TU Delft

Delft University of Technology

Document Version
Final published version

Citation (APA)
Stoimenova, N. (2023). *Designing AI systems*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:c53b6902-f347-4694-9702-6bc167fb8c04

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
In case the licence states “Dutch Copyright Act (Article 25fa)”, this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.

NIYA STOIMENOVA



DESIGNING AI SYSTEMS

Designing AI Systems

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus prof.dr.ir. T.H.J.J. van der Hagen
Chair of the Board for Doctorates
to be defended publicly on
Monday 9 October 2023 at 17:30 o'clock

by

Niya Evtimova STOIMENOVA
Master of Science in Strategic Product Design,
Delft University of Technology, the Netherlands

born in Petrich, Bulgaria

This dissertation has been approved by the promotor and copromotor

Composition of the doctoral committee:

Rector Magnificus	chairperson
Prof. dr. ir. M.S. Kleinsmann	Delft University of Technology, promotor
Dr. M.G. Gonçalves	Delft University of Technology, copromotor

Independent members:

Prof. dr. ir. A. Bozzon	Delft University of Technology
Prof. dr. P. A. Lloyd	Delft University of Technology
Prof. dr. S. Roeser	Delft University of Technology
Prof. dr. Ph. Cash	Northumbria University, UK
Prof. dr. A. Dong	Oregon State University, USA

keywords: abduction, AI systems, unintended consequences, reasoning, behaviour-use interdependence, design theory, design cognition
ISBN: 978-94-6366-742-5
©Copyright Niya Stoimenova, 2023

на семейството ми
to my family

Table of Contents

SUMMARY: DESIGNING AI SYSTEMS **IX**

Part I: Set the scene xi

Part II: Initial theoretical model xii

Part III: Extended theoretical model xiii

PREFACE: EVIL MICE, SISYPHUS, AND THE TRUTH **2**

Evil mice 2

Sisyphus 3

The truth 3

The existential crisis 4

INTRODUCTION: AI AND COMPLEX CONTEXTS **7**

AI and complex contexts 10

The initial research question 12

Research method 12

Dissertation outline 15

PART I: SET THE SCENE

CHAPTER 1: CHARACTERISTICS OF CONTEMPORARY AI SYSTEMS **25**

Four research directions to intelligence 26

Machine Learning 29

Characteristics of contemporary AI systems 34

Conclusion 35

CHAPTER 2: CHALLENGES FOR IMPLEMENTING AI SYSTEMS INTO COMPLEX CONTEXTS 37

The four principles for “good” AI 39

Transparency principle 40

Fairness principle 42

Responsibility principle 44

Non-maleficence principle 47

Main research question of the dissertation 49

PART II: INITIAL THEORETICAL MODEL

CHAPTER 3: THE FUNDAMENTS OF DESIGN THEORY 57

Section I: The two defining paradigms of Design Theory 58

Section II: The core of design 64

Conclusion 72

CHAPTER 4: PROTOTYPING FOR EARLY SIMULATION OF BEHAVIOUR AND USE 75

The design practice that supports early simulation 76

Using prototypes to simulate behaviour and use 78

Discussion 83

Prototypes and abduction 84

Conclusion 85

CHAPTER 5: INITIAL THEORETICAL MODEL 87

Theoretical model 89

A note on values and prototypes 94

Conclusion 96

PART III: EXTENDED THEORETICAL MODEL

CHAPTER 6: EARLY SIMULATION OF AI SYSTEM'S BEHAVIOUR-USE INTERDEPENDENCE 101

Example's background 103

Design process 104

Mapping the design process onto the theoretical model 108

Discussion 112

Conclusion and further research 117

CHAPTER 7: CASE STUDY - DESIGNING A SMART IN-CAR SYSTEM CONCEPT FOR A LARGE AUTOMAKER 121

Case background 123

Data collection 124

Data analysis 126

Results 133

Discussion 149

Conclusion 156

CHAPTER 8: FINAL THEORETICAL MODEL 159

Domain 161

Variables 161

Relationships 162

Predictions and future research 171

Good theory criteria 176

The implications for Design(ers) 178

EPILOGUE: BEYOND DESIGNING CONCEPTS FOR AI SYSTEMS 185

The strive towards general AI 188

“Only stupid people do a lot of calculations.” 192

APPENDIX

CHAPTER 6 197

CHAPTER 7 201

REFERENCES 206

ACKNOWLEDGEMENTS 229

SAMENVATTING 239

CURRICULUM VITAE 247

SUMMARY

DESIGNING AI SYSTEMS

Since its inception, the field of AI has been focused on devising systems that can provide clear answers. AI systems can tell us how to move a chess piece, translate a word, fold a protein, and predict whether a person would buy a book. These systems have a clear objective, clear outcome, and in some cases, clear reward, and punishment functions, too. The answers they provide can easily be classified as correct or incorrect. After a chess piece has been moved, one can easily check whether the move was correct and has increased the odds of winning. These characteristics have propelled the adoption of AI systems in areas such as modes of manufacturing and transportation, the way people receive information, select movies and songs, the way they date, trade at the stock exchange market, and the way social institutions such as hospitals, banks, police departments and courtrooms make decisions (e.g., Rahwan et al., 2019).

