup1, S4591-S4599. DOI: 10.1080/14606925.2017.1352956
- EU GDPD (2019a). GDPR Regulation: Art. 15 GDPR: Right of access by the data subject. Accessed 20/02/2020. Available on: https://gdpr-info.eu/art-15-gdpr/
- EU (2019b). Liability for AI (and other emerging digital technologies). EU Commission. doi:10.2838/25362

F

- Floridi L (2012) Distributed morality in an information society. Sci Eng Ethics 19(3):727–743. https://doi.org/10.1007/s11948-012-9413-4
- Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. Harvard Data Science Review, 1(1). Available from: https://doi.org/10.1162/99608f92.8cd550d1

- Floridi, L. (2019b) Translating principles into practices of digital ethics: five risks of being unethical. Philos Technol 32(2):185–193. https://doi.org/10.1007/s13347-019-00354-x
- Foucault, M. (2004). Sécurité, Territoire, Population. Cours au Collège de France (1977-1978), Paris, Seuil, coll. « Hautes Études », p. 50. In Claisse, F. (2012). Contr(ôl)e-fiction: de l'Empire à l'Interzone. Multitudes, 48(1), 106-117. doi:10.3917/mult.048.0106. Available from: https://doi.org/10.3917/mult.048.0106.
- Fulmer, C. A., & Gelfand, M. J. (2012). At what level (and in whom) we trust: trust across multiple organizational levels. Journal of Management, 38(4), 1167–1230.
- Fuller, R. B. (1957). "Comprehensive Anticipatory Design Science". Royal Architectural Institute of Canada Journal. J. F. Sullivan. 34 (9), 357–361. Retrieved 2019-12-21.
- Future, I. (2009). Anticipatory governance. Retrieved: 25 March 2019. Available from: https://www.iftf.org/uploads/media/SR-1272_anticip_govern-1.pdf

G

- Garcia, D., Kreutzer, C., Badillo-Urquiola, K., & Mouloua, M. (2015). Measuring trust of autonomous vehicles: a development and validation study. In: C. Stephanidis (Ed.), HCI International 2015 - Posters' Extended Abstracts, 529, 610–615. Cham: Springer International Available from: https://doi.org/10.1007/978-3-319-21383-5_102
- Galdon, F., & Wang, S. J. (2019a). Designing trust in highly automated virtual assistants: A taxonomy of levels of autonomy. Artificial Intelligence. In: Industry 4.0: A collection of innovative research case-studies. International Conference on Industry 4.0 and Artificial Intelligence Technologies IAIT. Cambridge, UK.
- Galdon, F., & Wang, S. J. (2019b). From apology to compensation; A multi-level taxonomy of trust reparation for highly automated virtual assistants. In: Proceedings of the 1st International Conference on Human Interaction and Emerging Technologies (IHIET 2019) conference August 22- 24, 2019, Nice, France.
- Galdon, F., & Wang, S. J. (2019c). Addressing accountability in highly autonomous virtual assistants. In: Proceedings of the 1st International Conference on Human Interaction and Emerging Technologies (IHIET 2019) August 22-24, 2019, Nice, France.
- Galdon, F., & Wang, S. J. (2019d). Optimising user engagement in highly automated virtual assistants to improve energy management and consumption. In: Proceedings of the 2019 Applied Energy Symposium AEAB Conference Proceedings, MIT, Boston. 22-24 May 2019.
- Galdon, F., & Wang, S. J. (2019e). Future development of AI Virtual Assistants (VAs) in Energy management and consumption. In: Proceedings of the 2019 Applied Energy Symposium AEAB Conference Proceedings, MIT. 22-24 May 2019
- Galdon, F., Hall, A. & Wang, S. J. (2019f). Prospective design: A future-led mixed-methodology to mitigate unintended consequences. In: Proceedings of the International

Association of Societies of Design Research Conference IASDR2019, The University of Manchester, UK.

- Galdon, F., Hall, A. (2019g). The ontological nature of design; prospecting new futures through probabilistic knowledge. In: Design Research for Change Symposium. Design Museum, London
- Galdon, F., Hall, A. (2020a). The right to reparations: a new digital right for repairing trust in the emerging era of highly autonomous systems. In: Proceedings of the 2nd International Conference on Human Interaction and Emerging Technologies: Future Applications (IHIET-AI 2020) Lausanne, Switzerland.
- Galdon, F., Hall, A. (2020b). Synthetic Consequential Reasoning: facilitating the design of synthetic morality in highly automated systems via a multidimensional-scalar framework. In: Proceedings of the 2nd International Conference on Human Interaction and Emerging Technologies: Future Applications (IHIET-AI 2020) Lausanne, Switzerland.
- Galdon, F., Hall, A., & Ferrarello, L. (2020c). Designing trust in artificial intelligence: A comparative study among specifications, principles and levels of control. In: Proceedings of the 2nd International Conference on Human Interaction and Emerging Technologies: Future Applications (IHIET-AI 2020) Lausanne, Switzerland.
- Galdon, F., Hall, A, Ferrarello, L. (2020d). Futuring and trust; A prospective approach to designing trusted futures via a comparative study among design future models. DCS 2020 symposium; Scenarios, Speculation, and Strategies. https://storage.googleapis.com/wzukusers/user-31742378/documents/38f9355a71a04e198259f33b96e02de8/DCS2020%20-%20EBOOK%20FINAL-2.pdf
- Galdon, F., Hall, A, Ferrarello, L. (2021a). Enhancing abductive reasoning in design and engineering education via probabilistic knowledge: a case study in AI. International conference on engineering and product design education (E&DPE2021).
- Galdon, F., Hall, A. (2021). (Un)Frayling design research in design education for the 21Cth. The 14th EAD Conference; Safe harbours, Lancaster, UK, 11-16 Oct 2021.
- Garfinkel, H. (1967). Studies in ethnomethodology, Englewood Cliffs, NJ: Prentice Hall. Available from; https://researchonline.rca.ac.uk/4907/2/SH_EAD2021-unfrayling-FG%20V3%20AUG.pdf
- Gefen, D., Karahanna, E., & Straub, D. W. (2003). Trust and TAM in online shopping: an integrated model. MIS Quarterly, 27, 51–90.
- GGDEC (2019). German Government Data Ethics Commission. Accessed 10/01/2020. Available from: https://datenethikkommission.de/wp-content/uploads/191023_DEK_Kurzfassung_en_bf.pdf
- Giddens, A. (1990). The consequences of modernity. Cambridge: Polity Press.
- Gidley, J. M. (2017). The future; A very short introduction. Oxford: Oxford University Press.
- Glanville, R. (2005). Design propositions. In: M. Belderbos and J. Verbeke, eds. The unthinkable doctorate: Brussels: Sint Lucas.
- Gong, L. (2008). How social is social responses to computers? The function of the degree of anthropomorphism in computer representations. Computers in Human Behavior, 24, 1494–1509

- Gonzatto, R. F., van Amstel, F. M., Merkle, L. E., & Hartmann, T. (2013). The ideology of the future in Design Fictions. Digital creativity, 24(1), 36-45.
- Gomez, F. J. and Schmidhuber, J. (2005). Evolving modular fast-weight networks for control. In W. Duch et al. (Eds.): Proc. ICANN'05, LNCS 3697, pp. 383-389, Cham: Springer.
- Green, B. D. (2010). Applying human characteristics of trust to animated anthropomorphic software agents Ph.D thesis, University at Buffalo.
- Gregson, N., Watkins, H., Broughton, L., Mackenzie, J., & Shepherd, J. (2011). Building bridges through performance and decision-making: schools, research and public engagement. Antipode, 44(2), 343–364. doi:10.1111/j.1467-8330.2010.00839.x
- Griffin, S. (2017). Facebook's artificial intelligence robots shut down after they start talking to each other in their own language. Independent, 31 July. Available from: https://www.independent.co.uk/life-style/gadgets-and-tech/news/facebook-artificial-intelligence-ai-chatbot-new-language-research-openai-google-a7869706.html . Accessed 31/07/2017.
- Guston, D. H. (2014). Understanding 'anticipatory governance.' Social Studies of Science, 44(2), 218–242. Available from: https://doi.org/10.1177/0306312713508669

H

- Hall, A. (2011). Experimental design: design experimentation. Design Issues, 27(2), Spring, 17-26
- Hall, A., Ferrarello, L., Anderson, P., Cooper, R., Ross, C., (2019). Designing Design for Safety. Manchester: International Association of Societies of Design Research
- Hancock, P. A. (2017). Imposing limits on autonomous systems. Ergonomics, 60(2), 284–291. DOI: 10.1080/00140139.2016.1190035
- Hancock, T and Bezold, C., (1994). Possible futures, preferable futures. Healthcare Forum Journal, vol. 37, no. 2, pp. 23-29.
- Harari, Y. N. (2019). Professor Yuval Noah Harari In conversation with Lord Hague of Richmond. RUSI, 13th November 2018. Accessed 15/11/2018. Available from: https://www.ynharari.com/wp-content/up- loads/2018/12/20181113-RUSI-Harari_Discussion_TRANSCRIPT.pdf
- Haslam, N. (2006). Dehumanisation: an integrative review. Personality and Social Psychology Review 10(3), 252–264.
- Haslam, N., Bain, P., Douge, L., Lee, M., & Bastian, B. (2005). More human than you: attributing humanness to self and others. Journal of Personality and Social Psychology, 89(6), 937–950.
- Hill RK (2016) What an algorithm is. Philos Technol 29(1):35–59. https://doi.org/10.1007/s13347-014-0184-5
- Hoff, K. A., & 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. Available from: https://doi.org/10.1177/0018720814547570

I

- IDC (2018). All categories of smart home devices forecast to deliver double-digit growth through 2022, says IDC. IDC, Accessed 12/03/2019. Available from: https://www.id- c.com/getdoc.jsp?containerId=prUS44361618
- IBM (2017). IBM Shoebox. Accessed 20/08/2017. Available from: https://www.ibm.com/ibm/history/exhibits/specialprod1/specialprod1_7.html
- Irwin, T, Kossoff, G., Tonkinwise, C., Scupelli, P. (2015). Transition Design 2015; A new area of design research, practice and study that propose design-led societal transition toward more sustainable futures. Pittsburgh: Carnegie Mellon University.

J

- Jackson, S. L. (2009). Research methods and statistics': a critical thinking approach. Third edition. Wadsworth: Cengage Learning.
- Jasanoff, S., (2016). The ethics of invention: technology and the human future. New York: W.W. Norton & Company.
- Jian, J., Bisantz, A. and Drury, C. (2000). Foundations for an empirically determined scale of trust in automated systems. International Journal of Cognitive Ergonomics, 4, 53-71. Available from: https://doi.org/10.1207/S15327566IJCE0401_04
- Jin, H., Wang, S. (2018). Voice-based determination of physical and emotional characteristics of users. US Patent 10681212. USPTO Patent Full-Text and Image Database Accessed 20/12/2018. Available from: http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO2&Sect2=HITOFF&u=%2Fne- tahtml%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
- Jones, J. C. (1992). Design methods. New York: Van Nostrand Reinhold.
- Johnson, C. D., Miller, M. E., Rusnock, C. F., & Jacques, D. R. (2017). A framework for understanding automation in terms of levels of human control abstraction (pp. 1145–1150). In 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC). Banff, Canada. Available from: https://doi.org/10.1109/SMC.2017.8122766

K

- 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. doi:10.1177/1555343417737203
- Kahneman, D. (2011). Thinking, fast and slow. New York: Farrar, Straus and Giroux.

- Kantar (2016). Brand and trust in a fragmented news environment. Oxford: Reuters Institute for the Study of Journalism.
- Kaplan, A. D., Kessler, T. T., & Hancock, P. A. (2020). How Trust is Defined and its use in Human-Human and Human-Machine Interaction. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 64(1), 1150–1154. https://doi.org/10.1177/1071181320641275
- Katz, G., Barrett, C. W., Dill, D. L., Julian, K. And Kochenderfer, M. J. (2017). 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.
- Kinsella, B. (2019). U.S. smart speaker ownership rises 40% in 2018 to 66.4 million and Amazon Echo maintains market share lead says new report from Voicebot. Voicebot, 7 March. Accessed 8/03/2019. Available from: https://voicebot.ai/2019/03/07/u-s-smart-speaker-ownership-rises-40-in-2018-to-66-4-million-and-amazon-echo-maintains-market-share-lead-says-new-report-from-voice- bot/
- Kim, P. H., Ferrin, D. L., Cooper, C. D., & Dirks, K. T. (2004). Removing the shadow of suspicion: the effects of apology versus denial for repairing competence-versus integrity-based trust violations. Journal of Applied Psychology, 89(1), 104.
- Kirkeby, O.F. (1990). "Abduktion", in Andersen, H. (Ed.), Vetenskapsteori och metodla ¨ra. Introduktion, (translated by Liungman, C.G.), Studentlitteratur, Lund.
- Kohli, P., Gowal, S., Dvijotham, K., and Uesato, J. (2019). Towards robust and verified AI: specification testing, robust training, and formal verification. Deepmind. Medium, 28 March 2019. Accessed 29/03/2019. Available from: https://deepmind.com/blog/robust-and-verified-ai/.
- Kohn, S. C., Quinn, D., Pak, R., de Visser, E. J., & Shaw, T. H. (2018). Trust repair strategies with self-driving vehicles: an exploratory study. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 62(1), 1108–1112. doi: 10.1177/1541931218621254
- Kohring, M. (2004). Vertrauen in Journalismus. Konstanz: UVK Verlagsgesellschaft
- Kovács, G. and Spens, K.M. (2005). Abductive reasoning in logistics research. International Journal of Physical Distribution & Logistics Management. Vol. 35 No. 2, pp. 132-144. https://doi.org/10.1108/09600030510590318
- Kurzweil, R. (2005). The singularity is near. New York: Viking Books.

L

- Lacson, F. C., Wiegmann, D. A., & Madhavan, P. (2005). Effects of attribute and goal framing on automation reliance and compliance. In: Proceedings of the Human Factors and Ergonomics Society 49th Annual Meeting (pp. 482–486). Santa Monica, CA: Human Factors and Ergonomics Society.
- Lane, C and Bachmann, R (1998). Trust within and between organisations: conceptual issues and empirical application. Oxford: Oxford University Press

- Leatherdale, C. T. (2019). Natural experiment methodology for research: a review of how different methods can support real-world research. International Journal of Social Research Methodology, 22 (1), 19-35, DOI: 10.1080/13645579.2018.1488449
- Lassalle, J. M. (2019). Ciberleviatán. El colapso de la democracia liberal frente a la revolución digital [Cyberleviathan. The Collapse of Liberal Democracy Against the Digital Revolution]. (1st ed.) Arpa.
- Lee, E.J. (2008). Flattery may get computers somewhere, some- times: The moderating role of output modality, computer gender, and user gender. International Journal of Human- Computer Studies, 66, 789–800.
- Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operator's adaptation to automation. International Journal of Human-Computer Studies, 40, 153–184. Available from: https://doi.org/10.1006/ijhc.1994.1007
- Lee, J., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. Ergonomics, 35, 1243–1270. Available from: https://doi.org/10.1080/00140139208967392
- Lee, J. D., & See, K. A. (2004). Trust in automation: designing for appropriate reliance. Human Factors, 46, 50–80. Available from: https://doi.org/10.1518/hfes.46.1.50_30392
- Levin, P H., (1966). 'Decision making in urban design': Building Research Station Note EN51/66 Garston, Herts: Building Research Station.
- Lewicki, R. J., Tomlinson, E. C., & Gillespie, N. (2006). Models of interpersonal trust development: Theoretical approaches, empirical evidence, and future directions. Journal of Management, 32(6), 991–1022.
- Lewis, M., Sycara, K., & Walker, P. (2018). The role of trust in human-robot interaction. In H. A. Abbass, J. Scholz, & D. J. Reid (Eds.), Foundations of Trusted Autonomy (pp. 135–159). Australia: Springer Open. Available from: https://doi.org/10.1007/978-3-319-64816-3_8
- Li, Y. M., & Yeh, Y. S. (2010, July). Increasing trust in mobile commerce through design aesthetics. Computers in Human Behavior, 26, 673–684.
- Liddell, H. G. & Scott, R (1940). λήθη. In A Greek–English Lexicon. Revised and augmented throughout by Sir Henry Stuart Jones with the assistance of Roderick McKenzie. Oxford. Clarendon Press 1940. Accessed 21/04/2020. Available on: https://www.perseus.tufts.edu/hopper/text?doc=Perseus:text:1999.04.0057:entry=lh/qh
- Liverpool Knowledge Quarter, (2018). Bringing the vision to life. Accessed 12/03/2019. Available from: https://www.kqliverpool.co.uk/pdf/kql_vision.pdf
- Lohmann, J.C. (2017). The Department of Seaweed; co-speculative design in a museum residency. Ph.D thesis, Royal College of Art. Available from: https://researchonline.rca.ac.uk/3704/4/JuliaLohmannPh.DThesis2018.pdf Accessed 10/05/2020.
- Lowerre, B. T. (1976). The HARPY speech recognition system. Ph.D thesis, Stanford University. Available from: https://stacks.stanford.edu/file/druid:rq916rn6924/rq916rn6924.pdf Accessed 20/08/2018.

M

- Madhavan, P., Wiegmann, D. A., & Lacson, F. C. (2006). Automation failures on tasks easily performed by operators undermine trust in automated aids. Human Factors, 48, 241–256.
- Madsen, M., & Gregor, S. (2000). Measuring human-computer trust. Gladstone, Australia, Central Queensland University: 12pp. Accessed 20/11/2018. Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.93.3874&rep=rep1&type=pdf
- Mallet, L., (2018). Creating quality living: the new Swedish town offering innovative solutions to London's housing crisis. Evening Standard, 10 January. Accessed 04/12/2018. Available from: https://www.homesandproperty.co.uk/property-news/the-new-swedish-town-offering-innovative-solutions-to-londons-housing-crisis-a116751.html
- Malpass, M. (2017): Critical design in context. London, New York: Bloomsbury.
- Manzey, D., Reichenbach, J., & Onnasch, L. (2012). Human performance consequences of automated decision aids: The impact of degree of automation and system experience. Journal of Cognitive Engineering and Decision Making, 6, 57–87.
- Marchen, R., (2018). Impact from critical research: what might it look like and what support is required?. LSE Impact Blog. September 4th, 2018. Accessed 14/03/2019. Available from: https://blogs.lse.ac.uk/impactofsocialsciences/2018/09/04/impact-from-critical-research-what-might-it-look-like-and-what-support-is-required/
- Marinik, A., Bishop, R., Fitchett, V., Morgan, J. F., Trimble, T. E., & 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
- Martin, R. (2009). The design of business. Cambridge, MA: Harvard Business Press.
- Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. Academy of Management Review, 20(3), 709–734
- Mayer, R. C. & Davis, J. H. (1999). The effect of the performance appraisal system on trust for management: A field quasi-experiment. Journal of Applied Psychology, 84(1), 123–136. Available from: https://psycnet.apa.org/doi/10.1037/0021-9010.84.1.123
- Michael, M. (2019). An introduction to unintended consequences. Mark Michael. Accessed 16/03/2019. Available from: https://www.markmichael.io/insights/mapping-mitigating-unintended-consequences/
- Mirman, M., Gehr, T. & Vechev, M. (2018). Differentiable abstract interpretation for provably robust neural networks. In: Proceedings of the 35th International Conference on Machine Learning, in PMLR 80:3578-3586
- Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. Nature Machine Intelligence, 4 November. doi:10.1038/s42256-019-0114-4
- Mittelstadt, B., Russell, C., and Wachter, S. (2019). Explaining explanations in AI. In FAT* ’19: Conference on Fairness, Accountability, and Transparency (FAT* ’19), January

29–31, 2019, Atlanta, GA, USA. New York, NY, ACM. Available from: https://doi.org/10.1145/3287560.3287574

- Mittelstadt BD, Allo P, Taddeo M, Wachter S, Floridi L. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society. doi:10.1177/2053951716679679
- Möllering, G. (2006). Trust: reason, routine, reflexivity. Oxford: Elsevier
- Moray, N., Hiskes, D., Lee, J. D., & Muir, B. (1995). Trust and human intervention in automated systems. In: J. M. Hoc, P. C. Cacciabue, & E. Hollnagel (Eds.) Expertise and Technology: Cognition & Human-Computer Cooperation. Hillsdale, NJ: Lawrence Erlbaum. (pp. 183–194). Accessed 20/08/2018. Available from: https://www.researchgate.net/publication/234830217_Trust_and_human_intervention_in_aut omated_systems?enrichId=rgreq-e749f9d23daf33523fdca3d2e6c70cd2-XXX&enrichSource=Y292ZXJQYWdlOzIzNDgzMDIxNztBUzo0MjYwODgwMTU4OTY1NzdAMTQ3ODU5ODcwNzg4NA%3D%3D&el=1_x_2&_esc=publicationCoverPdf
- Moray, N., & Inagaki, T. (1999). Laboratory studies of trust between humans and machines in automated systems. Transactions of the Institute of Measurement and Control, 21, 203–211. Available from: https://doi.org/10.1177/014233129902100408
- Moray, N., Inagaki, T., and Itoh, M. (2000). Adaptive automation, trust and self-confidence in fault management of time-critical tasks. J. Exp. Psychol. Appl. 6(1), 44–58. Available from: http://dx.doi.org/10.1037/1076-898X.6.1.44
- Morris, H. and Warman. G. (2015). Using design thinking in higher education. Educause Review, Monday, January 12, 2015. Accessed 14/03/2019. Available from: https://er.educause.edu/articles/2015/1/using-design-thinking-in-higher-education
- Mortier, R., Haddadi, H., Henderson, T., McAuley, D., and Crowcroft, J. (2014). Human-data interaction: The human face of the data-driven society. Available at SSRN 2508051 (2014).
- Morton T. (2013). Hyperobjects: philosophy and ecology after the end of the world. Minneapolis: University of Minnesota Press; 2013.
- Mullins, P. A. (2013). Ground-Breaking project to brain-scan shoppers. News. Bangor University. Accessed 20/10/2017. Available from: https://www.bangor.ac.uk/news/university/ground-breaking-project-to- brain- scan-shoppers-16874
- Multitudes (2012). Political Counter-Fictions – Fukushima: Voices of Rebels. Multitudes, 48.

N

- Naslund, D. (2002). “Logistics needs qualitative research – especially action research”, International Journal of Physical Distribution & Logistics Management, Vol. 32 No. 5, pp. 321-38.
- NCBI, (2016, July 2). National Center for Biotechnology Information: attention span statistics. Statisticbrain. Accessed 20/08/2017. Available from: http://www.statisticbrain.com/attention-span-statistics/

- Neyedli, H. F., Hollands, J. G., & Jamieson, G. A. (2011). Beyond identity: Incorporating system reliability information into an automated combat identification system. Human Factors, 53, 338–355.
- Nienaber, A.-M., Romeike, P., Searle, R., & Schewe, G. (2015). What makes the glue sticky? A qualitative meta-analysis of antecedents and consequences of trust in supervisor-subordinate relationships. Journal of Managerial Psychology, 30(5), 507–534.
- NPR and Edison Research (2019). The smart audio report. Accessed 20/01/2020. Available from: https:// www.nationalpublicmedia.com/wp-content/uploads/2019/01/ Smart- Audio-Report-Winter-2018.pdf

O

- Ou, C. X., & Sia, C. L. (2010). Consumer trust and distrust: An issue of website design. International Journal of Human-Computer Studies, 68, 913–934.
- Ortega, B. P. A. (2018). Building safe artificial intelligence: specification, robustness, and assurance specification: design the purpose of the system. Medium. Available from: https://medium.com/@deepmindsafetyresearch/building-safe-artificial-intelligence-52f5f75058f Accessed 20/08/2018.
- Olssen, M., (2015). Neoliberal competition in higher education today: research, accountability and impact. British Journal of Sociology of Education, 37(1), 129–148. doi:10.1080/01425692.2015.1100530

P

- Pain, R., Kesby, M., & Askins, K. (2010). Geographies of impact: power, participation and potential. Area, 43(2), 183–188. doi:10.1111/j.1475-4762.2010.00978.x
- Pain, R. (2014). Impact: striking a blow or walking together? ACME: An International Journal for Critical Geographies 13 (1), 19-23. Available from: https://www.acme-journal.org/index.php/acme/article/view/986.
- Parasuraman, R., and Riley, V. (1997). Humans and automation; use misuse, disuse, abuse. Hum. Factors 39, 230–253. Available from: https://doi.org/ 10.1518/001872097778543886
- 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. Available from: https://doi.org/10.1109/3468.844354
- Parasuraman, R., Sheridan, T. B., and Wickens, C. D. (2008). Situation awareness, mental workload and trust in automation: viable, empirically supported cognitive engineering constructs. J. Cogn. Eng. Decis. Mak. 2, 140–160. Available from: https:// doi.org/10.1518/155534308X284417

- Parasuraman, R., and Manzey, D. (2010). Complacency and bias in human use of automation: an attentional integration. Hum. Factors 52, 381–410. Available from: https://doi.org/10.1177/0018720810376055
- Parasuraman, R., & Miller, C. A. (2004). Trust and etiquette in high-criticality automated systems. Communications of the ACM, 47(4), 51–55.
- Pak, R., Fink, N., Price, M., Bass, B., & Sturre, L. (2012). Decision support aids with anthropomorphic characteristics influence trust and performance in younger and older adults. Ergonomics, 55, 1059–1072.
- Pelegrini Morita, P., & Burns, C. M. (2014). Trust tokens in team development. Team Performance Management, 20(1/2), 39–64.

R

- Raphael, J. (2018). This Is exactly how much information Google has on you. Esquire, 28 March. Accessed 12/10/2018. Available from: https://www.esquire.com/uk/latest-news/a19614546/this-exactly-how-much-information-google-has-on-you/
- Rabinowitz, N. C., Perbet, F., Song, H. F., Zhang, C., Eslami, S.M. A., Botvinick, M. (2018). Machine theory of mind. Cornell University. https://arxiv.org/abs/1802.07740
- Rittel, H.W.J. & Webber, (1973). Dilemmas in a General Theory of Planning, M.M. Policy Sci 4: 155.
- Resnick, E. (2019). The social design reader. London: Bloomsbury, 2019.
- Rotter, J. B. (1967). A new scale for the measurement of interpersonal trust. Journal of Personality, 35(4), 651-665.
- Rosenzweig, M. R.; Wolpin, K. I. (2000). 'Natural experiments' in economics. Journal of Economic Literature. 38 (4), 827–874. doi:10.1257/jel.38.4.827.
- Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). Introduction to special topic forum: not so different after all: a cross-discipline view of trust. Academy of Management Review, 23(3), 393–404
- Rubel A, Castro C, Pham A (2019) Agency laundering and information technologies. Ethical Theory Moral Pract 22(4):1017–1041. https://doi.org/10.1007/s10677-019-10030-w
- Ruttkay, Z., and Pelachaud, C. (2004). From brows to trust: evaluating embodied conversational agents. Vol. 7. Springer Science & Business Media. Doi: 10.1007/1-4020-2730-3

S

- Sanchez, J. (2006). Factors that affect trust and reliance on an automated aid. Ph.D thesis, Georgia Institute of Technology, Atlanta.

- Saunders, L. (2008). An evolving map of design practice and design research. ACM — Interactions — Volume XV.6. Accessed 12/10/2019. Available from: http://www.dubberly.com/wp-content/uploads/2008/11/ddo_article_evolvingmap.pdf
- Sarter, N. B., & Woods, D. D. (1997). Team play with a powerful and independent agent: operational experiences and automation surprises on the Airbus A-320. Human Factors, 39(4), 553–569. Available on: https://doi.org/10.1518/001872097778667997
- Sciuto, A., Saini, A., Forlizzi, J., & Hong, J. I. (2018). "Hey Alexa, what's up?": a mixed-methods studies of in-home conversational agent usage. In: Proceedings of the 2018 on Designing Interactive Systems Conference 2018 - DIS '18, 857–868. Available from: https://doi.org/10.1145/3196709.3196772
- Schoorman, F. D., Wood, M. M., & Breuer, C. (2015). Would trust by any other name smell as sweet? Reflections on the meanings and uses of trust across disciplines and context. In B. Bornstein & A. Tomkins (Eds.), Motivating cooperation and compliance with authority (pp. 13–35). New York: Springer International Publishing.
- Seeger, A.M., Heinzl, A. (2018) Human versus machine: contingency factors of anthropomorphism as a trust-inducing design strategy for conversational agents. In: Davis F., Riedl R., vom Brocke J., Léger PM., Randolph A. (eds.) Information Systems and Neuroscience. Lecture Notes in Information Systems and Organisation, vol 25. Springer, Cham
- Seeger, A.M., Pfeiffer, J., & Heinzl, A. (2017). When Do We Need a Human? Anthropomorphic Design and Trustworthiness of Conversational Agents. SIGHCI 2017 Proceedings 15. http://aisel.aisnet.org/sighci2017/15
- Seppanen, R., K. Blomqvist and S. Sundqvist (2007) Measuring inter- organisational trust: a critical review of the empirical research in 1990–2003 .Industrial Marketing Management, 36 (2), 249–65.
- Shadish, William R., et al., Experimental and quasi-experimental designs for generalized causal inference. Boston: Houghton Mifflin.
- Sheridan, T. B., & Verplank, W. L. (1978). Human and computer control of undersea teleoperators. Fort Belvoir, VA: Defence Technical Information Centre. Available on: https://doi.org/10.21236/ADA057655
- Silverstone, R. and Haddon, L. (1996) Design and the domestication of information and communication technologies: technical change and everyday life. In: Mansell, Robin and Silverstone, Roger, (eds.) Communication by design: the politics of information and communication technologies. Oxford: Oxford University Press, pp. 44-74
- Simon, H. (1996). The sciences of the artificial. Cambridge, MA: MIT Press.
- Simpson, A., Brander, G. N., & Portsdown, D. R. A. (1995). Seaworthy trust: confidence in automated data fusion. In: R. M. Taylor & J. Reising (Eds.) The human-electronic crew: can we trust the team, (pp 77–81). Hampshire, UK: Defence Research Academy. Available on: http://www.dtic.mil/dtic/tr/fulltext/u2/a308589.pdf . Accessed 12/02/2018.
- Snow, C. P. (1959). The two cultures and scientific revolution. New York: Cambridge University Press, 2013.

- Spain, R. D., & Madhavan, P. (2009). The role of automation etiquette and pedigree in trust and dependence. In: Proceedings of the Human Factors and Ergonomics Society 54th Annual Meeting (pp. 339–343). Santa Monica, CA: Human Factors and Ergonomics Society

T

- Taylor, S.S., Fisher, D. and Dufresne, R.L. (2002). "The aesthetics of management storytelling: a key to organizational learning", Management Learning, Vol. 33 No. 3, pp. 313-30
- Thropp, J. E. (2006). Individual preferences using automation. Ph.D thesis, University of Central Florida, Orlando
- Tsamados, A., Aggarwal, N., Cowls, J. et al. (2021). The ethics of algorithms: key problems and solutions. AI & Soc. https://doi.org/10.1007/s00146-021-01154-8

U

- Uesato, J., O'Donoghue, B., van den Oord, A., and Kohli, P. (2018). Adversarial risk and the dangers of evaluating against weak attacks. In: Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018

V

- Voros, J. (2001). A primer on futures studies, foresight and the use of scenarios. Prospect: The Foresight. Bulletin, 6 (1).

W

- Wachter, S., & and Mittelstadt, B. (2018). A right to reasonable inferences: re-thinking data protection law in the age of big data and AI. Columbia Business Law Review, Forthcoming. Accessed 01/02/2019. Available on: https://ssrn.com/abstract=3248829
- Wang, L., Jamieson, G. A., and Hollands, J. G. (2009). Trust and reliance on an automated combat identification system. Hum. Factors 51, 281–291. Available on: https://doi.org/10.1177%2F0018720809338842
- Weber, R. H. (2011). The right to be forgotten: more than a Pandora's box? JIPITEC, 120, para 1.
- Westin, C. A., Borst, C., and Hilburn, B. (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. Accessed 12/03/2018. Available on: https://www.sesarju.eu/sites/default/files/documents/sid/2011/SID%202011-MUFASA.pdf

- White, H., & S. Sabarwal (2014). Quasi-experimental design and methods. Methodological Briefs: Impact Evaluation 8. Florence: UNICEF Office of Research,
- Wickens, C. D., and Dixon, S. R. (2007). The benefits of imperfect diagnostic automation: a synthesis of the literature. Theor. Issues Ergon. Sci. 8, 201–212. Available on: https://doi.org/10.1080/14639220500370105
- Wickens, C. D., Li, H., Santamaria, A., Sebok, A., & 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). Available on: https://doi.org/10.1177%2F154193121005400425
- Wigblad, R. (2003). “Praktikteori – en mo¨jlig forskningsstrategi?”, paper prepared for the SIRA Conference ”Interaktiv forskning – utmaningar fo¨r akademin”, available at: www.ehv.vxu. se/forskn/utb/kurser/3fei014/forelasningsmat/p-wigblad18_0.pdf (accessed March 17, 2004).
- Winfield, A. (2019). Ethical standards in robotics and AI. Nat Electron 2, 46–48. Available on: https://doi.org/10.1038/s41928-019-0213-6
- WHO. (2018). List of Blueprint priority diseases. Geneva: World Health Organization [Online] Accessed 01/03/2018. Available on: http://www.who.int/blueprint/priority-diseases/en/
- Welt, Die (2012). From IBM Shoebox to Siri: 50 years of speech recognition, Die Welt. 20 April, Available at:https://www.welt.de/newsticker/dpa_nt/infoline_nt/computer_nt/article106206488/Von-IBM-Shoebox-bis-Siri-50-Jahre-Spracherkennung.html

Y

- Yagoda, R. E., & Gillan, D. J. (2012). You want me to trust a ROBOT? The development of a human–robot interaction trust scale. International Journal of Social Robotics, 4, 235–248. Available on: https://doi.org/10.1007/s12369-012-0144-0

Z

- Zand, D.E. (1972). Trust and managerial problem solving , Administrative Science Quarterly, 17 (2), 229–39.
- Zhou, T. (2011). The effect of initial trust on user adoption of mobile payment. Information Development, 27, 290–300.

1

—

APPENDIX

CHRONOLOGY OF TRUST DESIGN

DESIGNING TRUST; A HISTORICAL ANALYSIS

Before focusing into how we can design trusted system in the fully automated paradigm, we need to build a contextual understanding on how trust has been design historically. Its strategies, its methods, its techniques or aesthetics.

In this area, I decided to map and categorise how trust has been implemented and designed to get a preliminary perspective.

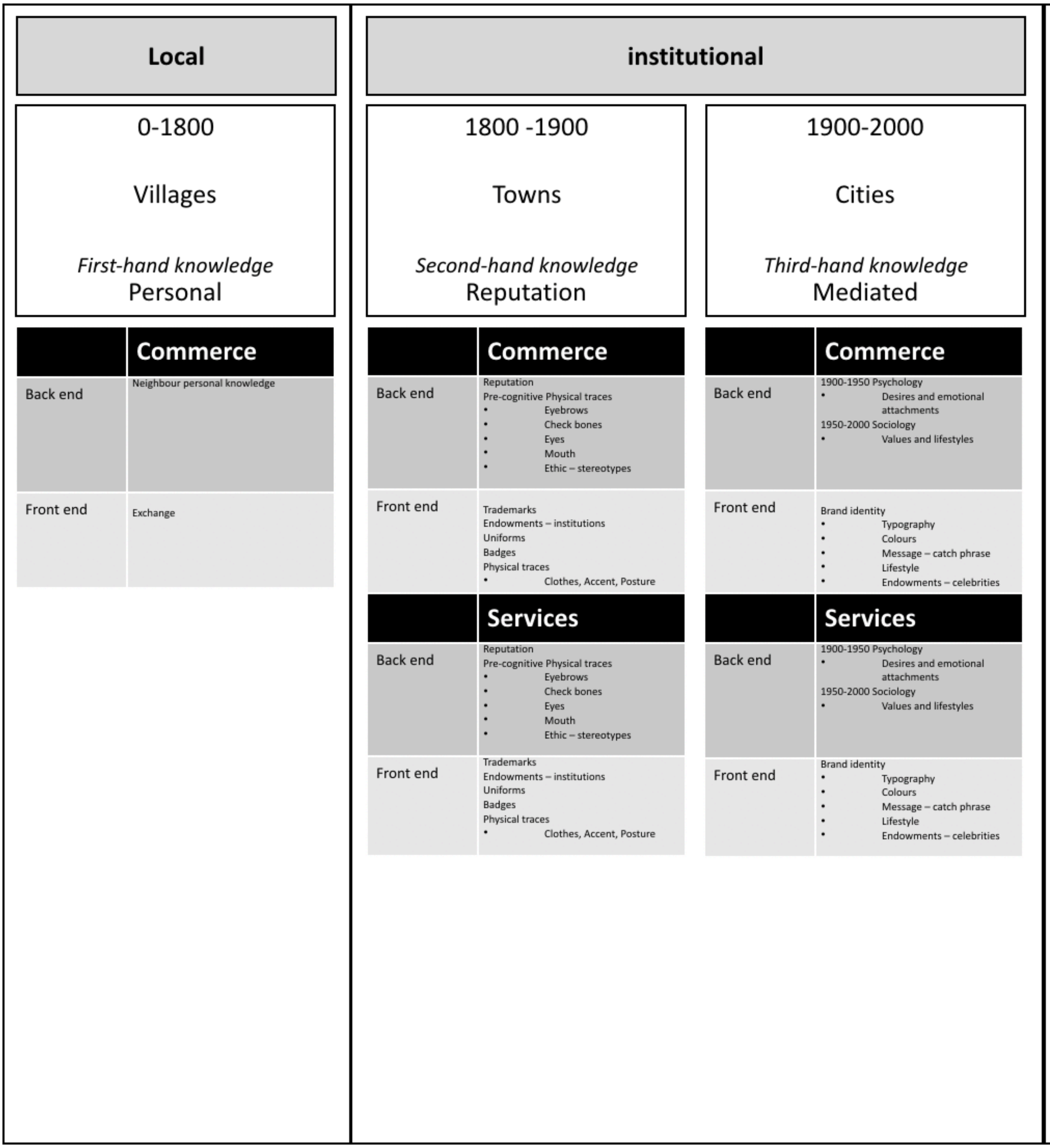
First I decided to categorise all the examples based on the three temporal paradigms introduced by Rachel Botsman (2017); Local, institutional and distributed. In the process, what I realise is that the analysis worked very well in the past, well in the near past, but very difficult in the digital paradigm.

This aspect forced me to reconsider categorisation. A preliminary analysis restructured my *a priory* position. It moved from three time partitions (local, institutional and distributed) to six subcategories. In this process local remained the same, the institutional was divided in two parts; social and corporate. In the third case I did change the denomination from distributed to global. The global paradigm is broken into three parts; static, dynamic and proactive. The main reason being that in the global paradigm they are radically different. Distributed is static, automated is semi-dynamic and delegated is totally dynamic.

Furthermore, I realise it was confusing due to a lack of domain based categorisation. Therefore I decided to cross-categorise with three main domains; commerce, services and social. The cross-categorisation varies with time. In the local only commerce prevails. In the institutional; commerce and service entangle. And in the digital three different categories emerge; e-commerce, social networks and service platforms.

Finally, building from literature, Building from Blobaum (2014), I realise that I needed to implement a multi-level categorisation. However, in order to simplify the task I decided to implement a front-end versus back-end partition.

The outcome clarified my analysis tremendously and provided a template to work with. This template worked as a foundation to build a catalogue of methods, processes and techniques. What I am presenting is a preliminary analysis. To complement this analysis, I aim implementing an archive analysis to underpin key principles and processes of trust design.



Local	institutional	
0-1800 Villages *First-hand knowledge* Personal	1800 -1900 Towns *Second-hand knowledge* Reputation	1900-2000 Cities *Third-hand knowledge* Mediated

Local – 0-1800

	Commerce
Back end	Neighbour personal knowledge
Front end	Exchange

1800 -1900

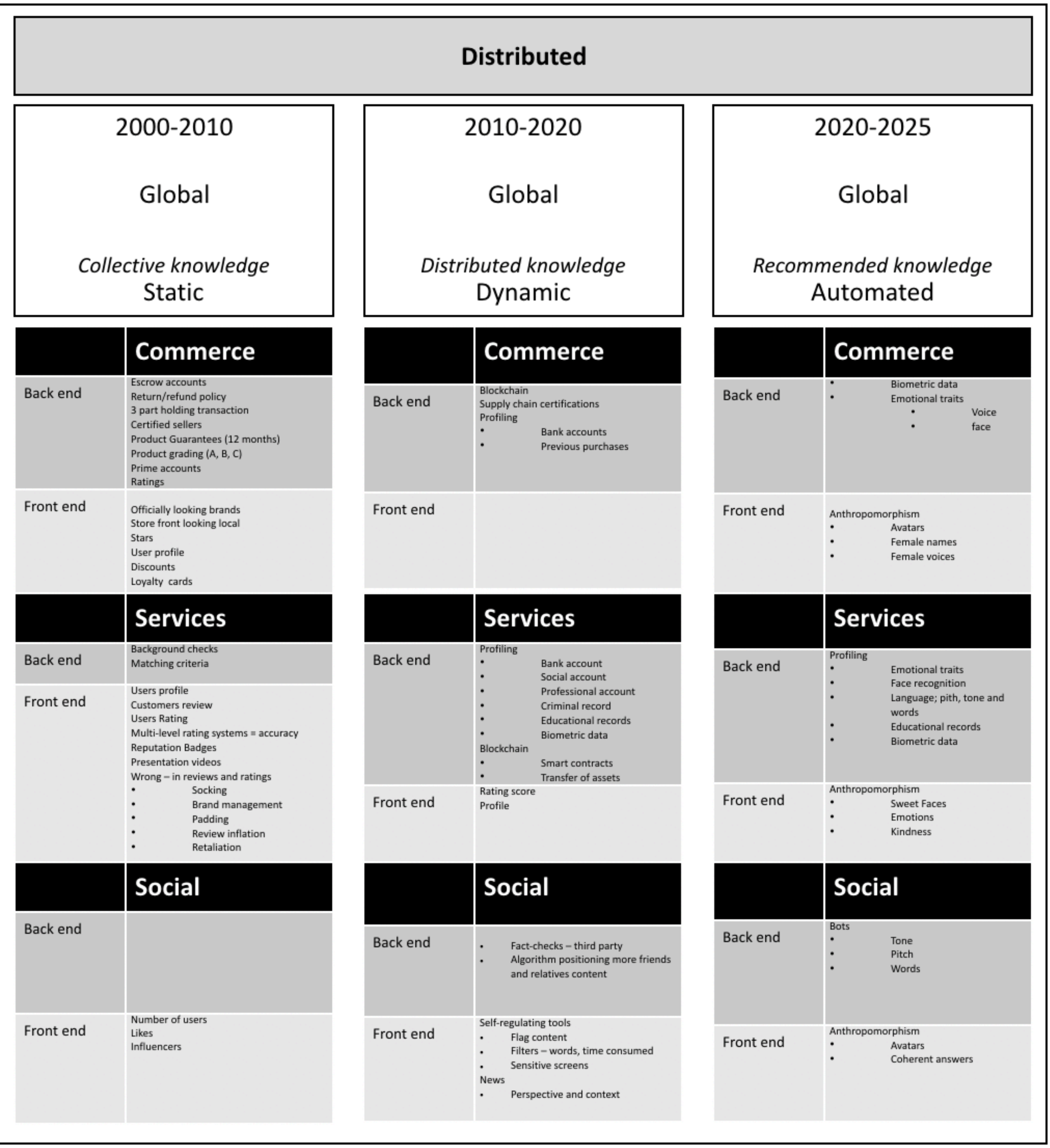
	Commerce
Back end	Reputation Pre-cognitive Physical traces • Eyebrows • Check bones • Eyes • Mouth • Ethic – stereotypes
Front end	Trademarks Endowments – institutions Uniforms Badges Physical traces • Clothes, Accent, Posture