However, once AI systems that need clear rules and objectives face the complexity of social contexts, they start to produce unintended and sometimes harmful outcomes (Rudin, 2019). For instance, an AI system used by the Dutch Tax Authorities led to the false persecution of thousands of families and the resignation of the entire government in 2021 (Hanley, 2021). AI systems

used by big technological companies such as Google, Meta and Twitter (now X) frequently mislabel black people as e.g., primates (BBC, 2021b). When the Covid-19 Pandemic hit, the system Zillow used to estimate house prices resulted in more than $300 million loss for the company and the firing of 2000 people (CNN, 2021). In such situations, neither the problems nor the objectives can be clearly defined. Further, instead of true or false answers, there is a multitude of potential solutions that can only be classified as good or bad. Therefore, if we are to design AI systems that do not produce far-reaching harmful consequences, we *cannot decouple/detach* them from the complex systems in which we are to embed them.

Historically, the field of Design developed as a response to large changes in society facilitated by rapid technological advancements (Calabretta & Kleinsmann, 2017). In the span of a century, Design gradually evolved from designing products to designing human-computer interactions, product-service-systems, and recently, complex sociotechnical systems (Norman & Stappers, 2015). This expansion of the field sparked an on-going debate whether Design theories should be adapted to address the increasingly complex contexts in which designers operate (e.g., Voûte et al., 2020). The debate should be extended to include the design of ***AI systems that are to be implemented into larger complex contexts***, too. Hence, the initial research question of this dissertation:

"How can Design theories support the design and implementation of AI systems into complex contexts?"

It is against this backdrop that we start the theoretical and empirical exploration presented in the dissertation. We follow the theory building part of the theory building/testing cycle proposed by Cash (2018). Namely, we address its first three stages: ***Discovery and description*** (i.e., detailing the important issues around which the theory is built), ***Definitions of variables and limitation of domain*** (i.e., presenting variables and their definitions as well as where and when a theory is to be applied), and ***Relationship building*** (i.e., describing the

conceptual relationships among the identified variables) (Cash, 2018). These stages are reflected into the three parts of the dissertation.

Part I: Set the scene

The wide-spread adoption of AI systems is a result of the significant progress in the field over the past decade (Rahwan et al., 2019). For instance, deep learning models can predict lung cancer with a 94.4% accuracy (Ardila et al., 2019), perform better than radiologists in detecting pneumonia (Rajpurkar et al., 2017), detect hypertrophic cardiomyopathy in asymptomatic patients (Green et al., 2019) and forecast the aftershock locations of earthquakes (DeVries et al., 2018). Machine Learning, which trains models through examples of desired input-output behaviour, fuels most of these achievements.

While there are various types of AI systems, they share common characteristics: (1) their decision-making processes are complex and not easily understood, (2) they are heavily dependent on human-generated data, and (3) they continuously learn from and adapt to human behaviour. These characteristics make it difficult to ensure AI systems are *transparent*, *fair*, *responsible*, and *non-maleficent* (Jobin et al., 2019). Consequently, there are four main challenges for implementing AI systems into existing contexts: (1) ensuring transparency of the innerworkings of AI systems, (2) mitigating implicit biases in the data used to train and retrain these systems, (3) the difficulty to align the behaviour of AI systems with human values, and (4) addressing the behaviour-use interdependence of these systems.

While research approaches exist for the first three challenges, the challenge of addressing the AI systems' behaviour-use interdependence has not been fully addressed in extant literature. Design theories are well-positioned to tackle it due to the central role both behaviour and use play in the core of Design reasoning – innovative abduction (Roozenburg, 1993). These insights serve as a way for us to reframe the main research question into:

"How can a theoretical model be designed that supports the early simulation of AI systems' behaviour-use interdependence by utilising Design theories?"

Part II: Initial theoretical model

We begin our exploration with a theoretical investigation into Design theories that could aid the early simulation of AI system's behaviour-use interdependence. First, we examine the paradigms that have largely defined the majority of Design theories – Rational Problem Solving (Simon, 1996) and Reflective Practice (Schön, 1983). Despite their differences, both paradigms contend that (1) establishing a feedback mechanism is instrumental to achieving a better understanding of the problem; (2) the designer has influence over the design process; and (3) unintended outcomes are a natural and needed result of the design process since they propel it forward.