	Services
Back end	Reputation Pre-cognitive Physical traces • Eyebrows • Check bones • Eyes • Mouth • Ethic – stereotypes
Front end	Trademarks Endowments – institutions Uniforms Badges Physical traces • Clothes, Accent, Posture

1900-2000

	Commerce
Back end	1900-1950 Psychology • Desires and emotional attachments 1950-2000 Sociology • Values and lifestyles
Front end	Brand identity • Typography • Colours • Message – catch phrase • Lifestyle • Endowments – celebrities

	Services
Back end	1900-1950 Psychology • Desires and emotional attachments 1950-2000 Sociology • Values and lifestyles
Front end	Brand identity • Typography • Colours • Message – catch phrase • Lifestyle • Endowments – celebrities



Distributed

		2000-2010	2010-2020	2020-2025
		Global	Global	Global
		Collective knowledge Static	*Distributed knowledge* Dynamic	*Recommended knowledge* Automated
Commerce	Back end	Escrow accounts Return/refund policy 3 part holding transaction Certified sellers Product Guarantees (12 months) Product grading (A, B, C) Prime accounts Ratings	Blockchain Supply chain certifications Profiling • Bank accounts • Previous purchases	• Biometric data • Emotional traits • Voice • face
	Front end	Officially looking brands Store front looking local Stars User profile Discounts Loyalty cards		Anthropomorphism • Avatars • Female names • Female voices
Services	Back end	Background checks Matching criteria	Profiling • Bank account • Social account • Professional account • Criminal record • Educational records • Biometric data Blockchain • Smart contracts • Transfer of assets	Profiling • Emotional traits • Face recognition • Language; pith, tone and words • Educational records • Biometric data
	Front end	Users profile Customers review Users Rating Multi-level rating systems = accuracy Reputation Badges Presentation videos Wrong – in reviews and ratings • Socking • Brand management • Padding • Review inflation • Retaliation	Rating score Profile	Anthropomorphism • Sweet Faces • Emotions • Kindness
Social	Back end		• Fact-checks – third party • Algorithm positioning more friends and relatives content	Bots • Tone • Pitch • Words
	Front end	Number of users Likes Influencers	Self-regulating tools • Flag content • Filters – words, time consumed • Sensitive screens News • Perspective and context	Anthropomorphism • Avatars • Coherent answers

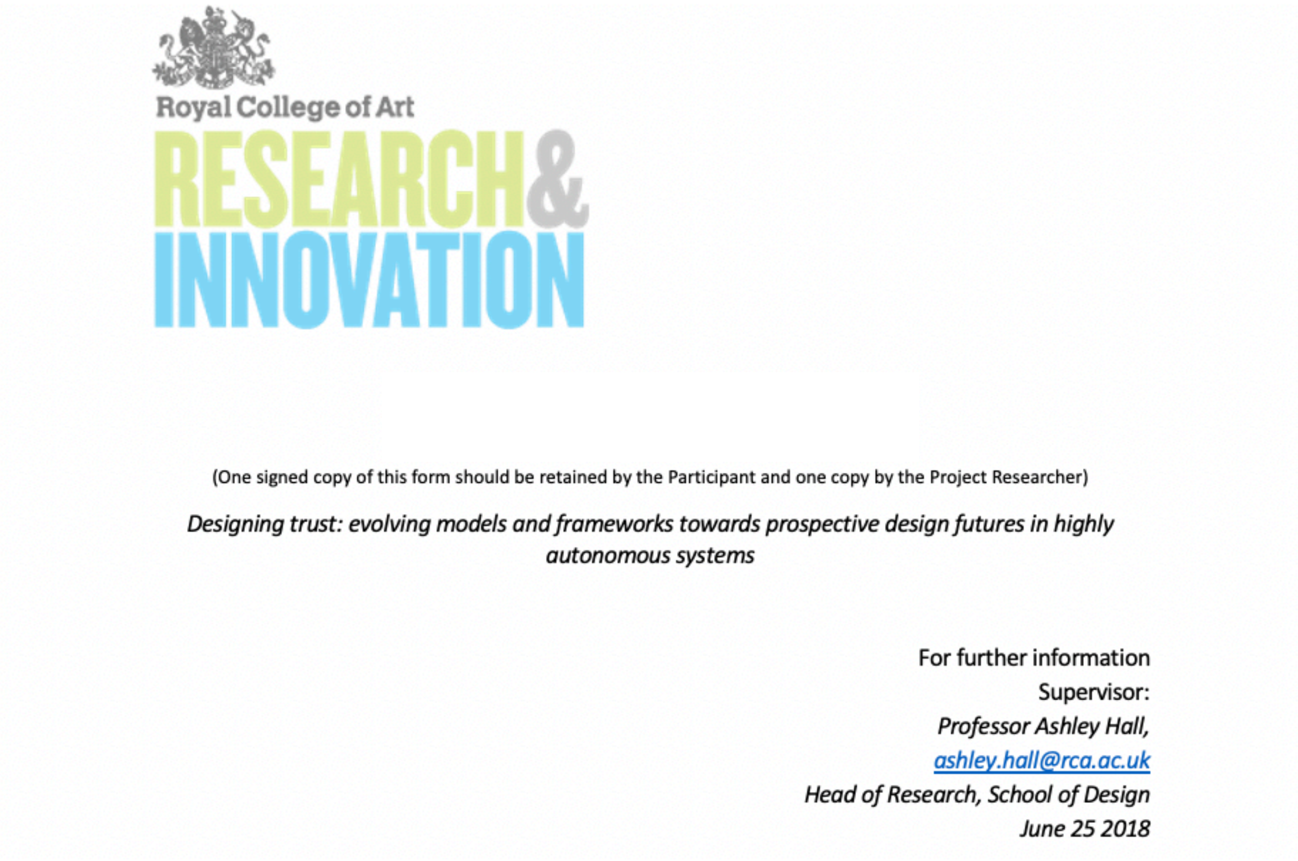
2

—

APPENDIX

ETHICS

ETHICS FORM



Royal College of Art

RESEARCH&
INNOVATION

(One signed copy of this form should be retained by the Participant and one copy by the Project Researcher)

Designing trust: evolving models and frameworks towards prospective design futures in highly autonomous systems

For further information
Supervisor:
Professor Ashley Hall,
ashley.hall@rca.ac.uk
Head of Research, School of Design
June 25 2018

Workshops call

Deconstructing the future; Designing for the unexpected
Workshop with Fernando Galdon

The future is a beautiful place, however, if not properly designed, it may become a nightmare.

Interested in learning how to access and influence the future?
RCA Ph.D research student in GID, Fernando Galdon, will run a workshop introducing a tool-kit to access and systematically analyse the future.

Though a case study on Virtual Assistants and their potential impact, we will investigate how to design for the unexpected.

This will be a hands-on workshop. All material will be provided.
No experience necessary, please come and join!

Limited to 16 attendees - so please make sure you can attend if you reply.
Email fernando.galdon@network.rca.ac.uk to reserve a spot

Day, Month,
Time,
Room STE109, Stevens Building
Royal College of Art, Kensington campus.

Workshop preliminary ethical statement

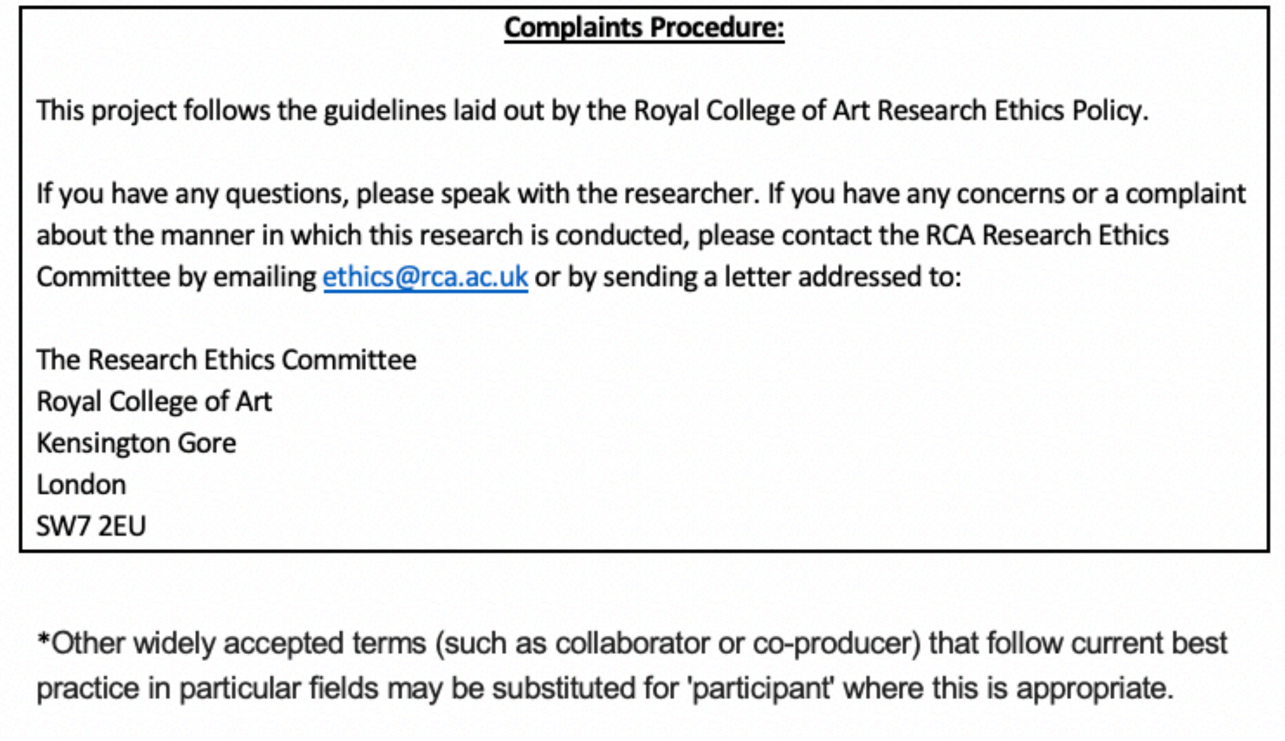
This workshop is part of my Ph.D. We will be implementing a methodology I have developed to design the unexpected.

We will implement this methodology in the context of Virtual assistants.
At the end of the workshop, I will collect all the work to analyse the validity of the methodology. If you want you can take photos to have a copy, you can, but I have to keep the original work for research purposes.

All the tasks are anonymous. Please do not write your name in any part of the pages. This is a standard practice designed to protect your privacy while collecting, analysing, and reporting data. Your name will not be present in any part of the research, outputs, or publications.

Participation & withdrawal

You can choose whether or not to be in this study. If you volunteer to be in this study, you may withdraw at any time without consequences of any kind or loss of benefits to which you are otherwise entitled. You may also refuse to answer any questions you do not want to answer. There is no penalty if you withdraw from the study.



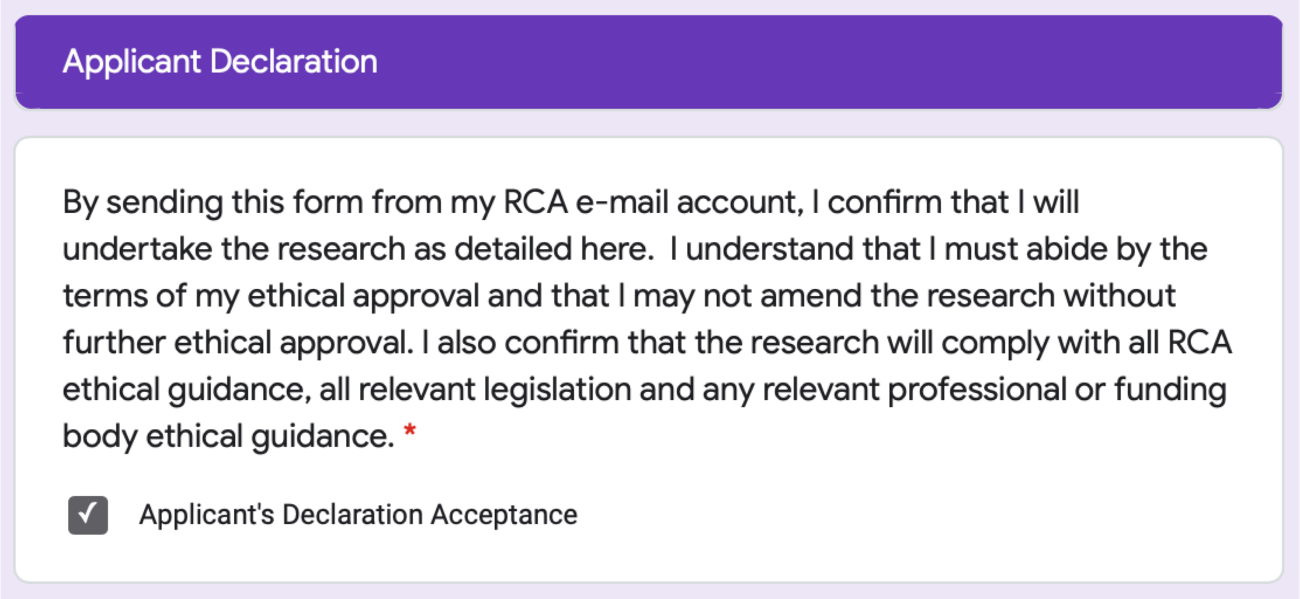
Complaints Procedure:

This project follows the guidelines laid out by the Royal College of Art Research Ethics Policy.

If you have any questions, please speak with the researcher. If you have any concerns or a complaint about the manner in which this research is conducted, please contact the RCA Research Ethics Committee by emailing ethics@rca.ac.uk or by sending a letter addressed to:

The Research Ethics Committee
Royal College of Art
Kensington Gore
London
SW7 2EU

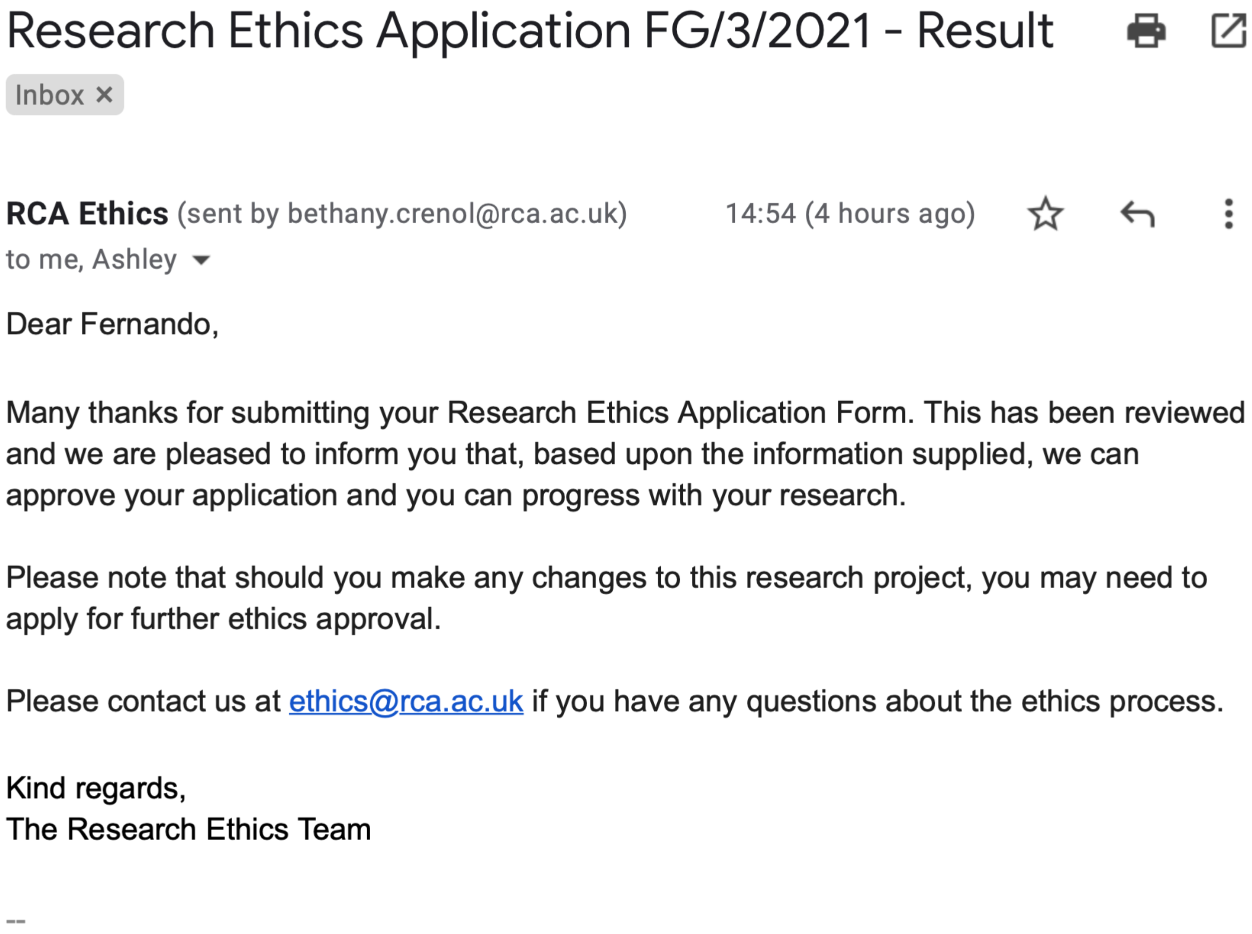
*Other widely accepted terms (such as collaborator or co-producer) that follow current best practice in particular fields may be substituted for 'participant' where this is appropriate.



Applicant Declaration

By sending this form from my RCA e-mail account, I confirm that I will undertake the research as detailed here. I understand that I must abide by the terms of my ethical approval and that I may not amend the research without further ethical approval. I also confirm that the research will comply with all RCA ethical guidance, all relevant legislation and any relevant professional or funding body ethical guidance. *

☑ Applicant's Declaration Acceptance



Research Ethics Application FG/3/2021 - Result

Inbox ×

RCA Ethics (sent by bethany.crenol@rca.ac.uk) 14:54 (4 hours ago)
to me, Ashley

Dear Fernando,

Many thanks for submitting your Research Ethics Application Form. This has been reviewed and we are pleased to inform you that, based upon the information supplied, we can approve your application and you can progress with your research.

Please note that should you make any changes to this research project, you may need to apply for further ethics approval.

Please contact us at ethics@rca.ac.uk if you have any questions about the ethics process.

Kind regards,
The Research Ethics Team

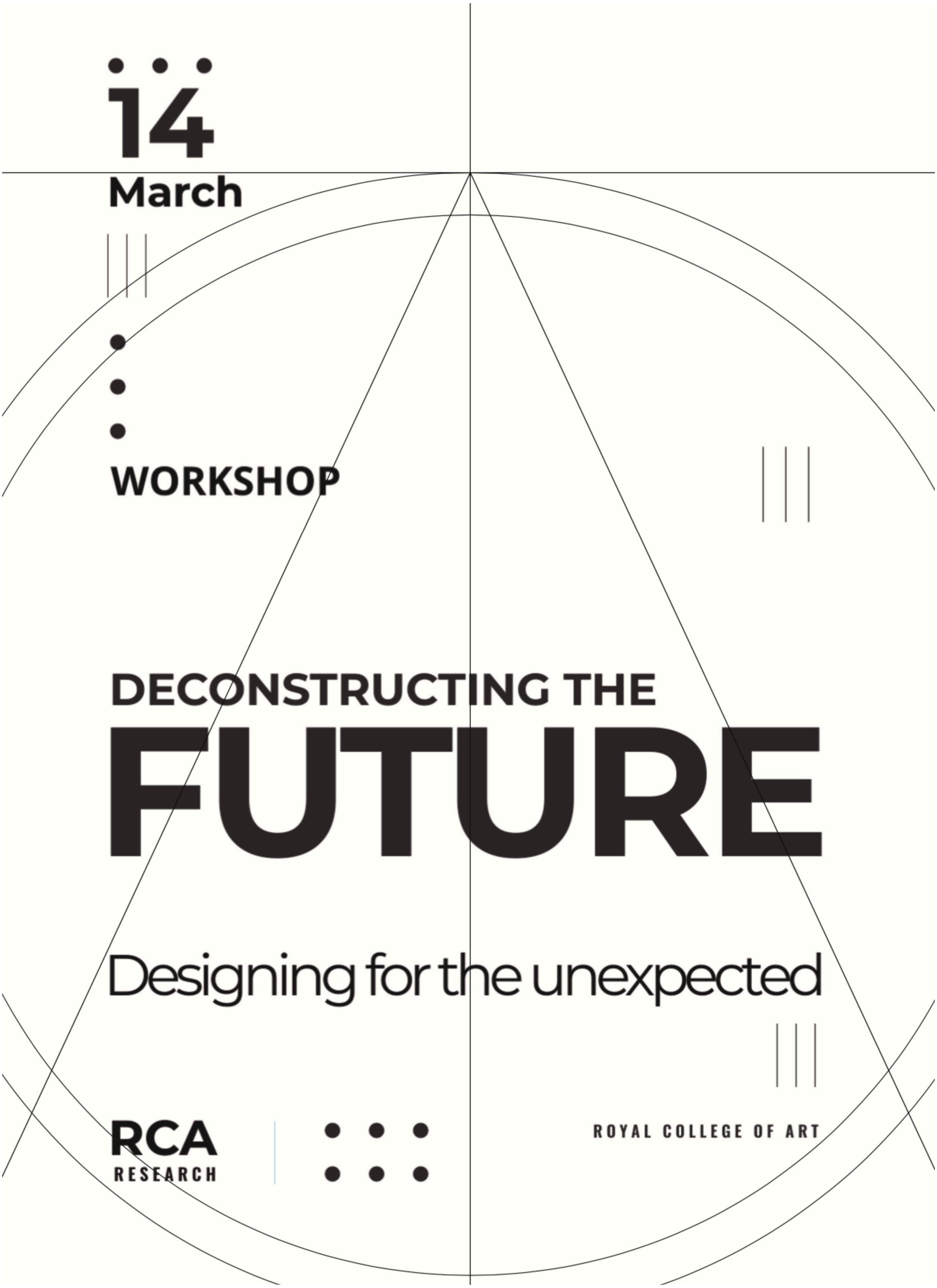
--

3

—

APPENDIX

WORKSHOP 1



14

March

WORKSHOP

DECONSTRUCTING THE

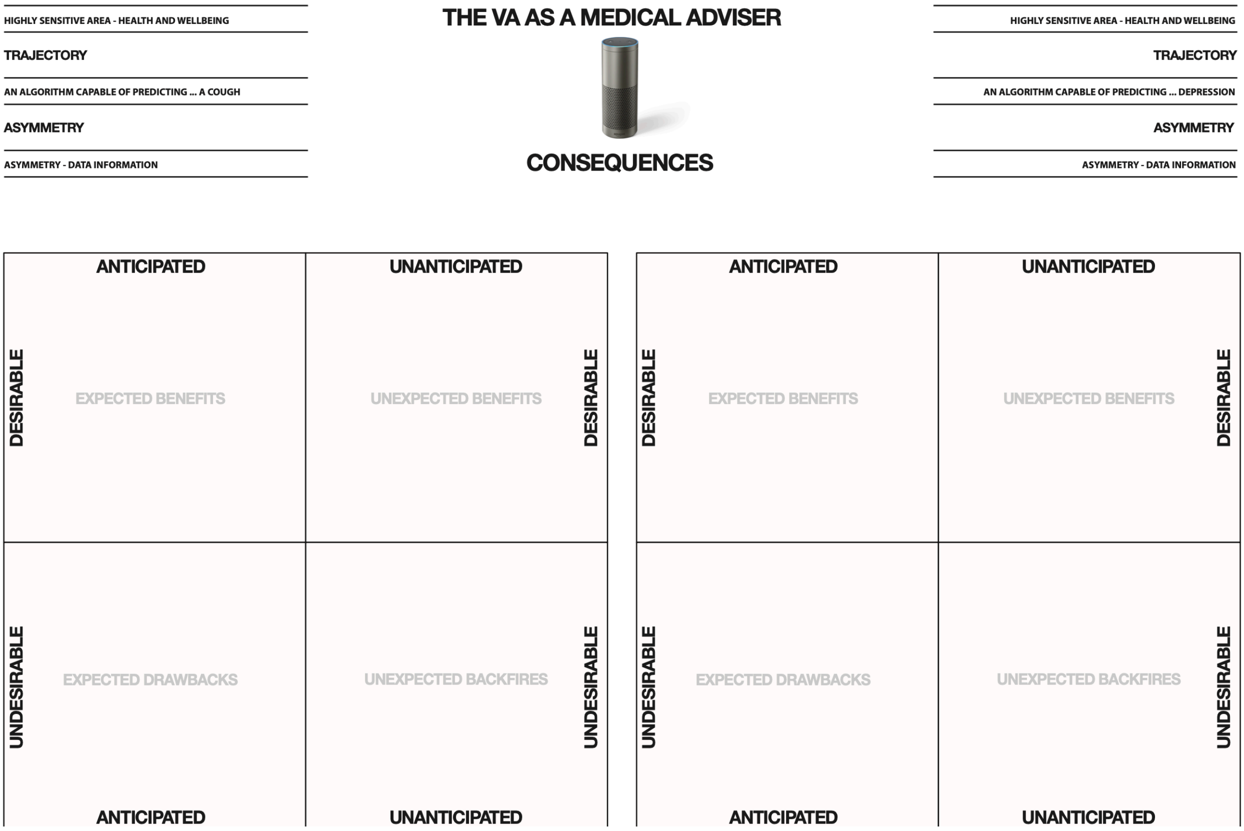
FUTURE

Designing for the unexpected

RCA RESEARCH

ROYAL COLLEGE OF ART

Type to enter a caption.



HIGHLY SENSITIVE AREA - HEALTH AND WELLBEING

TRAJECTORY

AN ALGORITHM CAPABLE OF PREDICTING ... A COUGH

ASYMMETRY

ASYMMETRY - DATA INFORMATION

THE VA AS A MEDICAL ADVISER

CONSEQUENCES

HIGHLY SENSITIVE AREA - HEALTH AND WELLBEING

TRAJECTORY

AN ALGORITHM CAPABLE OF PREDICTING ... DEPRESSION

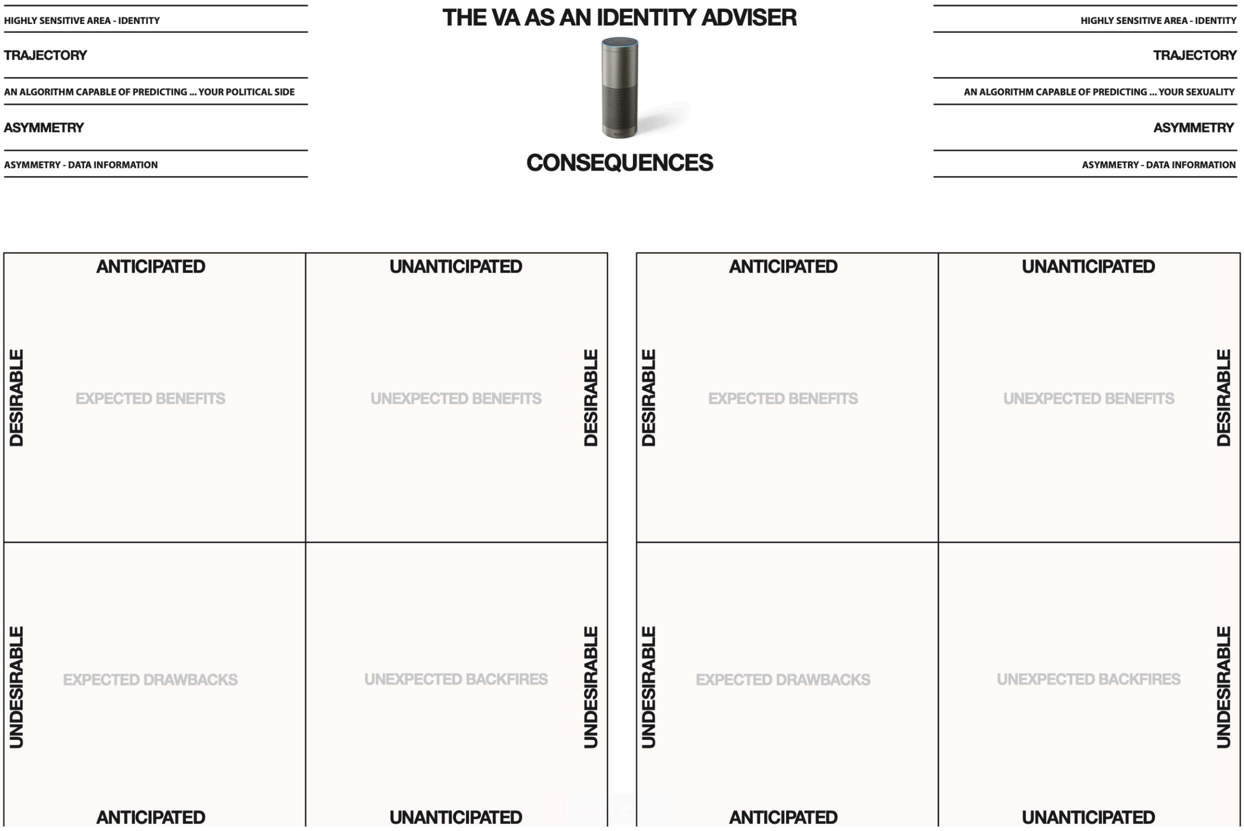
ASYMMETRY

ASYMMETRY - DATA INFORMATION

	ANTICIPATED	UNANTICIPATED	
DESIRABLE	EXPECTED BENEFITS	UNEXPECTED BENEFITS	DESIRABLE
UNDESIRABLE	EXPECTED DRAWBACKS	UNEXPECTED BACKFIRES	UNDESIRABLE
	ANTICIPATED	UNANTICIPATED	

	ANTICIPATED	UNANTICIPATED	
DESIRABLE	EXPECTED BENEFITS	UNEXPECTED BENEFITS	DESIRABLE
UNDESIRABLE	EXPECTED DRAWBACKS	UNEXPECTED BACKFIRES	UNDESIRABLE
	ANTICIPATED	UNANTICIPATED	

Type to enter a caption.



HIGHLY SENSITIVE AREA - IDENTITY

TRAJECTORY

AN ALGORITHM CAPABLE OF PREDICTING ... YOUR POLITICAL SIDE

ASYMMETRY

ASYMMETRY - DATA INFORMATION

THE VA AS AN IDENTITY ADVISER

CONSEQUENCES

HIGHLY SENSITIVE AREA - IDENTITY

TRAJECTORY

AN ALGORITHM CAPABLE OF PREDICTING ... YOUR SEXUALITY

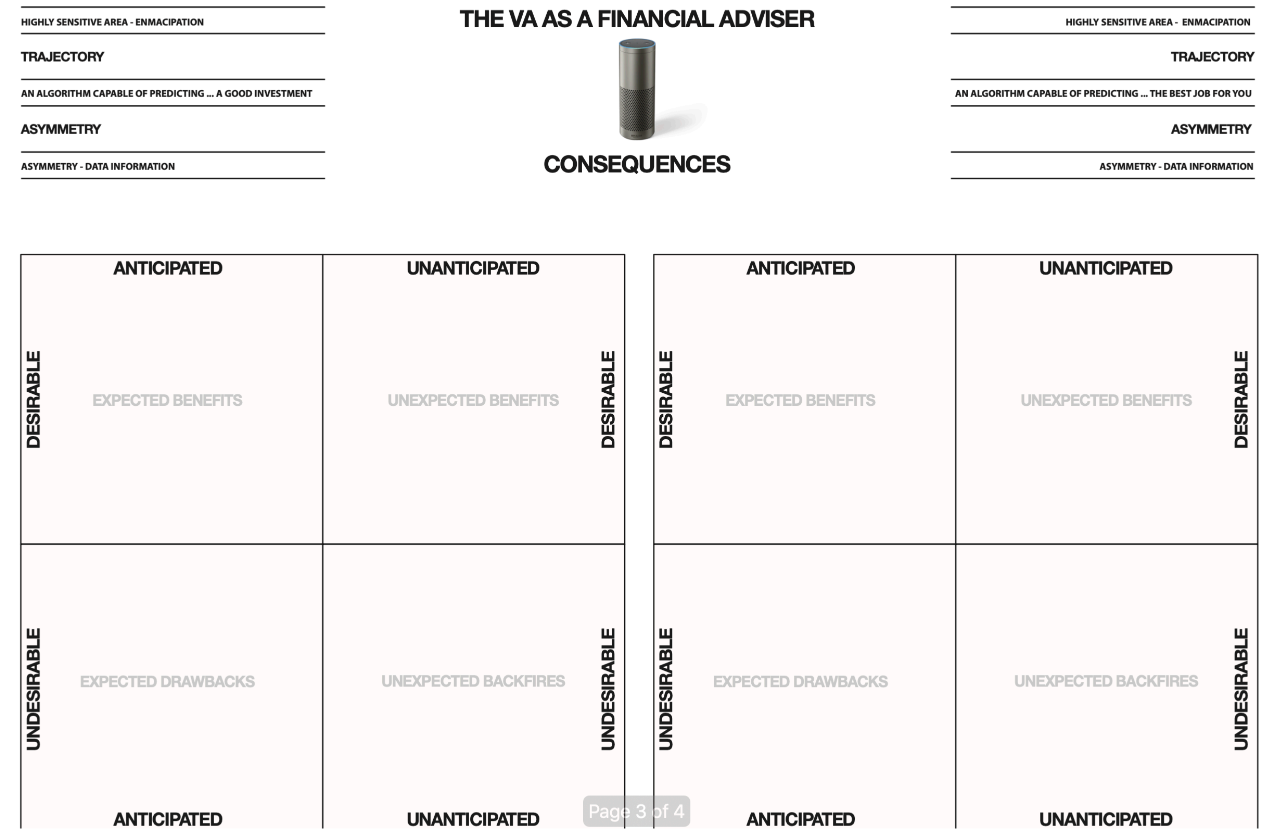
ASYMMETRY

ASYMMETRY - DATA INFORMATION

	ANTICIPATED	UNANTICIPATED	
DESIRABLE	EXPECTED BENEFITS	UNEXPECTED BENEFITS	DESIRABLE
UNDESIRABLE	EXPECTED DRAWBACKS	UNEXPECTED BACKFIRES	UNDESIRABLE
	ANTICIPATED	UNANTICIPATED	

	ANTICIPATED	UNANTICIPATED	
DESIRABLE	EXPECTED BENEFITS	UNEXPECTED BENEFITS	DESIRABLE
UNDESIRABLE	EXPECTED DRAWBACKS	UNEXPECTED BACKFIRES	UNDESIRABLE
	ANTICIPATED	UNANTICIPATED	

Type to enter a caption.



HIGHLY SENSITIVE AREA - ENMACIPATION

TRAJECTORY

AN ALGORITHM CAPABLE OF PREDICTING ... A GOOD INVESTMENT

ASYMMETRY

ASYMMETRY - DATA INFORMATION

THE VA AS A FINANCIAL ADVISER

CONSEQUENCES

HIGHLY SENSITIVE AREA - ENMACIPATION

TRAJECTORY

AN ALGORITHM CAPABLE OF PREDICTING ... THE BEST JOB FOR YOU

ASYMMETRY

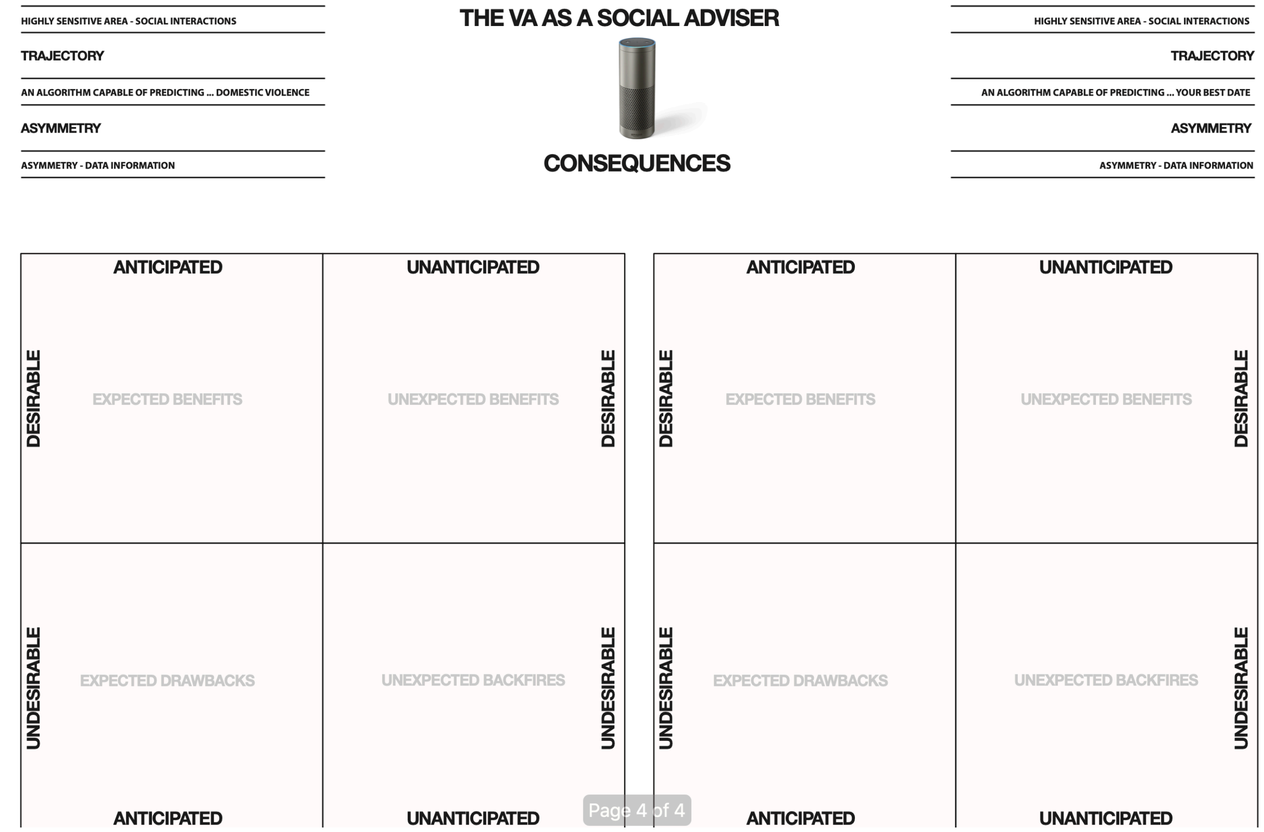
ASYMMETRY - DATA INFORMATION

	ANTICIPATED	UNANTICIPATED	
DESIRABLE	EXPECTED BENEFITS	UNEXPECTED BENEFITS	DESIRABLE
UNDESIRABLE	EXPECTED DRAWBACKS	UNEXPECTED BACKFIRES	UNDESIRABLE
	ANTICIPATED	UNANTICIPATED	

	ANTICIPATED	UNANTICIPATED	
DESIRABLE	EXPECTED BENEFITS	UNEXPECTED BENEFITS	DESIRABLE
UNDESIRABLE	EXPECTED DRAWBACKS	UNEXPECTED BACKFIRES	UNDESIRABLE
	ANTICIPATED	UNANTICIPATED	

Page 3 of 4

Type to enter a caption.