Second, we explore the Design theory of innovative abduction which provides formally defined relationships between the variables of behaviour and use. It also introduces the manners in which the two can be used to design new solutions. Two widely-agreed upon models exist that define the relationship – one introduced by Roozenburg (1993) and one by Dorst (2011). Despite their differences, they both start from (1) an initially agreed-upon starting point (*purpose* and *value*), (2) which they use to define the behaviour and use of the potential solution (*mode of action* and *actuation*, and *how*), and (3) the combination of these leads to a tangible solution – either a *form* or an *object, service, or a system*.

Third, prototypes can support the continuous simulation of the behaviour and use of a solution that is to be implemented into complex contexts. In fact, they (1) support us in observing the different types of outcomes and uses the behaviour can uncover; (2) serve as a bridge between behaviour and use; and (3) enable what Magnani (2007) terms manipulative abduction. We illustrate

these insights with two examples from a project conducted between November 2015 and September 2016 for a large European airline.

Using this foundation, we theoretically examine the suitability of Design theories for the early simulation of AI system's behaviour-use interdependence. We then introduce an initial version of a theoretical model, which proposes a set of relationships among the variables we identified.

purpose + data → frame[1]
frame + mode of action → prototype
prototype + actuation → outcomes

The model aids us to adapt the system's behaviour to trigger the desired use and outcomes. Consequently, it is best suited for the early stages of conceptual design when neither the behaviour nor the desired use (or outcomes) are clear. Hence, providing a preliminary answer to the main research question of the dissertation.

Part III: Extended theoretical model

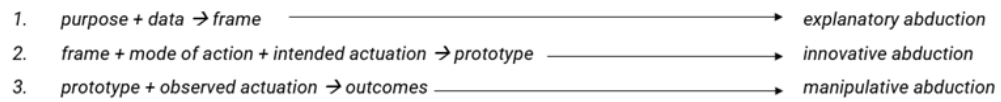
The Design theories we utilised, however, have been developed for the design of products and services. Unlike these, the behaviour of an AI system is continuously influenced by and learns from user-generated data. Therefore, in Part III, we explore further how to simulate the behaviour-use interdependence in the context of designing AI systems.

We begin our exploration by presenting an example of a three-person student team who successfully simulated the behaviour-use interdependence of an AI system four times by using simple prototypes. The team elicited multiple (intended and unintended) outcomes, which served as a robust feedback mechanism. Three things aided the team. Firstly, they explicitly identified intended actuations after deciding on their mode of action and before they built

1 A cognitive act of looking at a problem situation from a specific viewpoint that informs how the problem can be solved. According to Dorst (2011), the formulation of frames follows the format: "IF we look at the problem situation from this viewpoint, and adopt the working principle associated with that position, THEN we will create the value we are striving for." (Dorst, 2011, p. 525).

a prototype. The addition of this variable allows us to connect each row of the theoretical model with a different type of abductive reasoning: explanatory, innovative, and manipulative (Figure 0). Secondly, current Design theories suggest that to design a new solution, one needs to apply innovative (and explanatory) abduction. Yet, the team made use of all three abduction types to simulate the behaviour-use interdependence of their concept. Thirdly, they also used non-abductive activities such as explicitly defining requirements and values. These played a pivotal supportive role in the development of the concept for an AI system.

Figure 0 Formulaic representation of the theoretical model

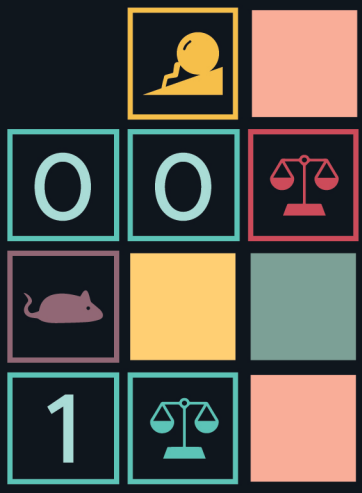


We address the uncovered insights with a five-month long case study. It discusses a design project for the devising of an in-car AI system conducted in collaboration with a large automaker. The case yields two main conclusions. Firstly, to support the early simulation of AI systems' behaviour-use interdependence, the three abduction types - explanatory, innovative, and manipulative - need to be applied. Secondly, existing Design theories need to be extended. Five insights can guide such extension: (1) explanatory abduction is usually followed by innovative abduction; (2) the inductive generation of new values and requirements informs the formulation of every variable of the model; (3) visuals generated as a result of inductive reasoning (e.g., data visualisations) facilitate explanatory abduction; (4) the deductive evaluation of each row's result against requirements and values supports the move from one abduction type to another; and (5) manipulative abduction plays a facilitative role while carrying out innovative abduction.

These insights form the basis of the final theoretical model which we name Theoretical Model for Prototyping AI or the **PAI model**. The PAI model is defined by relationships among abduction (explanatory, innovative and manipulative), induction and deduction. A model that provides us with a manner to support *the early simulation of AI systems' behaviour-use interdependence*. Furthermore,

it is our contention that the PAI model will be applied in the same manner by different AI developers regardless of their background or skill level. Finally, the model also provides an indication on how three different data types can be used to update AI system's behaviour during model development and deployment. These can serve as the starting point for the theory-testing part of the cycle Cash (2018) introduced.

Following the three theory building steps prescribed by Cash (2018), the devising of the PAI model allows us to shed light into how Design theories could contribute to the design of better AI systems. It also allows us to extend these theories and identify directions in which the field might (or should) go in a future defined by intelligent agents. Thereby, the model provides us with a way to approach the conceptual design of ever-evolving AI systems through the early simulation of their behaviour-use interdependence. Finally, these formalised relationships could also provide us with initial indications on how new AI models can be devised. AI models that do not rely solely on vast numbers of data points, but instead, allow for the creation of highly configurable world models AI agents can use.


0
0
1

PREFACE

EVIL MICE, SISYPHUS, AND THE TRUTH

It appears I am prone to existential crises. So much so that the dissertation you are about to read is the direct result of one. And doing what a researcher does, I have boiled down the triggers for my crises to three components: **evil mice**, **Sisyphus,** and **the truth**. If one is to understand this dissertation, they need to take stock of the three components first.

Evil mice

When I was around two years old, I hated going to the kindergarten. The reasons behind my vehement hatred are not remembered neither by my parents, nor by me. However, we can all recall how the hatred was substituted with cautious excitement. One day, I came home recounting tales of how I singlehandedly managed to defeat the evil mice and protected all the other children in my kindergarten. I am not sure what triggered this story. Yet, I remember what my mother did with it – she cleverly picked up on my overly active imagination and innate (and highly delusional) desire to see myself as the hero. So, every day she would tell me ever more fantastical tales about my adventures and great victories against the ever-elusive evil mice and their king.

From that day onward, my mother's task to get me to the kindergarten on time was made significantly easier.

Sisyphus

The second component contributing to my crisis-prone disposition is Sisyphus, and more specifically, the Sisyphus described by Albert Camus. Sisyphus was a mortal man who cheated Death. Thus, not unlike Prometheus, he was eternally punished for no one should dare cheat the gods. Sisyphus' punishment was more creative, however. While Prometheus played the rather passive role of having to wait for an eagle to eat his liver, Sisyphus had to push a boulder up a hill and just when it has reached the peak, the boulder would roll down and he had to start his journey anew. Every fibre of his body utilised and exerted for one purpose only – to accomplish nothing. For the gods decided, there is no worse punishment than meaningless labour.

The act of returning down the hill to get to the stone and begin his meaningless labour again, Camus sees as the "hour of consciousness". It is, Camus claims, this specific part of the punishment when Sisyphus becomes superior to his fate. For it is the struggle towards the heights that matter, not the number of times one must return back to their rock. One must imagine Sisyphus happy, Camus concluded. I also imagined him happy, although not for the same reasons. To me, there was something strangely liberating in knowing one can control her fate even in the presence of what seem like immutable constraints and predefined outcomes.

The truth

There is a Zen koan that goes something like that. A teacher pushed his student's head underwater and watched as the bubbles of oxygen dissipate. He then pulled the student's head out and told him that the moment he wishes to know the truth as much as he wished for oxygen, would be the moment in which he would find it. I, much like that student, have always been fascinated by the truth. In fact, throughout all my years in philosophy competitions, I wrote about truth almost exclusively. I wrote about it so much that my philosophy teacher would beg me to pick any other topic. Yet, to me the only worthy pursuit

has always been that of truth – the ability to get to the crux of a problem; to understand why things transpire; to know what is real and what is not. Of course, I soon realised the truth can be subjective and even the objective truth can be distorted. Yet, my fascination with the truth and what is real never really dissipated.