HIGHLY SENSITIVE AREA - SOCIAL INTERACTIONS

TRAJECTORY

AN ALGORITHM CAPABLE OF PREDICTING ... DOMESTIC VIOLENCE

ASYMMETRY

ASYMMETRY - DATA INFORMATION

THE VA AS A SOCIAL ADVISER

CONSEQUENCES

HIGHLY SENSITIVE AREA - SOCIAL INTERACTIONS

TRAJECTORY

AN ALGORITHM CAPABLE OF PREDICTING ... YOUR BEST DATE

ASYMMETRY

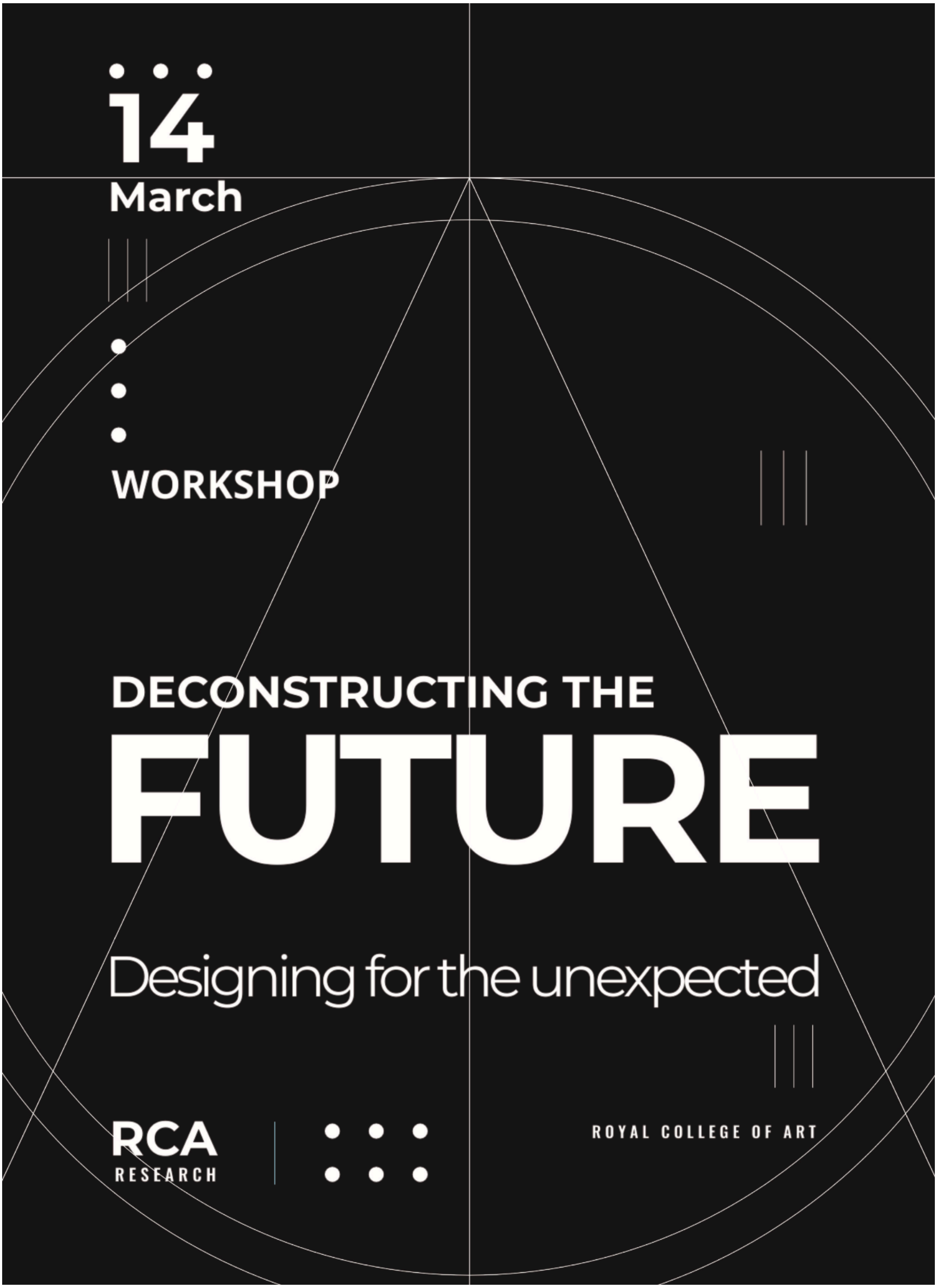
ASYMMETRY - DATA INFORMATION

	ANTICIPATED	UNANTICIPATED	
DESIRABLE	EXPECTED BENEFITS	UNEXPECTED BENEFITS	DESIRABLE
UNDESIRABLE	EXPECTED DRAWBACKS	UNEXPECTED BACKFIRES	UNDESIRABLE
	ANTICIPATED	UNANTICIPATED	

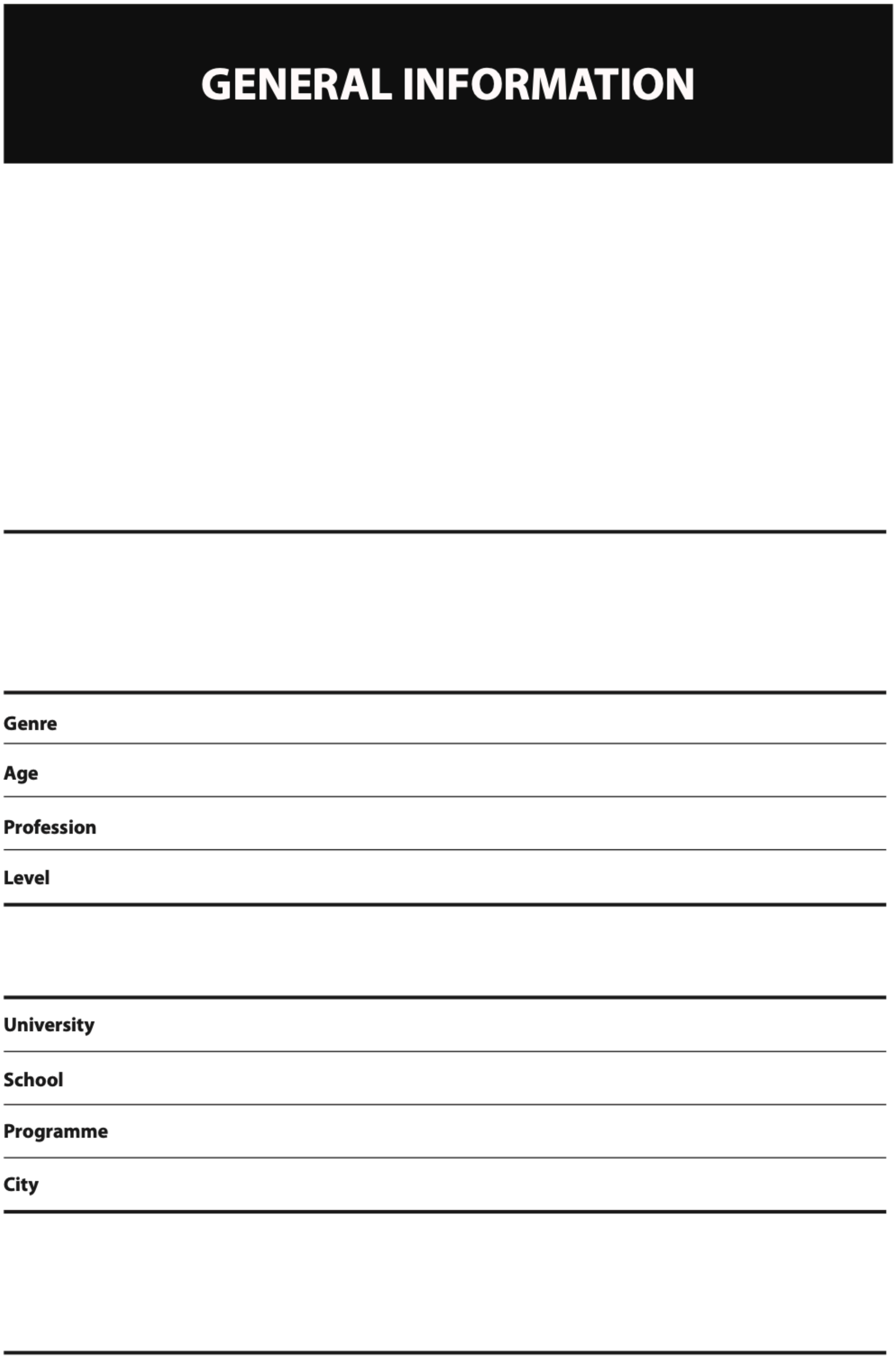
	ANTICIPATED	UNANTICIPATED	
DESIRABLE	EXPECTED BENEFITS	UNEXPECTED BENEFITS	DESIRABLE
UNDESIRABLE	EXPECTED DRAWBACKS	UNEXPECTED BACKFIRES	UNDESIRABLE
	ANTICIPATED	UNANTICIPATED	

Page 4 of 4

Type to enter a caption.



Workshop 1 - Individual task - Cover



GENERAL INFORMATION

Genre

Age

Profession

Level

University

School

Programme

City

Type to enter a caption.

TECHNOLOGY

HIGHLY SENSITIVE AREA

YOUR DISCIPLINE

TRAJECTORY

AN ALGORITHM CAPABLE OF PREDICTING ...

ASYMMETRY

ASYMMETRY -

CONSEQUENCES

	ANTICIPATED	UNANTICIPATED	
DESIRABLE	EXPECTED BENEFITS	UNEXPECTED BENEFITS	DESIRABLE
UNDESIRABLE	EXPECTED DRAWBACKS	UNEXPECTED BACKFIRES	UNDESIRABLE
	ANTICIPATED	UNANTICIPATED	

Type to enter a caption.

SKETCH

DESIGN A PRODUCT WITH AN ANTICIPATED AND UNDESIRABLE CONSEQUENCE IN MIND

Type to enter a caption.

FINAL PRODUCT/SERVICE

NAME OF PRODUCT -

WHAT KIND OF REPARATION STRATEGY WOULD YOU APPLY TO MITIGATE UNINTENDED CONSEQUENCES

APOLOGY	WHAT KIND?	
COMPENSATION	HOW MUCH?	
OTHER	WHAT ELSE?	

Type to enter a caption.

Type to enter a caption.

Type to enter a caption.

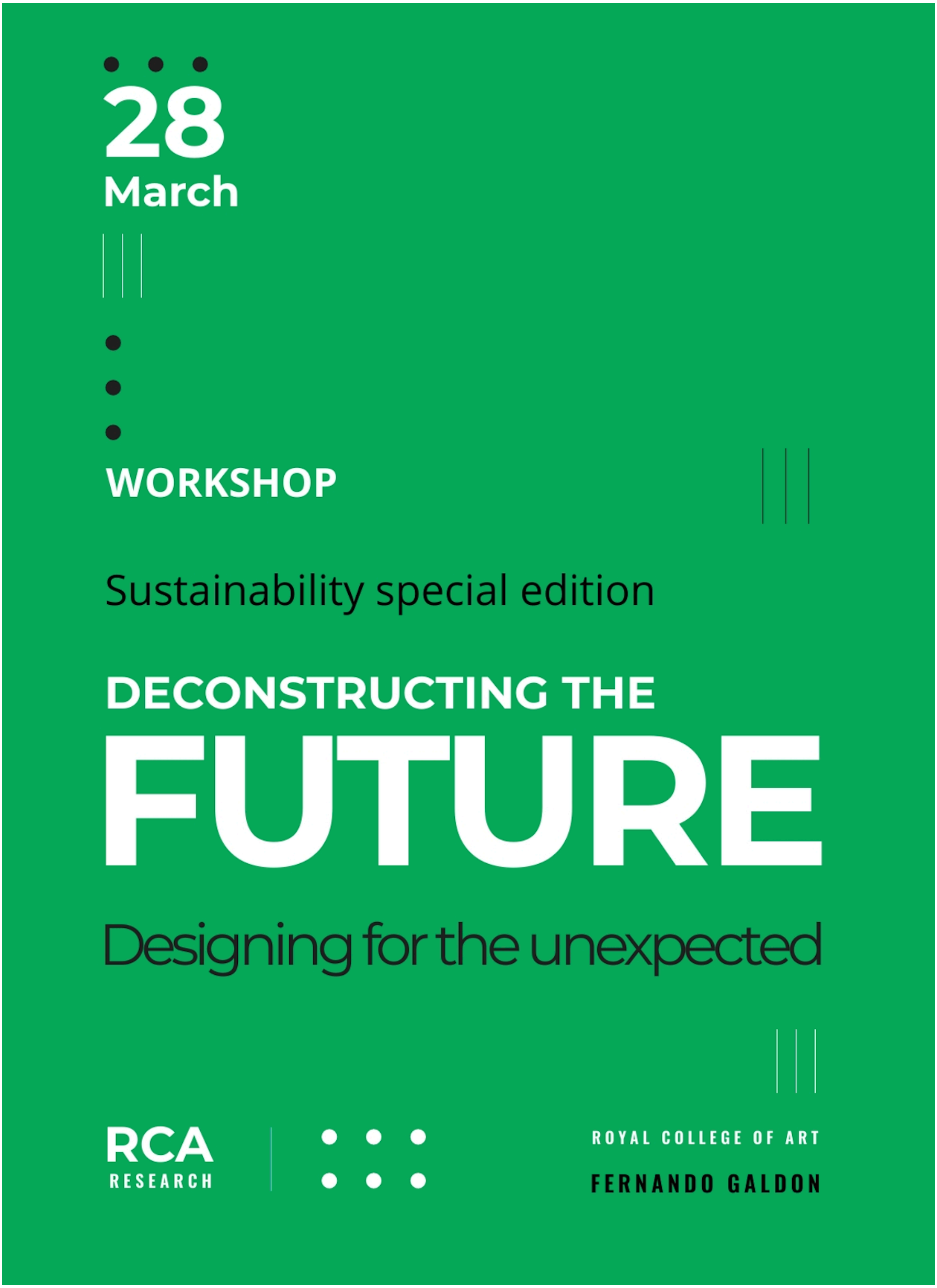
4

—

APPENDIX

WORKSHOP

2



28
March

WORKSHOP

Sustainability special edition

DECONSTRUCTING THE

FUTURE

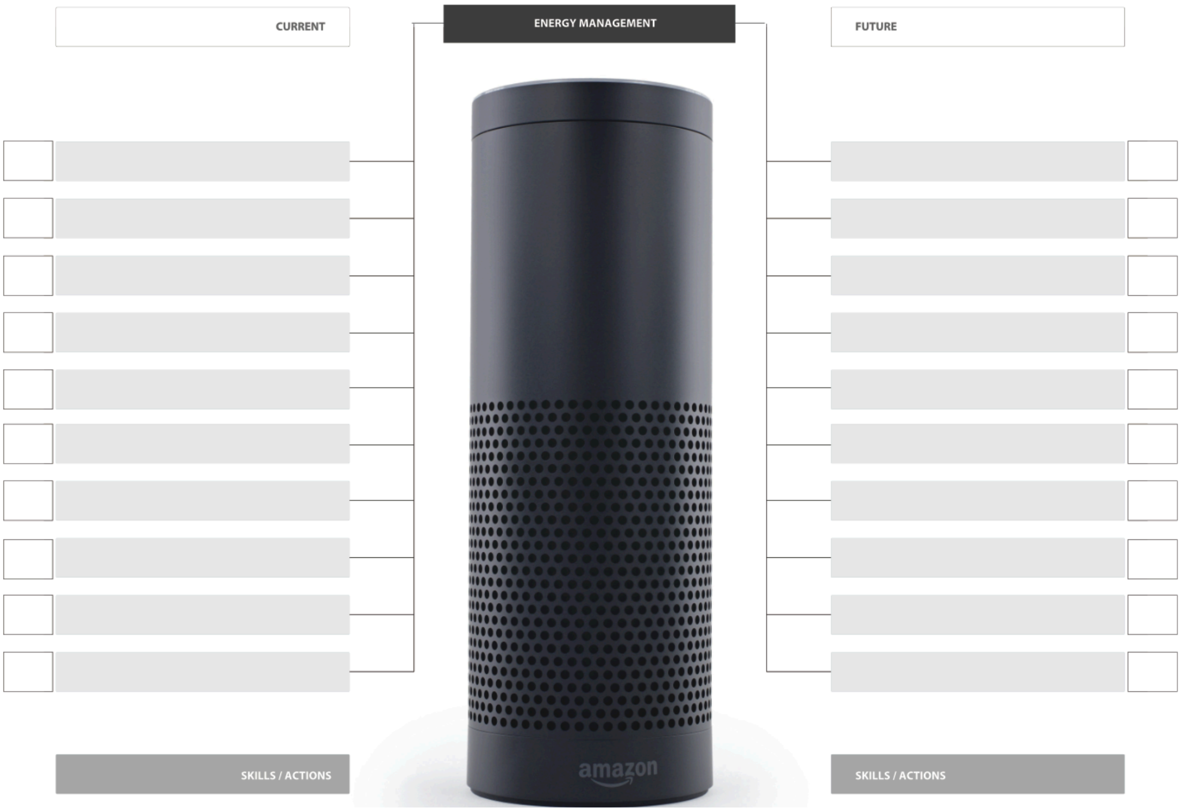
Designing for the unexpected

RCA
RESEARCH

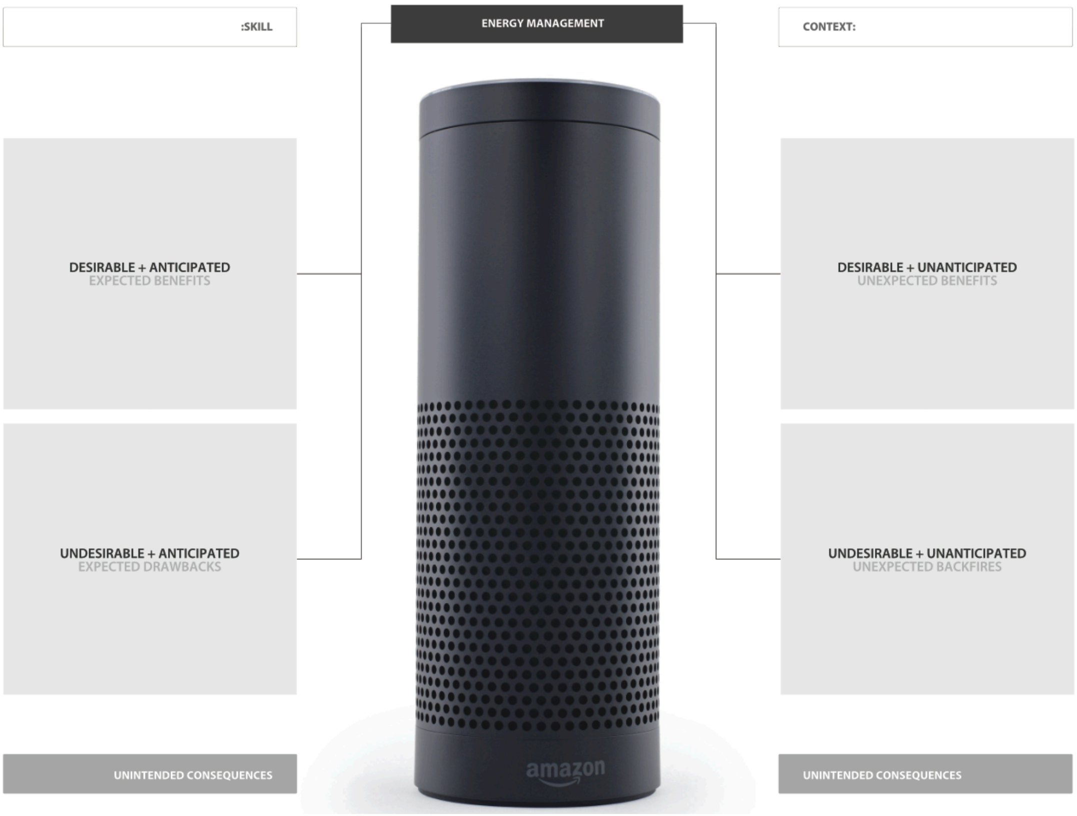
ROYAL COLLEGE OF ART

FERNANDO GALDON

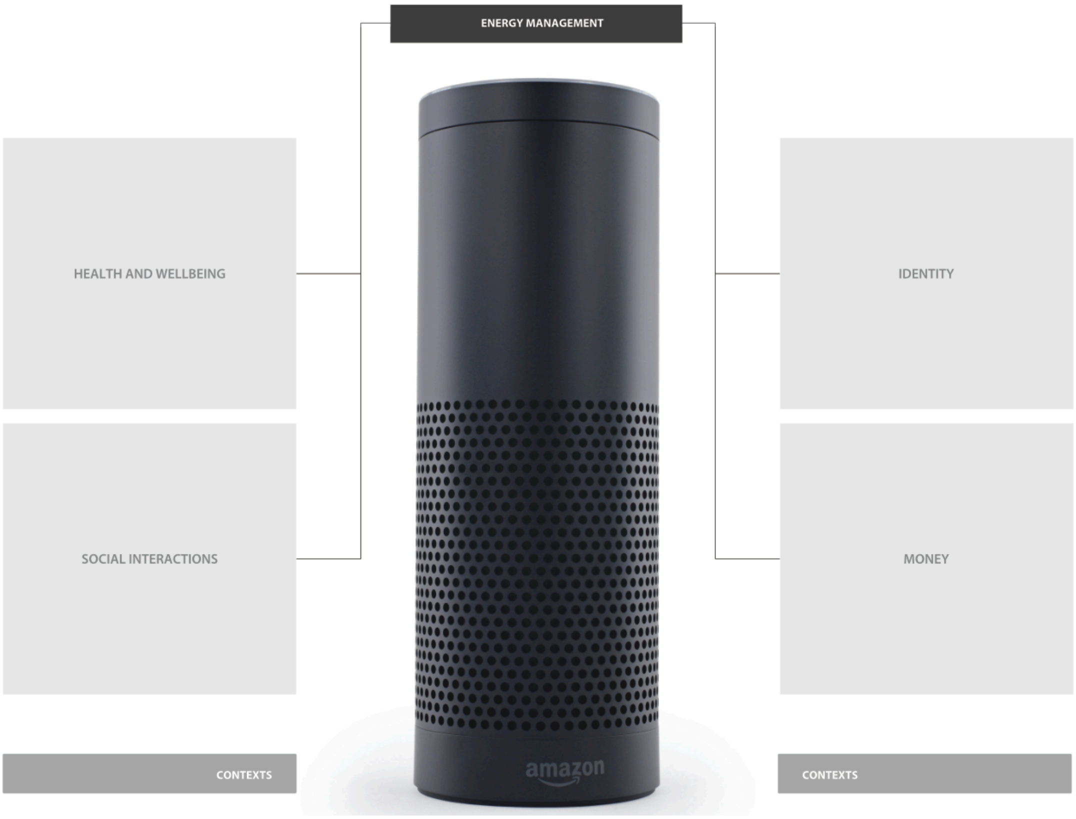
Type to enter a caption.