The existential crisis

Around the time I graduated from my master's degree, I was supposed to choose one of the many more than generous job offers I received. However, this was also the year when we, as a society, started getting glimpses on how AI systems (and in particular Facebook and Google's recommendation algorithms) could empower a slew of demagogues to "repackage" their propaganda as the truth, while squandering scientific evidence. We also saw how people were being discriminated against due to their skin colour, gender, or the neighbourhood they live in – systemic unfairness made worse due to models that are essentially black boxes. And all that was facilitated by technology – the very entity I had considered my entire life to be an objective source of progress and good. I, naturally, had the strong urge to do something about it. Yet, the problem seemed so insurmountable that I doubted anything can be done to ensure the negative unintended and unanticipated consequences of AI can be somehow mitigated. Certainly, not anything I was equipped to do. So, my strongest to date existential crisis settled in.

At the height of my crisis, I decided to climb the second highest peak in Bulgaria (with no previous experience) (Image 1). As one can expect from a very skill-less climber, while we were coming down the mountain, I made a mistake, and I thought I am about to die. When I did not, a very long few hours of Sisyphean consciousness and acute awareness of my mortality and insignificance followed. It was then when the three components came together and provided my existential crisis with a (delusional) purpose of sorts. I decided I want to be one of the people who are actively trying to design better AI. I did not want to be complacent with the development of systems that exponentially amplify pre-

existing societal biases and the voices of people who squander the truth. Yet, I had no idea how to do any of that.

Fittingly, I was coming down a mountain. I did not push a rock uphill, and I was not coming back to one either. Still, just like my two-year-old self, I was eager to confront an even more elusive nemesis than the evil mice. Soon after, I met Maaike Kleinsmann who employed me for two days a week and allowed me to research any topic within the area of AI – no constraints. This research work then slowly evolved to a full-fledged PhD research on the manners in which we can ensure AI systems will perform as intended once they get deployed into complex contexts. The work was supported by Maaike, Christine de Lille, Milene Gonçalves and Dirk Snelders. A work that often resembled the absurd Sisyphean labour – just when I thought I was about to push the boulder to the top, it rolled down.

Still, one must always imagine Sisyphus happy!



Image 1 Peak Vihren (2914 m) – the highest point of the Pirin Mountain in Bulgaria

INTRODUCTION

AI AND COMPLEX CONTEXTS

Self-learning Artificial Intelligence (AI)[1] systems are already deeply embedded into people's everyday lives and have a prominent role in their daily activities: from modes of manufacturing and transportation through the way people[2] receive information from news, select movies and songs, the way they date, trade at the stock exchange market, to the manner in which critical social institutions such as hospitals, banks, police departments and courtrooms make decisions (e.g., Crawford & Calo, 2016; Rahwan et al., 2019). The widespread adoption of AI systems is a direct result from the impressive strides the field has witnessed in the past decade in solving technical problems that

1 In this dissertation, we mainly refer to the type of AI that is currently making the biggest strides and affecting people's lives the most – Machine Learning (ML). AI and ML are two different things (i.e., ML is a branch of AI – see Chapter 2), yet we chose to retain both of them as the research presented here can be applied to different types of AI. Still, most of the examples we use are based on ML algorithms. Moreover, the aim of this dissertation is not to explain the technical differences of both approaches – many scholars have done that already (e.g., see Russel & Norvig, 2021). Rather, we focus on the behaviour an AI system exhibits once released into a broader context. Therefore, although ML and AI are not the same, we will use the terms interchangeably.

2 In this dissertation, unless explicitly stated, the pronoun "we" includes myself and my doctoral team. In the rare occasions in which the pronoun "I" is used, it refers to my personal opinions or activities I carried out as part of a case study (see Chapter 7).

have resisted the attempts of the AI community for decades (Ching et al., 2018; Rahwan et al., 2019). This statement rings true in the fields of speech and visual object recognition, object detection, drug discovery, physics, and genomics (Ching et al., 2018). For instance, a coordinated effort of multiple machine learning (ML) algorithms helped in converting telescopic data into the first-ever photo of a black hole's silhouette (BBC, 2019). Deep learning algorithms have been shown to predict lung cancer with a 94.4% accuracy[3] (Ardila et al., 2019), detect signs of autism in the human DNA (Zhou et al., 2019), accurately predict presidential elections (Kahn, 2020), and perform better than radiologists in detecting pneumonia from front-view chest X-ray images (Rajpurkar et al., 2017). Further, an ML classifier was used to non-invasively detect hypertrophic cardiomyopathy, even in asymptomatic patients (Green et al., 2019) and another classifier learnt to forecast the aftershock locations of earthquakes (DeVries et al., 2018). These and other similar achievements prompted the widely accepted notion that AI could provide a viable path to exponential societal betterment (e.g., Rahwan et al., 2019).

The wide-spread implementation of AI systems is a relatively new phenomenon. Since the inception of the field in the 1950s, AI systems have been designed, developed, and deployed in controlled environments (i.e., in labs or simulators). In a controlled environment, an application can be shut down or reprogrammed if it does not perform as intended without producing long-lasting effects on its environment and users (Russell & Norvig, 2021). However, as AI systems become an instrumental part of our everyday lives, they also start being embedded in complex contexts. In this dissertation, we use the term *complex contexts/systems* to denote a collection of interconnected and interdependent social, physical, and technical elements that exhibit emergent behaviours and properties not directly predictable from the behaviour of individual components. These systems often involve nonlinear interactions, feedback loops, and intricate relationships among their elements, resulting in behaviours that are difficult to fully understand and predict. An example of such comes from

3 The model used in this study performed on-par with radiologists when prior computer tomography imagining was available, and it was able to reduce the rate of false positives with 11% when no computer tomography imaging was available.

Microsoft. In 2016, the company introduced an AI chatbot that was designed to "emulate the casual, jokey speech patterns of a stereotypical millennial" (Price, 2016). They called it Tay and stated that its aim was to engage millennials by learning from its conversations with people on Twitter and get smarter over time (Hunt, 2016). When Tay was deployed, however, people started tweeting racist statements using the designated hashtags (i.e., #taytweets). Doing what it was designed to do (i.e., learn from conversations with people), the bot started generating racist tweets defending white supremacy, denying the Holocaust, and praising Nazis (ibid). Microsoft promptly took down Tay and apologised for the damage it created (ibid). It is of utmost importance to ensure AI systems like Tay behave as intended because they can scale quickly over short periods of time (i.e., many people can be affected by them) and transcend the platform for which they were designed (Russell & Norvig, 2021). Hence, their unintended consequences may cause far-reaching irreversible damage both to their direct users and to the companies that created them.

There is a growing awareness of the need to address the plethora of negative unintended consequences that arise once an AI system becomes an integral part of complex contexts. The plea for additional research on the subject has permeated across academic fields including, but not limited to, cognitive systems engineering; human factors; science, technology and society; safety engineering; ethics; legislation (i.e., for a comprehensive overview see Johnson et al., 2013; Crawford & Calo, 2016; Amodei et al., 2016; Rahwan et al., 2019), and design (e.g., Human-centered ML and approaches based on the work of Friedman's Value Sensitive Design (e.g., Friedman et al., 2002)). In addition, different civil rights groups (e.g., Chee, 2021) and governmental bodies (e.g., in the United States[4], the European Union[5], and China[6]) have proposed regulations to mitigate the risks an AI system might produce. As such, AI systems and their potential influence over existing systems have become one of the largest

4 Algorithmic Accountability Act of 2019

5 It is worth noting that this proposed regulation has been criticized by civil rights groups for its multiple loopholes which could lead to the abuse of AI systems by repressive governments (Chee, 2021).

6 In August 2021, the Cyberspace Administration of China (CAC) released draft regulation guidelines for algorithmic recommender systems (i.e., AI systems) (Singh, 2021).

technological and societal challenges humans face today (e.g., Rahwan et al., 2019; Russell & Norvig, 2021). The next section of the Introduction elaborates further on this challenge.

AI and complex contexts

AI systems are software programmes defined by their continuously self-learning nature that "can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with" (Artificial Intelligence Act, 2021)[7]. These systems are devised to answer questions that are clearly defined, involve an enumerable set of solutions, clear rules, and inherently binary decision mechanisms (their output is either true or false) (Russell & Norvig, 2021). As such, they inherently solve what Rittel and Webber (1973) call "tame problems". However, once these systems that are devised to solve tame problems transcend the boundaries of the platform for which they were designed, they start facing problems which do not have binary answers, clear rules, or enumerable set of solutions (Stoimenova & Price, 2020). Rittel and Webber (1973) call these types of problems "wicked". Wicked problems are ill-formulated unique problems that can never be fully defined[8], have no stopping rule or permissible set of operations and their solutions can only be classified as good or bad (instead of true or false) (ibid). To exemplify this tension between an AI system designed to provide binary answers (i.e., "yes" and "no") and a complex context where there are no clear-cut optimal answers, we use the work of White and colleagues (2018).