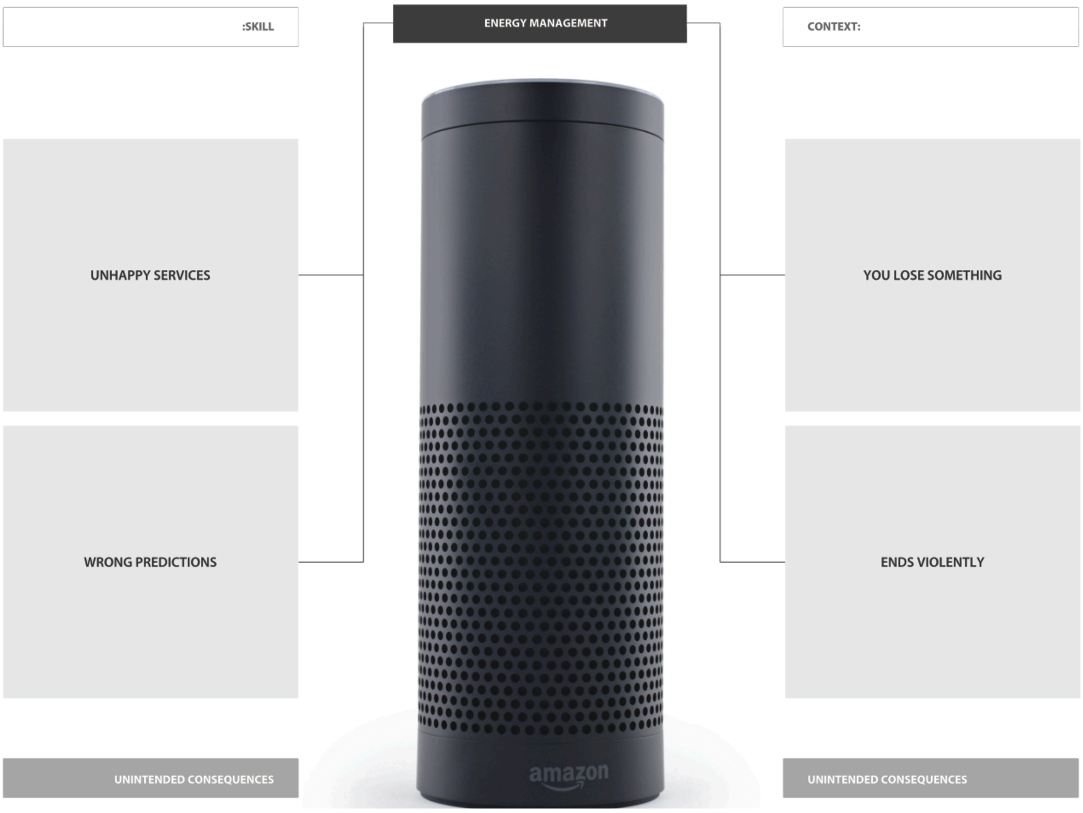
Exercise 1 - Present and future protections



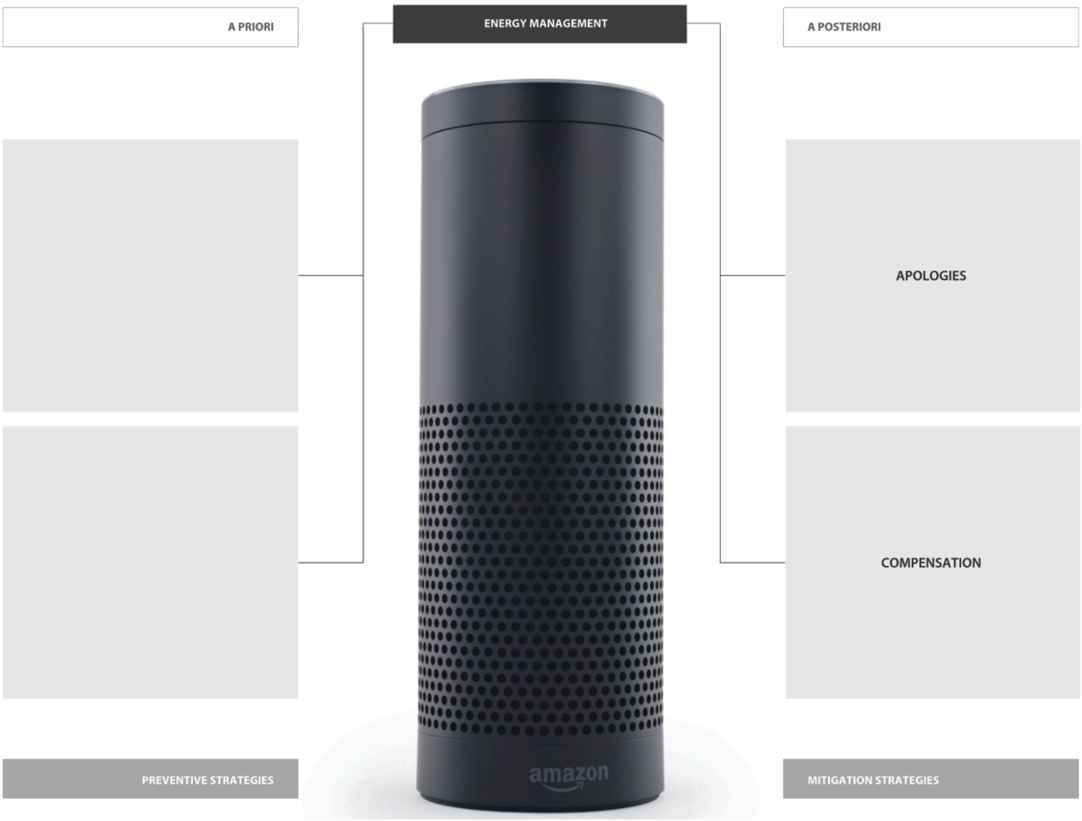
Exercise 2 - Consequential analysis



Exercise 3 - Highly sensitive area mapping



Exercise 4 - Actions mapping



Exercise 5 - Developing strategies

Type to enter a caption.

Type to enter a caption.

5

—

APPENDIX

WORKSHOP

3

2019

WORKSHOP VA

Fernando Galdon

Type to enter a caption.

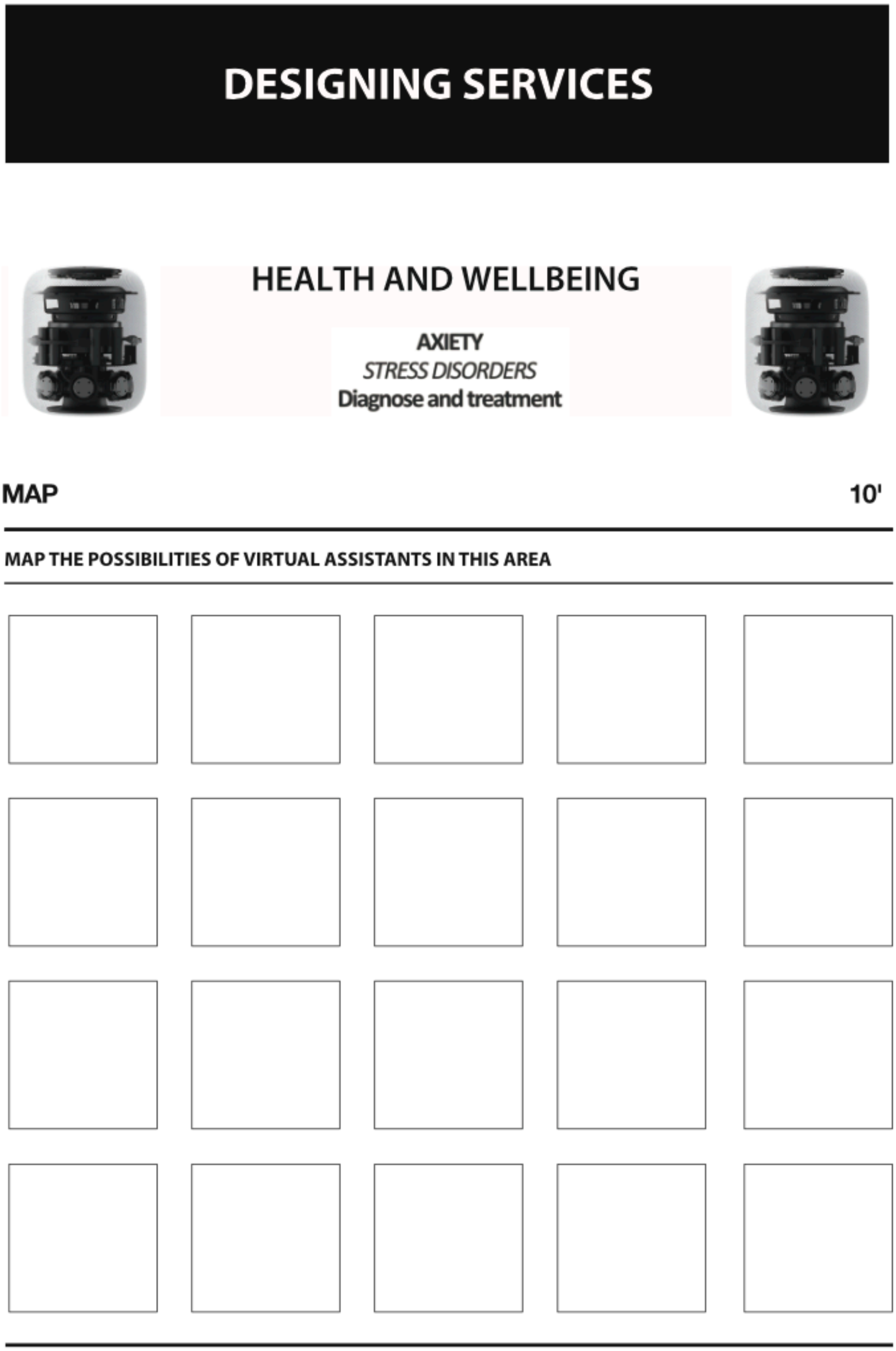
Type to enter a caption.

DESIGNING SERVICES

DESIGNING SERVICES
Health and Wellbeing

AXIETY
STRESS DISORDERS
Diagnose and treatment

Task 1 - No Framework



DESIGNING SERVICES

HEALTH AND WELLBEING

AXIETY

STRESS DISORDERS

Diagnose and treatment

MAP **10'**

MAP THE POSSIBILITIES OF VIRTUAL ASSISTANTS IN THIS AREA

Task 1 - Exercise 1

DESIGNING SERVICES

HEALTH AND WELLBEING

AXIETY

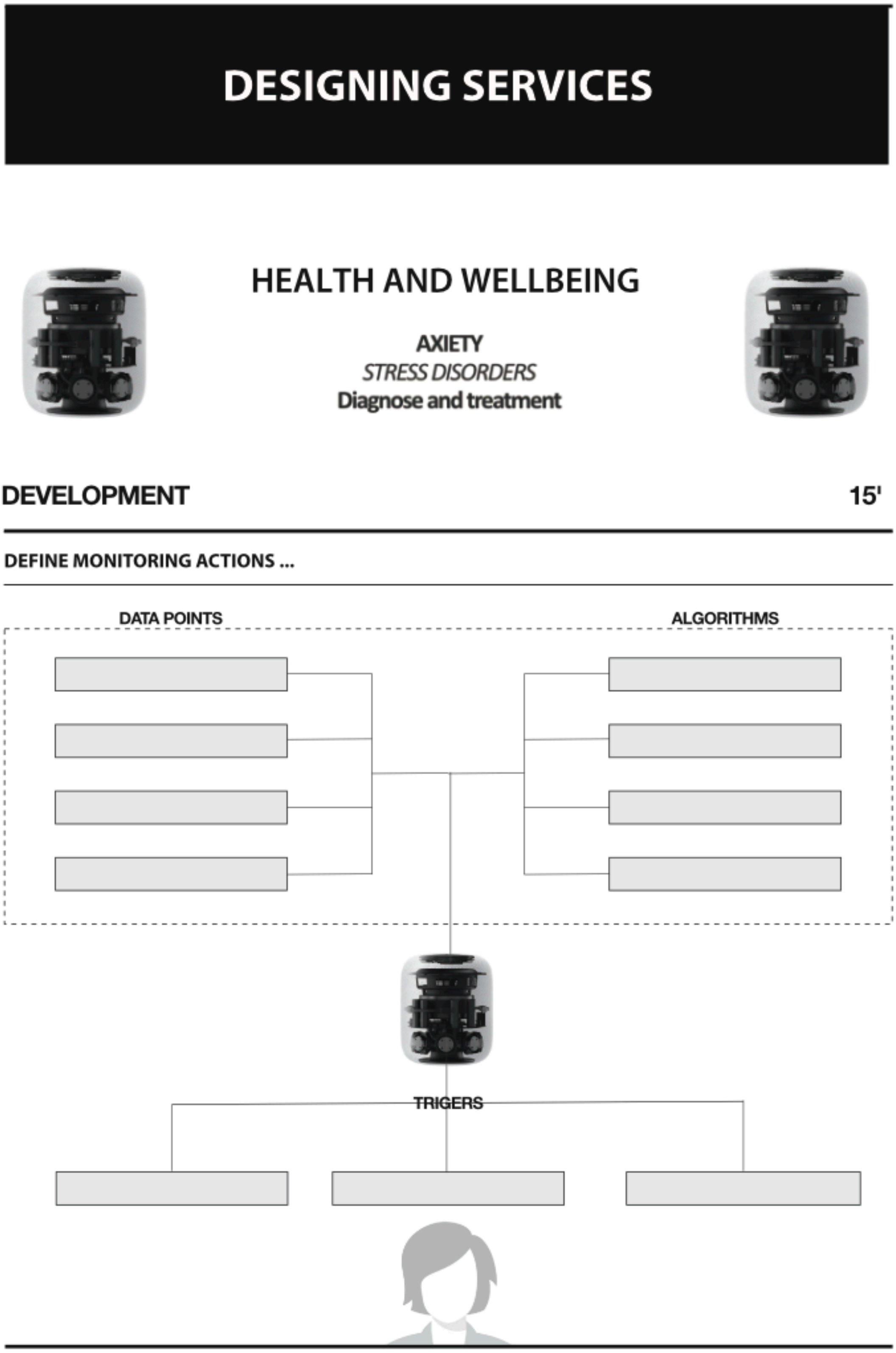
STRESS DISORDERS

Diagnose and treatment

- **10'**

DESIGN ...

Task 1 - Exercise 2



DESIGNING SERVICES

HEALTH AND WELLBEING

AXIETY

STRESS DISORDERS

Diagnose and treatment

DEVELOPMENT 15'

DEFINE MONITORING ACTIONS ...

Task 1 - Exercise 3

DESIGNING SERVICES

HEALTH AND WELLBEING

AXIETY
STRESS DISORDERS
Diagnose and treatment

DEVELOPMENT 15'

CUSTOMER JOURNEY MAP AND DESIGN INTERVENTIONS

PHASE 1 EXPECTATION Before the interaction	**PHASE 2** EXPERIMENTATION During the interaction	**PHASE 3** RELIABILITY After the interaction
CUSTOMER JOURNEY	CUSTOMER JOURNEY	CUSTOMER JOURNEY
DESIGN INTERVENTION	DESIGN INTERVENTION	DESIGN INTERVENTION

Task 1 - Exercise 4

DESIGNING SERVICES

DESIGNING SERVICES
Health and Wellbeing

ADDICTION
DRINKING DISORDERS
Diagnose and treatment

Task 2 - Specifications

DESIGNING SERVICES

HEALTH AND WELLBEING

ADDICTION
DRINKING DISORDERS
Diagnose and treatment

MAP **10'**

MAP THE POSSIBILITIES OF VIRTUAL ASSISTANTS IN THIS AREA

Task 2 - Exercise 1

DESIGNING SERVICES

HEALTH AND WELLBEING

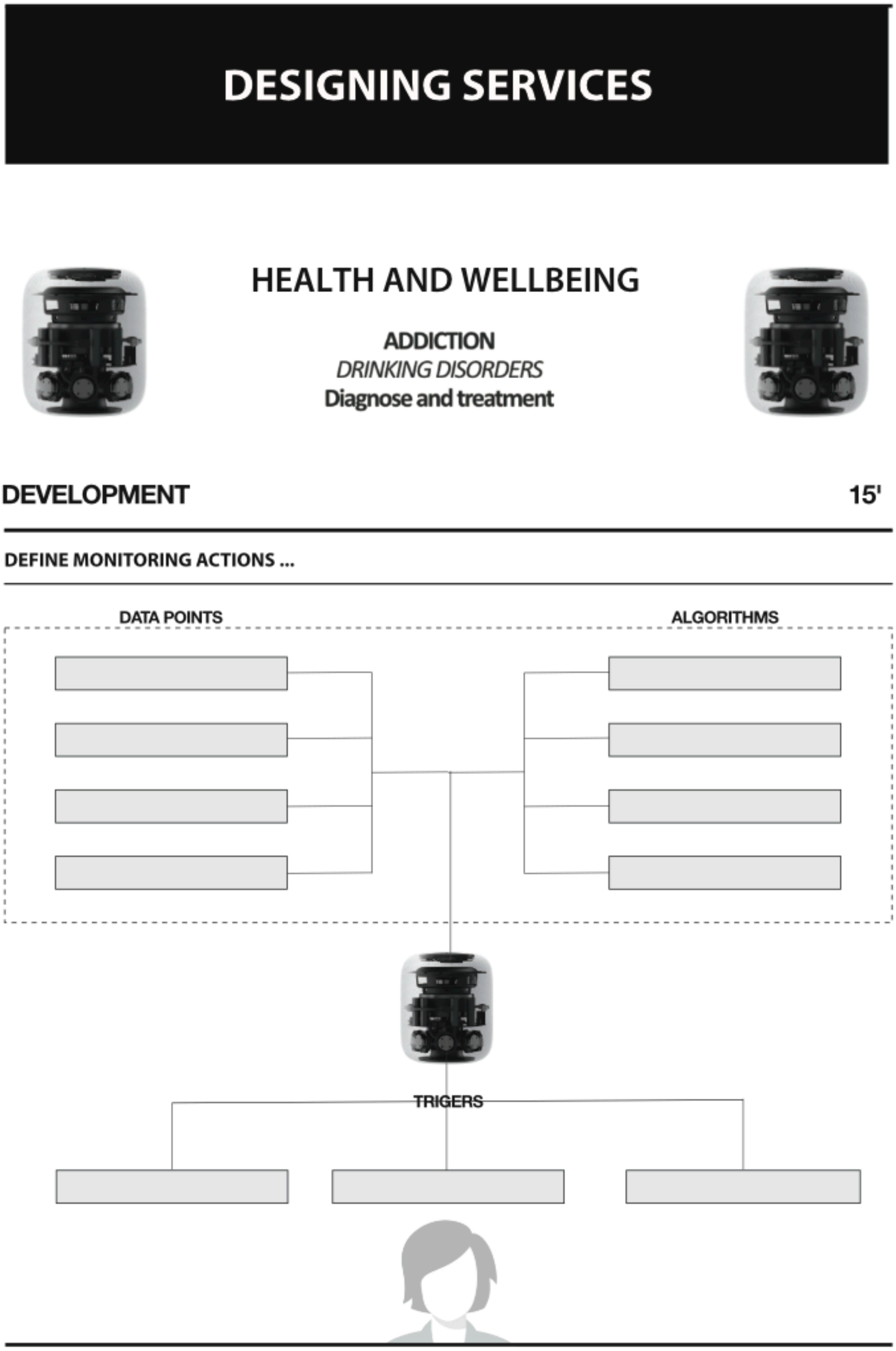
ADDICTION
DRINKING DISORDERS
Diagnose and treatment

SPECIFICATIONS 15'

DEFINE THE PURPOSE OF THE SYSTEM.

1. NAME	
2. DESCRIPTION	
3. CONTEXT	
4. AMBITIONS	
5. OBJECTIVES	
6. ATTITUDES & BEHAVIOUR	
7. GOALS	
8. CHALLENGES	

Task 2 - Exercise 2



DESIGNING SERVICES

HEALTH AND WELLBEING

ADDICTION

DRINKING DISORDERS

Diagnose and treatment

DEVELOPMENT 15'

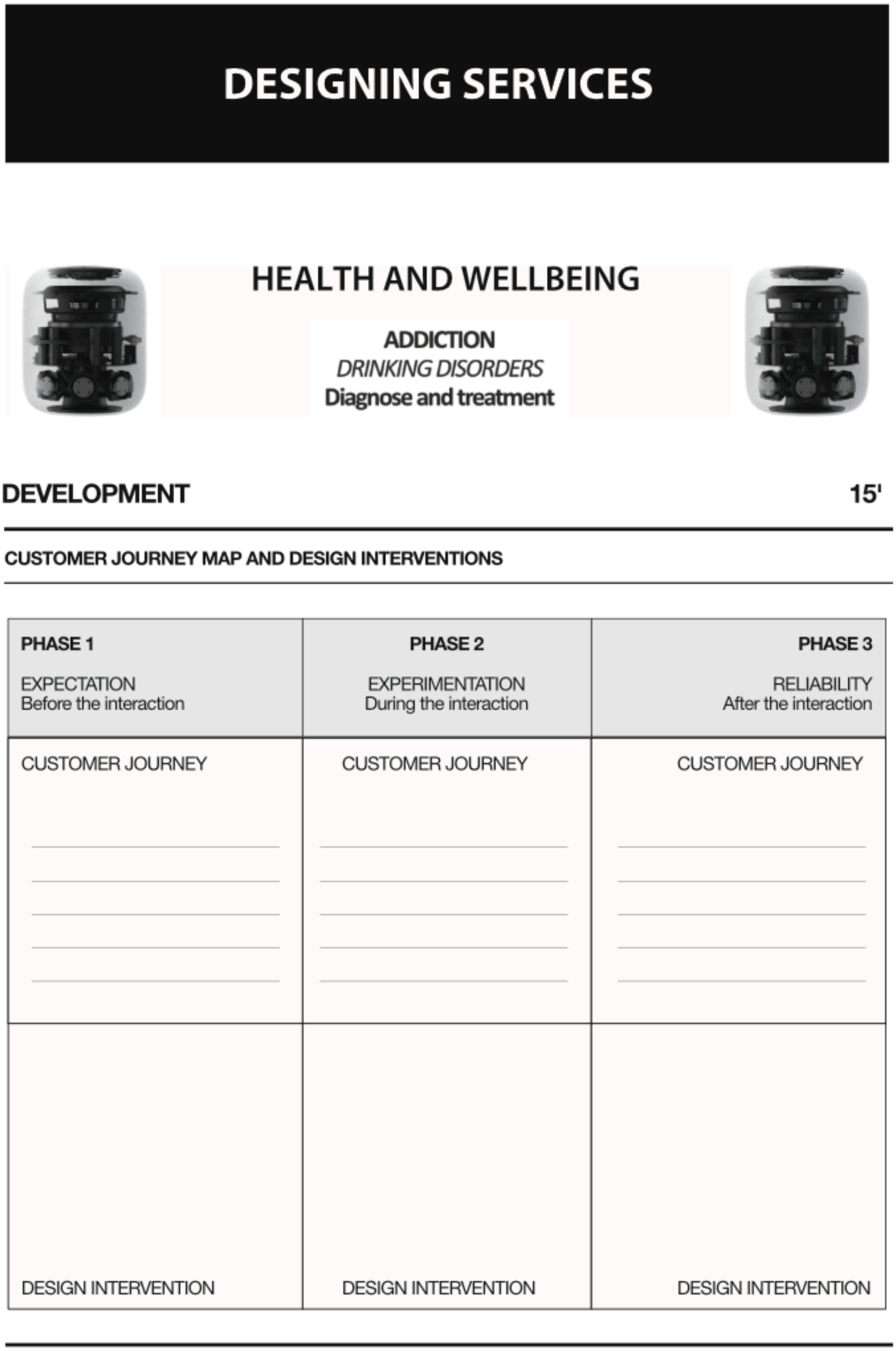
DEFINE MONITORING ACTIONS ...