In April 2018, White and co-authors from Microsoft Research and Duke University published a paper reporting their initial attempts to create a "simple scalable test that can be used for screening of Parkinson's disease in the community or at home." (ibid). The researchers used longitudinal log data from Microsoft's search engine, Bing, to look into the presence and frequency of

7 The definition was put forward by the European Union in their proposal regulation titled "Laying down harmonised rules on artificial intelligence", published on April 21, 2021 (Title 1, Article 3).
8 According to Rittel and Webber (1973), "the information needed to understand the problem depends on one's idea of solving it" (p. 161).

symptom-related query terms; motor symptoms, such as the speed, direction, and tremors of cursor movements; and presence of risk factors. Despite still being in a testing phase, their AI system showed promise in detecting a disease (with an accuracy of 94.2 %) that has a current clinical early diagnosis accuracy of approximately 80 percent (ibid). The focus of the underlying AI algorithm was to accurately predict the presence of factors that potentially signal Parkinson's disease. In effect, it answered a binary question – there could either be early signs of Parkinson's in each individual's data or not. However, the team's ambition to create a test for community and home use (i.e., implement the AI into a complex context) inevitably posited a wicked problem (i.e., the information delivery of this type of solution requires careful consideration of perspectives in allied health systems). How does the home test deliver a diagnosis? What role do doctors play when diagnosis is outsourced? How does the test connect the diagnosed within the allied health service? How might the family require support after a diagnosis? In such situations, there are innumerable set of potential solutions and no well-described set of permissible operations. Moreover, there are many stakeholders that are interested to judge the solution, but none has the power to set formal decision rules to determine correctness. Finally, there is no immediate and ultimate test of a solution to the problem that ensures positive impact. Yet, delivering a false diagnosis or delivering a diagnosis insensitively could cause significant distress to the community.

What the examples we discussed so far show is that there is a technology that has the potential to exponentially better human life and science. Yet, once it starts to affect the context it is situated in, this technology designed to answer relatively well-defined (i.e., tame) questions also raises many ill-defined or wicked ones due to its self-learning nature and the complexity of the context in which it is implemented (Rudin, 2019; Rahwan et al., 2019). Therefore, if we are to design AI systems that do not produce far-reaching irreversible negative consequences, we *cannot decouple/detach* them from the complex contexts in

which we are to embed them. This statement brings us to the initial research question of this dissertation.

The initial research question

Historically, the field of Design developed as a response to large changes in society facilitated by rapid technological advancements (Calabretta & Kleinsmann, 2017). In the span of a century, Design gradually evolved from designing products for the highly industrialised production lines to the creation of optimal human-computer interactions and product-service-systems (ibid). Then, with the wide-spread adoption of human-centred design methods and design thinking, also came the increasing involvement of designers in complex sociotechnical arenas such as healthcare, sustainability, and transportation (Norman & Stappers, 2015; Stoimenova & Price, 2020). This expansion of the Design field sparked an on-going debate whether existing Design theories and methods need to be further developed to address the increasingly complex contexts in which designers operate (e.g., Voûte et al., 2020).

It is our contention that the same type of methodological exploration is needed for the design of ***AI systems that are to be implemented*** into larger complex contexts. This contention brings forward the initial research question of this dissertation:

"How can Design theories support the design and implementation of AI systems into complex contexts?"

An exploration prompted by this research question would shed light into the manners in which Design theories could contribute to our understanding of how AI systems can be designed. Further, it could identify possible directions in which the field might (or should) go in a future defined by intelligent agents.

Research method

In order to answer the aforementioned research question, we adopt theory-driven research and, in particular, the theory building/testing cycle proposed by Cash (2018). This type of research is at the core of robust scientific knowledge

that is “valuable internally and accessible externally” (ibid, p. 87). It also allows for the creation of a foundation that can be clearly communicated across different fields (ibid).

The cycle Cash (2018) introduces is comprised of five distinct stages that guide the designer in gradually building up a robust scientific theory (Figure 1). This dissertation only addresses the theory-building part of the cycle (highlighted in the figure). Namely, the stages of ***Discovery and description***, ***Definitions of variables and limitation of domain*** and ***Relationship building***. It is our contention that by focusing on the theory-building part of the cycle, we will be able to devise a theoretical model that can serve as the foundation on which predictions can be generated. Thus, allowing for the rigorous testing and falsification of the theory by different scholars, in different contexts.

During the first stage, one details the important issues around which the theory will be built. This is done by establishing the general characteristics of the issues and the potential importance of research in this area (Cash, 2018). During the second stage, one identifies the variables that will underpin the theory and carefully crafts their definitions (ibid). According to Wacker (1998, 2008), for a theory to be good, its definitions need to fulfil three criteria. First, they need to be *conservative* – use new terms only if they can be clearly distinguished from existing ones. Second, they should be *unique* – if new terms are introduced, they should not borrow from existing conceptual definitions. Third, they should be *parsimonious* (i.e., only short definitions should be used). During the second stage one also elaborates upon the limitations of the domain (i.e., where and when a theory is to be applied (Wacker, 2008)). As such, the domain should be *generalisable* (i.e., “the degree to which a theory can be applied to existing populations” (ibid, p. 10)) and *abstract* (i.e., the theory can be applied across different times and places).