DATA POINTS

ALGORITHMS

TRIGERS

Task 2 - Exercise 3



DESIGNING SERVICES

HEALTH AND WELLBEING

ADDICTION
DRINKING DISORDERS
Diagnose and treatment

DEVELOPMENT **15'**

CUSTOMER JOURNEY MAP AND DESIGN INTERVENTIONS

PHASE 1 EXPECTATION Before the interaction	**PHASE 2** EXPERIMENTATION During the interaction	**PHASE 3** RELIABILITY After the interaction
CUSTOMER JOURNEY	CUSTOMER JOURNEY	CUSTOMER JOURNEY
DESIGN INTERVENTION	DESIGN INTERVENTION	DESIGN INTERVENTION

Task 2 - Exercise 4

DESIGNING SERVICES

DESIGNING SERVICES
Health and Wellbeing

EXERCISE
OVERWEIGHT DISORDERS
Diagnose and treatment

DESIGNING SERVICES

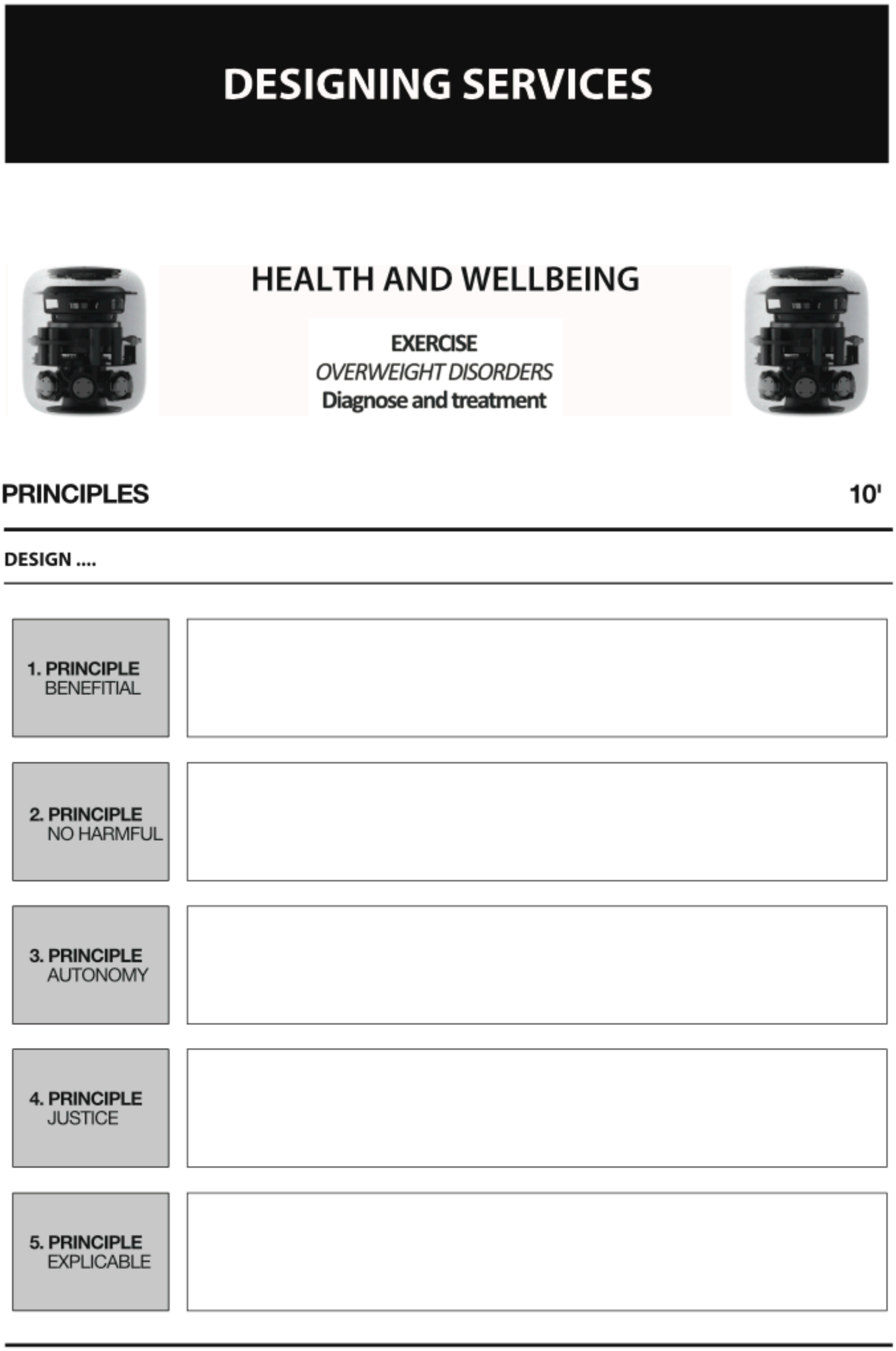
HEALTH AND WELLBEING

EXERCISE
OVERWEIGHT DISORDERS
Diagnose and treatment

MAP **10'**

MAP THE POSSIBILITIES OF VIRTUAL ASSISTANTS IN THIS AREA

Task 3 - Exercise 1



DESIGNING SERVICES

HEALTH AND WELLBEING

EXERCISE
OVERWEIGHT DISORDERS
Diagnose and treatment

PRINCIPLES **10'**

DESIGN ....

1. PRINCIPLE BENEFITIAL	
2. PRINCIPLE NO HARMFUL	
3. PRINCIPLE AUTONOMY	
4. PRINCIPLE JUSTICE	
5. PRINCIPLE EXPLICABLE	

Task 3 - Exercise 2

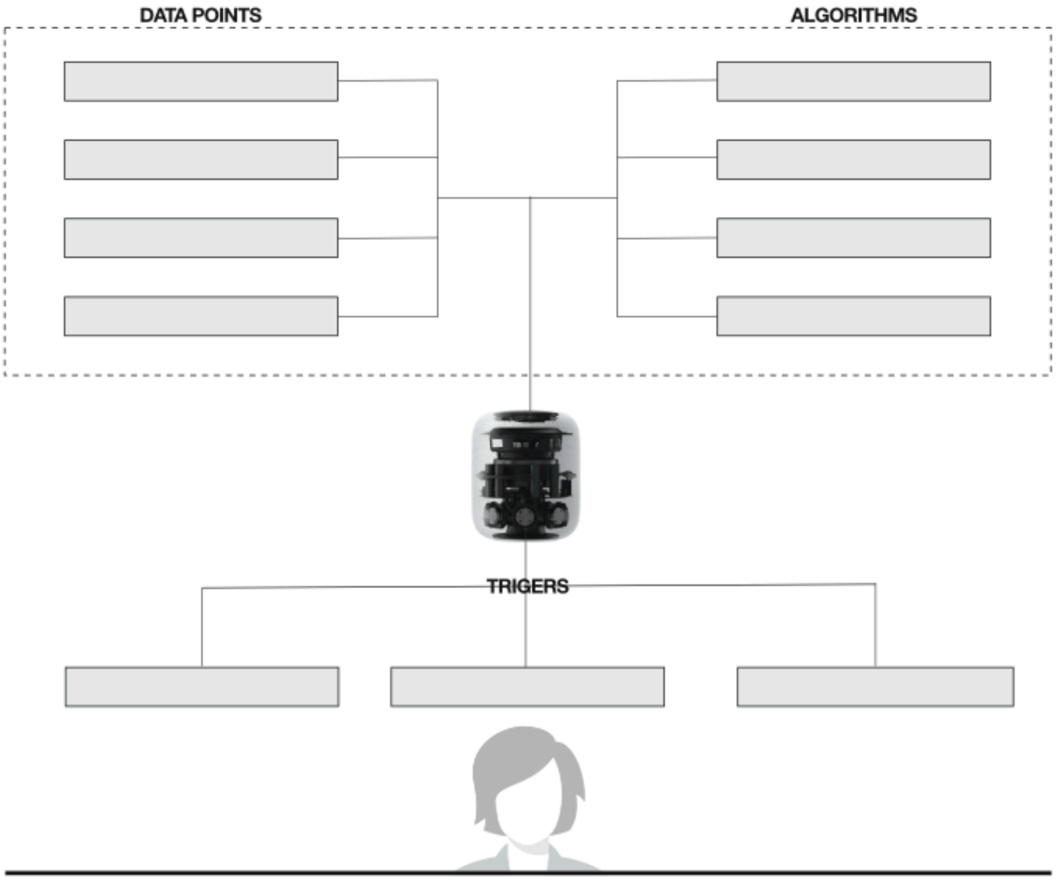
DESIGNING SERVICES

HEALTH AND WELLBEING

EXERCISE
OVERWEIGHT DISORDERS
Diagnose and treatment

DEVELOPMENT 15'

DEFINE MONITORING ACTIONS ...

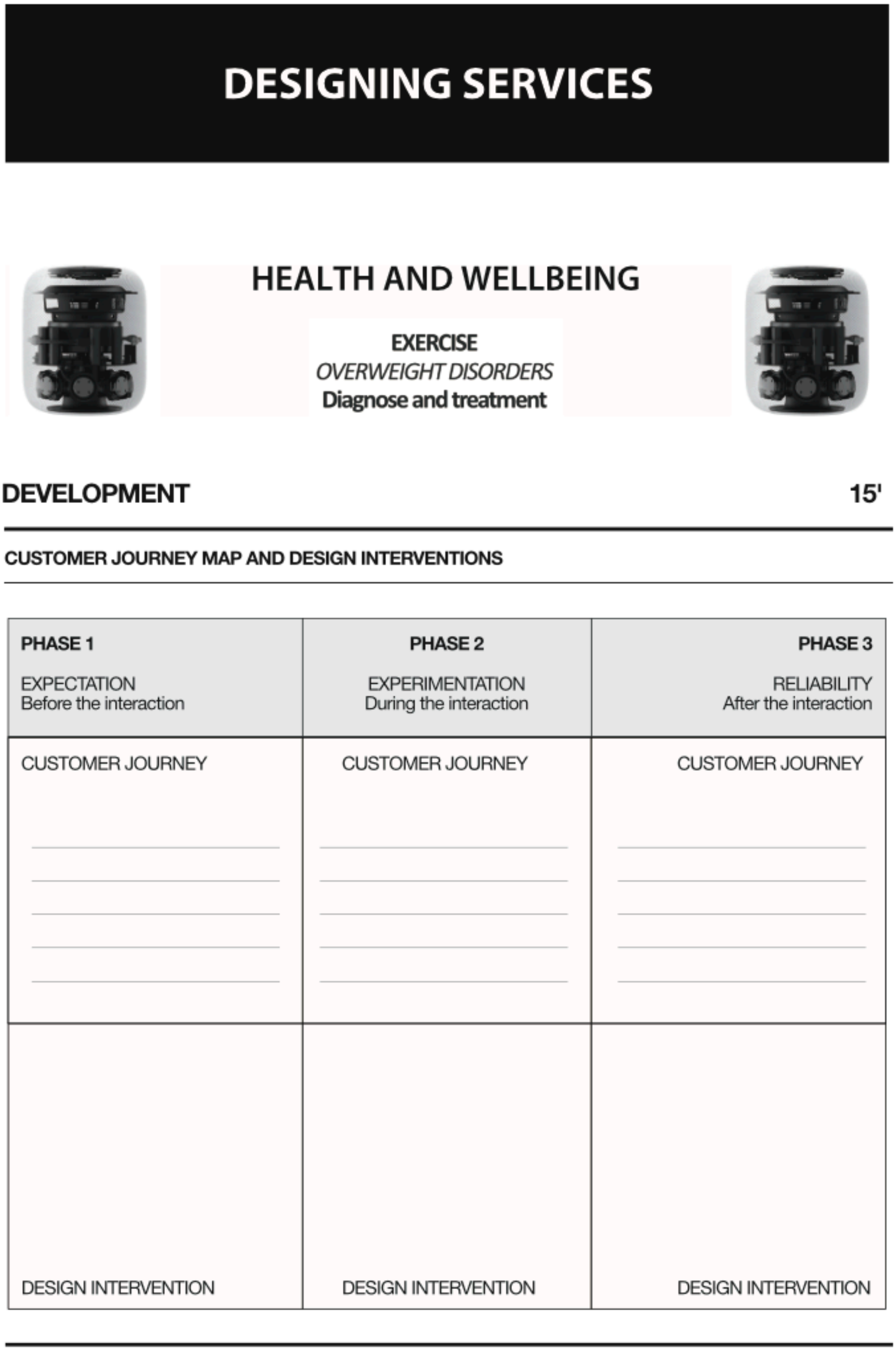


DATA POINTS

ALGORITHMS

TRIGERS

Task 3 - Exercise 3



DESIGNING SERVICES

HEALTH AND WELLBEING

EXERCISE
OVERWEIGHT DISORDERS
Diagnose and treatment

DEVELOPMENT **15'**

CUSTOMER JOURNEY MAP AND DESIGN INTERVENTIONS

PHASE 1 EXPECTATION Before the interaction	PHASE 2 EXPERIMENTATION During the interaction	PHASE 3 RELIABILITY After the interaction
CUSTOMER JOURNEY	CUSTOMER JOURNEY	CUSTOMER JOURNEY
DESIGN INTERVENTION	DESIGN INTERVENTION	DESIGN INTERVENTION

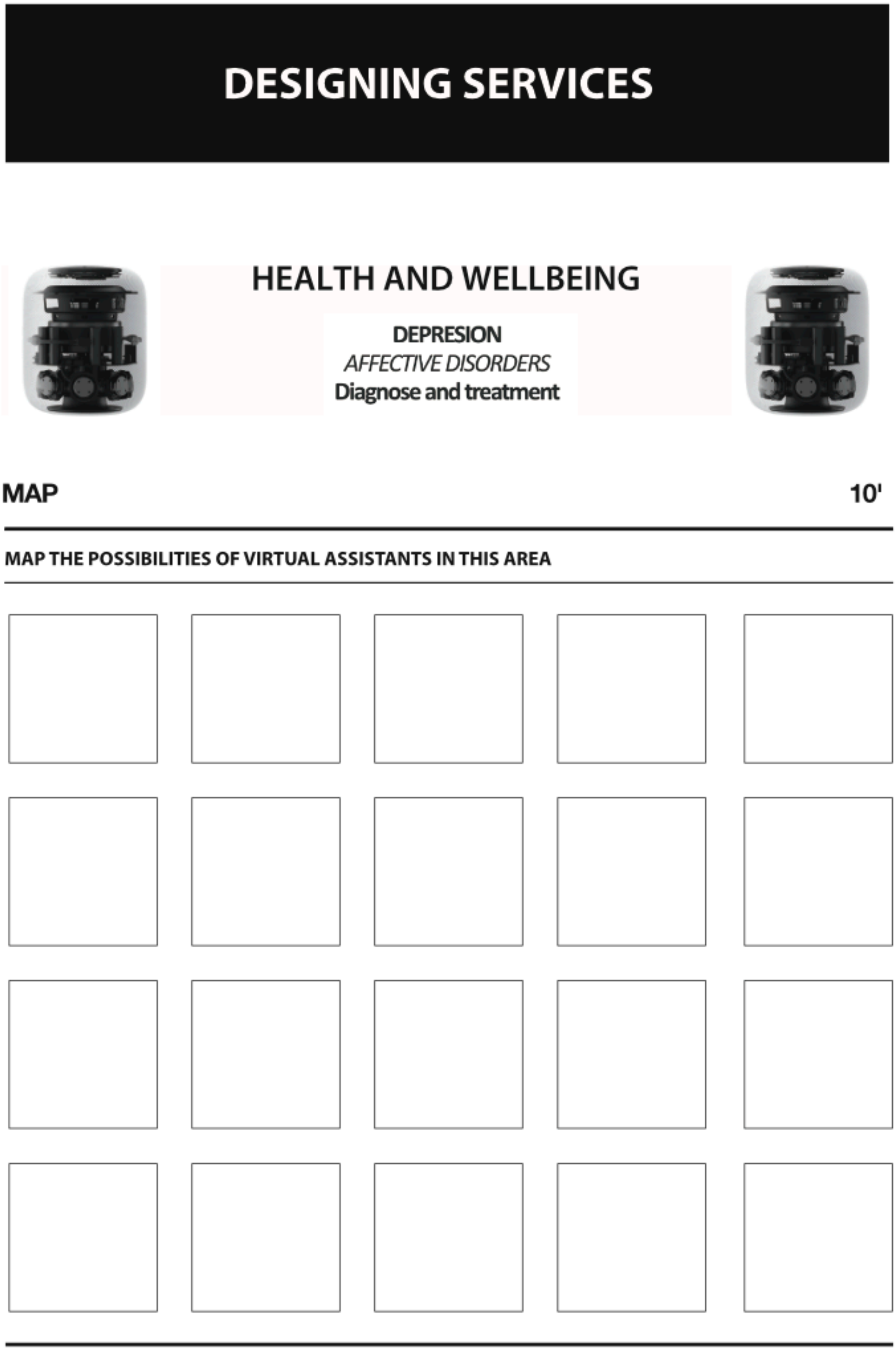
Task 3 - Exercise 4

DESIGNING SERVICES

DESIGNING SERVICES
Health and Wellbeing

DEPRESION
AFFECTIVE DISORDERS
Diagnose and treatment

Task 4 - Exercise 1



DESIGNING SERVICES

HEALTH AND WELLBEING

DEPRESION

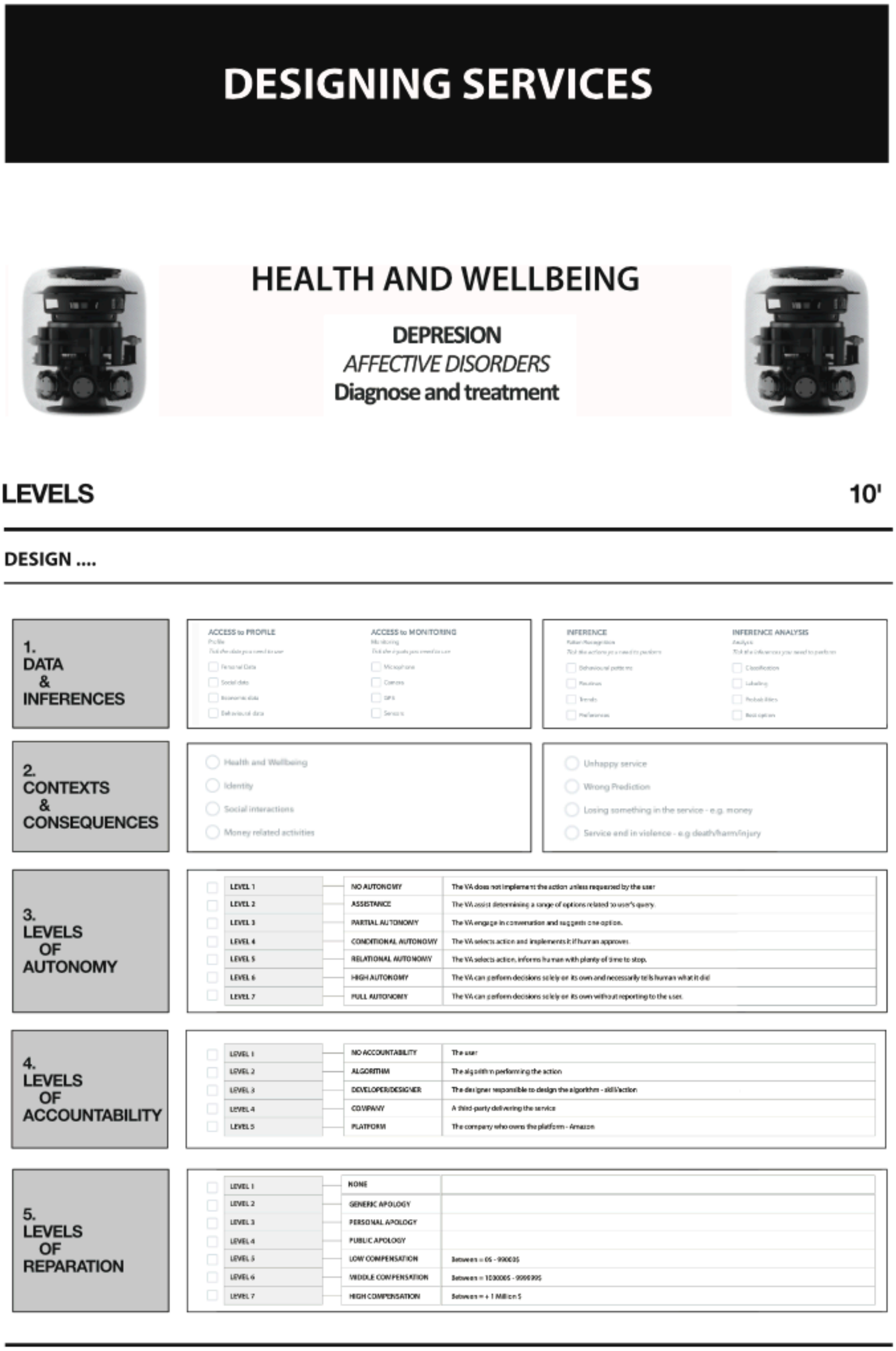
AFFECTIVE DISORDERS

Diagnose and treatment

MAP 10'

MAP THE POSSIBILITIES OF VIRTUAL ASSISTANTS IN THIS AREA

Task 4 - Exercise 1



DESIGNING SERVICES

HEALTH AND WELLBEING

DEPRESION

AFFECTIVE DISORDERS

Diagnose and treatment

LEVELS

10'

DESIGN ....

1. DATA & INFERENCES

ACCESS to PROFILE
Profile
Tick the data you need to use
- □ Personal Data
- □ Social data
- □ Economic data
- □ Behavioural data

ACCESS to MONITORING
Monitoring
Tick the inputs you need to use
- □ Microphone
- □ Camera
- □ GPS
- □ Sensors

INFERENCE
Pattern Recognition
Tick the actions you need to perform
- □ Behavioural patterns
- □ Routines
- □ Trends
- □ Preferences

INFERENCE ANALYSIS
Analysis
Tick the inferences you need to perform
- □ Classification
- □ Labeling
- □ Probabilities
- □ Best option

2. CONTEXTS & CONSEQUENCES

- ○ Health and Wellbeing
- ○ Identity
- ○ Social interactions
- ○ Money related activities

- ○ Unhappy service
- ○ Wrong Prediction
- ○ Losing something in the service - e.g. money
- ○ Service end in violence - e.g death/harm/injury

3. LEVELS OF AUTONOMY

□	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.

4. LEVELS OF ACCOUNTABILITY

□	LEVEL 1	NO ACCOUNTABILITY	The user
□	LEVEL 2	ALGORITHM	The algorithm performing the action
□	LEVEL 3	DEVELOPER/DESIGNER	The designer responsible to design the algorithm - skill/action
□	LEVEL 4	COMPANY	A third-party delivering the service
□	LEVEL 5	PLATFORM	The company who owns the platform - Amazon

5. LEVELS OF REPARATION

□	LEVEL 1	NONE	
□	LEVEL 2	GENERIC APOLOGY	
□	LEVEL 3	PERSONAL APOLOGY	
□	LEVEL 4	PUBLIC APOLOGY	
□	LEVEL 5	LOW COMPENSATION	Between = 0$ - 99000$
□	LEVEL 6	MIDDLE COMPENSATION	Between = 100000$ - 999999$
□	LEVEL 7	HIGH COMPENSATION	Between = + 1 Million $

Task 4 - Exercise 2 - Calibrated using the calculator bellow.

DESIGNING SERVICES

HEALTH AND WELLBEING

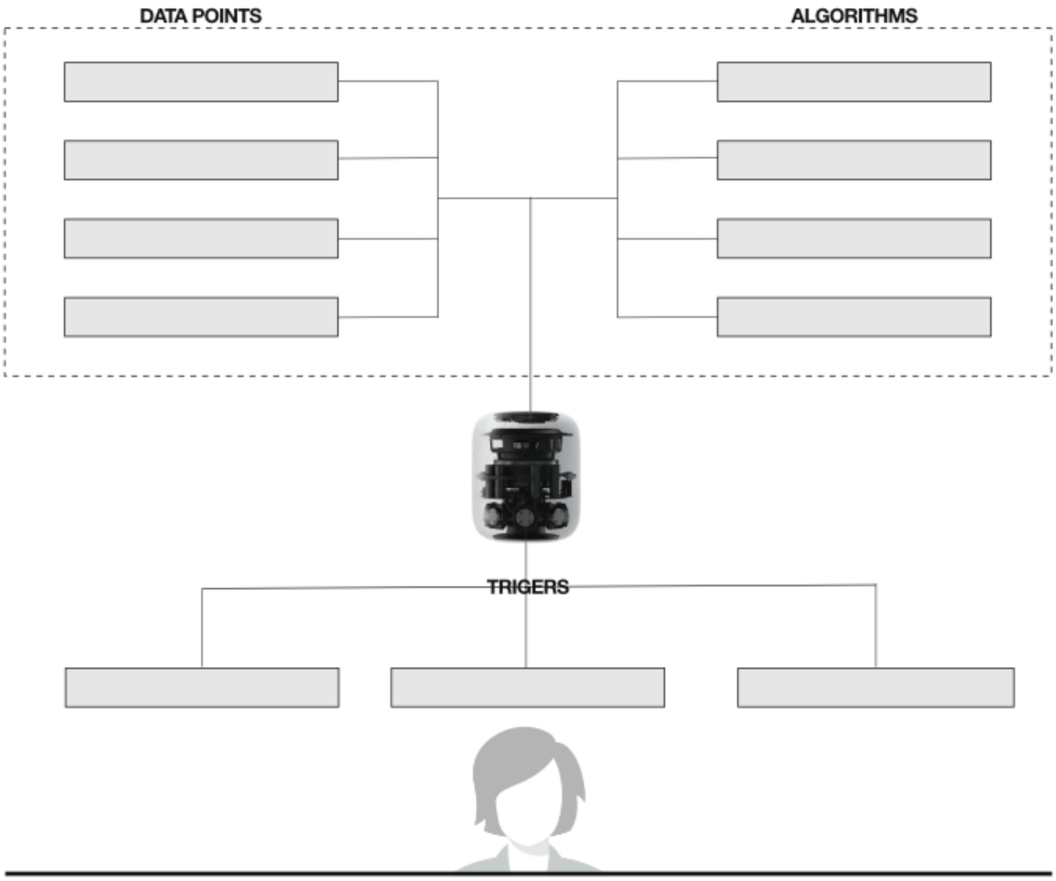
DEPRESION

AFFECTIVE DISORDERS

Diagnose and treatment

DEVELOPMENT **15'**

DEFINE MONITORING ACTIONS ...

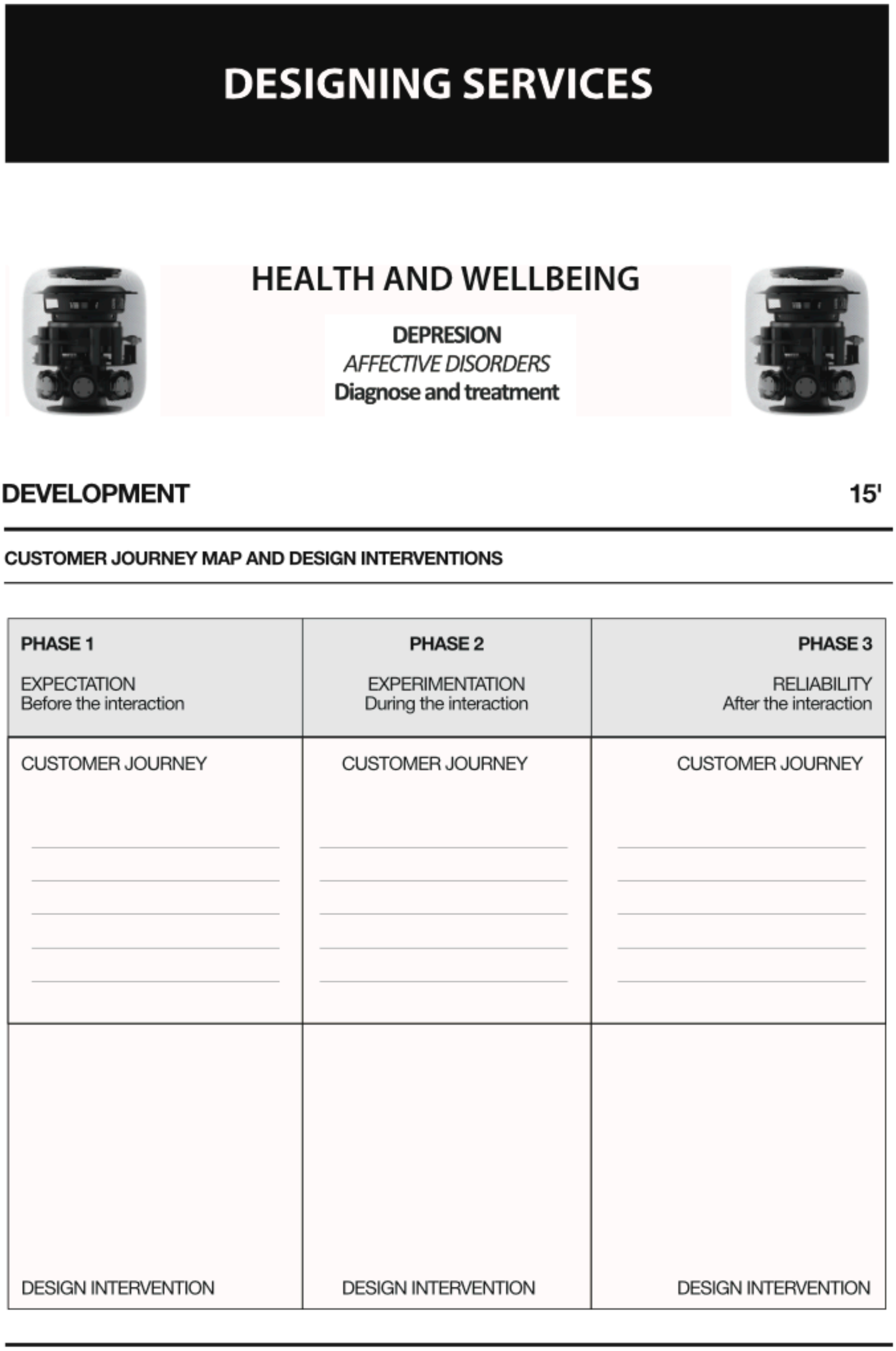


DATA POINTS

ALGORITHMS

TRIGERS

Task 4 - Exercise 3



DESIGNING SERVICES

HEALTH AND WELLBEING

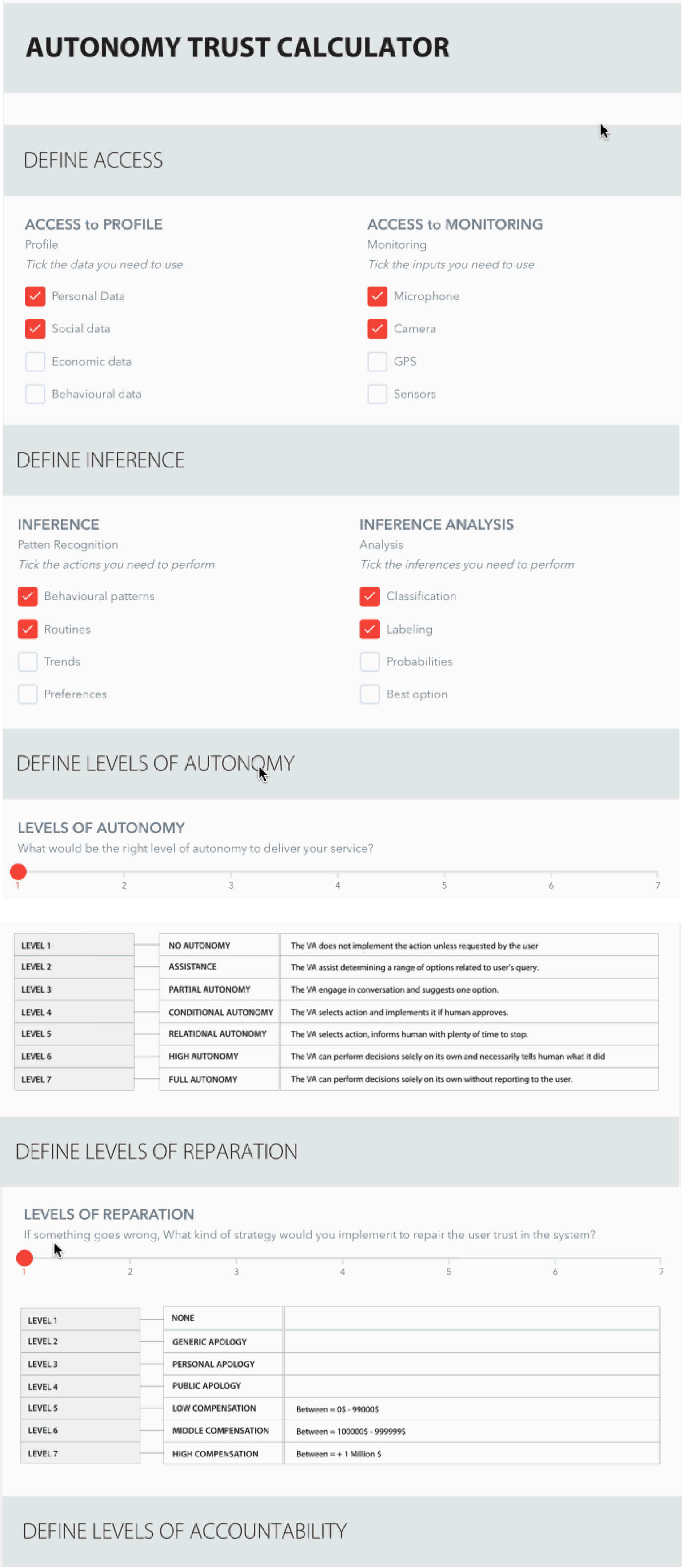
DEPRESION
AFFECTIVE DISORDERS
Diagnose and treatment

DEVELOPMENT **15'**

CUSTOMER JOURNEY MAP AND DESIGN INTERVENTIONS

PHASE 1 EXPECTATION Before the interaction	**PHASE 2** EXPERIMENTATION During the interaction	**PHASE 3** RELIABILITY After the interaction
CUSTOMER JOURNEY ________________ ________________ ________________ ________________ ________________	CUSTOMER JOURNEY ________________ ________________ ________________ ________________ ________________	CUSTOMER JOURNEY ________________ ________________ ________________ ________________ ________________
DESIGN INTERVENTION	DESIGN INTERVENTION	DESIGN INTERVENTION

Task 4 - Exercise 4



AUTONOMY TRUST CALCULATOR

DEFINE ACCESS

ACCESS to PROFILE

Profile

Tick the data you need to use

☑ Personal Data

☑ Social data

□ Economic data

□ Behavioural data

ACCESS to MONITORING

Monitoring

Tick the inputs you need to use

☑ Microphone

☑ Camera

□ GPS

□ Sensors

DEFINE INFERENCE

INFERENCE

Patten Recognition

Tick the actions you need to perform

☑ Behavioural patterns

☑ Routines

□ Trends

□ Preferences

INFERENCE ANALYSIS

Analysis

Tick the inferences you need to perform

☑ Classification

☑ Labeling

□ Probabilities

□ Best option

DEFINE LEVELS OF AUTONOMY

LEVELS OF AUTONOMY

What would be the right level of autonomy to deliver your service?

● 1 2 3 4 5 6 7

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.

DEFINE LEVELS OF REPARATION

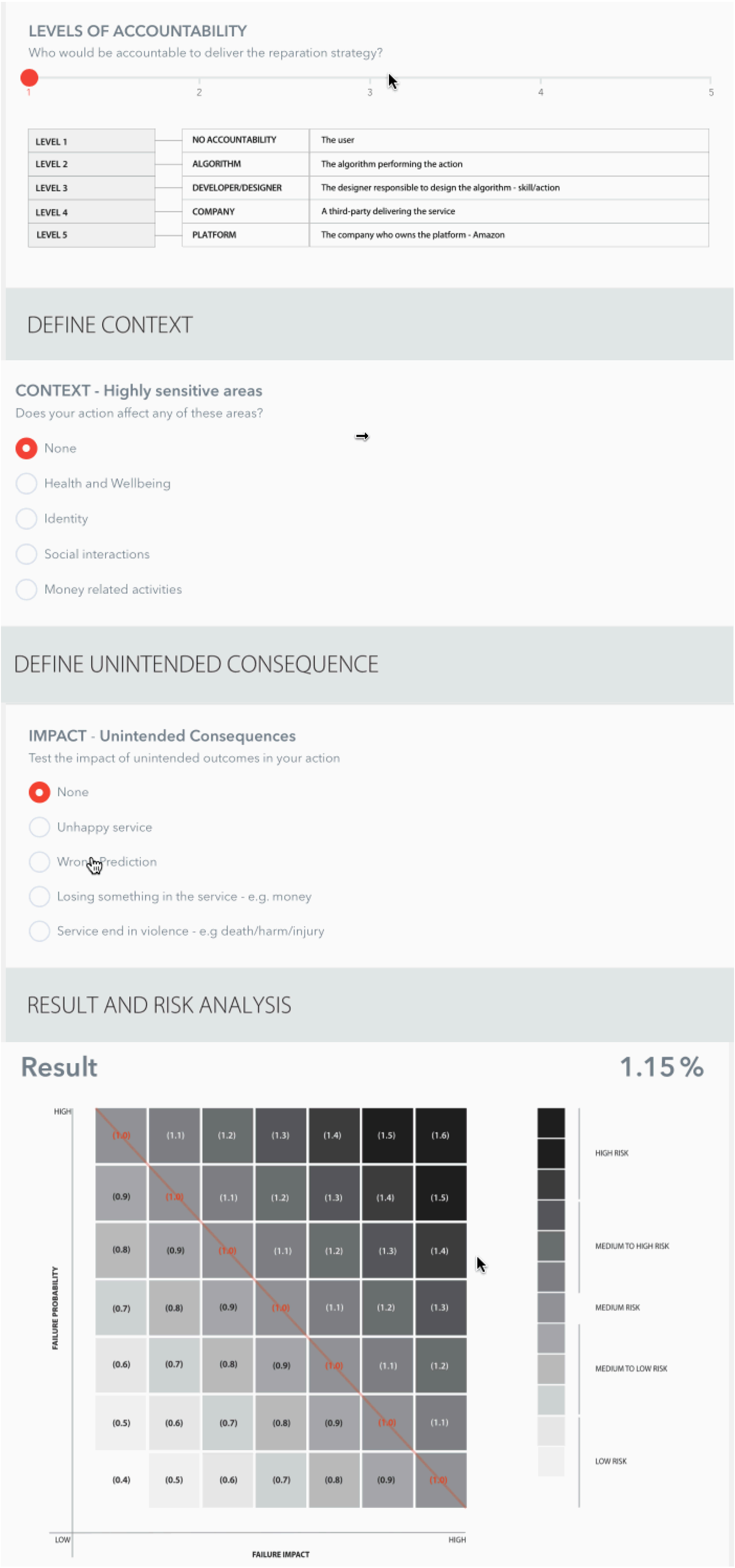
LEVELS OF REPARATION

If something goes wrong, What kind of strategy would you implement to repair the user trust in the system?

● 1 2 3 4 5 6 7

LEVEL 1	NONE	
LEVEL 2	GENERIC APOLOGY	
LEVEL 3	PERSONAL APOLOGY	
LEVEL 4	PUBLIC APOLOGY	
LEVEL 5	LOW COMPENSATION	Between = 0$ - 99000$
LEVEL 6	MIDDLE COMPENSATION	Between = 100000$ - 999999$
LEVEL 7	HIGH COMPENSATION	Between = + 1 Million $

DEFINE LEVELS OF ACCOUNTABILITY



LEVELS OF ACCOUNTABILITY

Who would be accountable to deliver the reparation strategy?

● 1 — 2 — 3 — 4 — 5

LEVEL 1	NO ACCOUNTABILITY	The user
LEVEL 2	ALGORITHM	The algorithm performing the action
LEVEL 3	DEVELOPER/DESIGNER	The designer responsible to design the algorithm - skill/action
LEVEL 4	COMPANY	A third-party delivering the service
LEVEL 5	PLATFORM	The company who owns the platform - Amazon

DEFINE CONTEXT

CONTEXT - Highly sensitive areas

Does your action affect any of these areas?

- ● None
- ○ Health and Wellbeing
- ○ Identity
- ○ Social interactions
- ○ Money related activities

DEFINE UNINTENDED CONSEQUENCE

IMPACT - Unintended Consequences

Test the impact of unintended outcomes in your action

- ● None
- ○ Unhappy service
- ○ Wrong Prediction
- ○ Losing something in the service - e.g. money
- ○ Service end in violence - e.g death/harm/injury

RESULT AND RISK ANALYSIS

Result 1.15 %

(1.0)	(1.1)	(1.2)	(1.3)	(1.4)	(1.5)	(1.6)
(0.9)	(1.0)	(1.1)	(1.2)	(1.3)	(1.4)	(1.5)
(0.8)	(0.9)	(1.0)	(1.1)	(1.2)	(1.3)	(1.4)
(0.7)	(0.8)	(0.9)	(1.0)	(1.1)	(1.2)	(1.3)
(0.6)	(0.7)	(0.8)	(0.9)	(1.0)	(1.1)	(1.2)
(0.5)	(0.6)	(0.7)	(0.8)	(0.9)	(1.0)	(1.1)
(0.4)	(0.5)	(0.6)	(0.7)	(0.8)	(0.9)	(1.0)

GLOSSARY

HAS	Highly Automated Systems
VA	Virtual Assistants
HMI	Human-Machine Interactions
MHI	Machine Human Interactions
MMI	Machine-Machine Interactions
ML	Machine Learning
DL	Deep Learning
HAuS	Highly Autonomous Systems
HAI	Human-Automation-Interactions
LoA	Levels of Automation
LoT	Levels of Trust
CSD	Critical and Speculative Design
SD	Speculative Design
CoS	Co-Speculative Design
TD	Transition Design
FP	Forecasting Planning
ABCD	ABCD Planning
SP	Scenario Planing