During the third stage, one describes the conceptual relationships among the variables identified in the previous stage, as well as the rationale behind these relationships in the light of the identified domain (Cash, 2018). In order to ensure that the relationship stage is executed well, one needs to observe the following three criteria, according to Wacker (1998, 2008). First, the

relationships should be *fecund* – i.e., they should build on existing theories and integrate them, so that new, wider research areas can be explored.

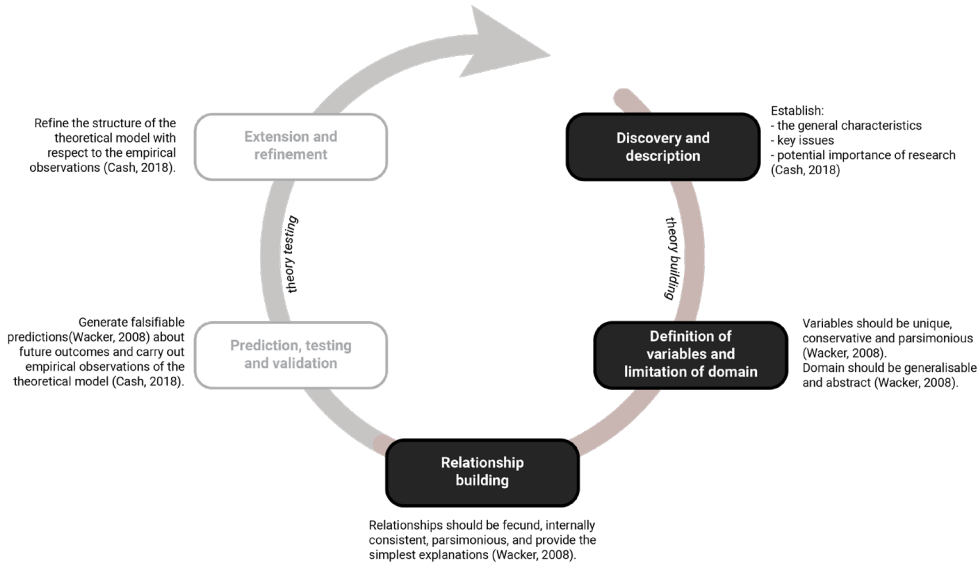


Figure 1 The theory building/testing cycle (Cash, 2018), paired with the requirements of good theory (Wacker, 2008)[9].

Second, the proposed relationships should be internally consistent: they should clearly explain how each variable is connected to the other variables. This can be done either through mathematics or symbolic logic (i.e., they should be logically consistent). Lastly, they should be parsimonious so that the resulting model is as simple as possible. Still, it should be able to accurately explain all the results.

Finally, as it can be seen from the requirements for good theory Wacker (1998, 2008) put forward, both the variables we use and the relationships among them should consider existing theory (i.e., the definitions should be conservative and the relationships – fecund). Therefore, the theory building approach we

9 All visualisations presented in this dissertation, unless explicitly referenced, are made by me.

adopt in this dissertation can be classified as *analytical conceptual research* (Wacker 1998,) which adds "new insights into traditional problems through logical relationship-building" (p. 373). However, since the design of AI systems is a new domain for the field of Design, we complement the analytical approach with empirical investigations. Doing so allows us to "raise the abstraction level" (p. 379) of the developed theory. Therefore, our theory-building approach combines both theoretical (e.g., Chapters 3, 4 and 5) and empirical (e.g., examples presented in Chapters 4 and 6, and a case study discussed in Chapter 7) explorations.

Dissertation outline

The dissertation consists of three parts that follow the stages of the theory-building method described previously. In Part I, **Set the scene**, we describe the context of this doctoral research and identify a gap in the literature. These then we use to formalise the dissertation's main research question. In Part II, **Initial theoretical model,** we define the variables, which could support us in designing a theoretical model that serves as a potential answer to the main research question of the dissertation. We then propose an initial set of relationships among these. Finally, in Part III, **Extended theoretical model,** we further detail these relationships to reflect the context we identified in Part I. As Cash's theory building cycle prescribes, these parts are connected and build on each other (see Figure 2). In order to reflect this theory-building choice, we visualised the chapters as interlocking building blocks. Below, a more detailed outline on each chapter can be found. In it, we elaborate upon the carried-out research and the elicited insights that allowed us to gradually build and define the final version of the theoretical model. Hence, it is positioned on the top of our tower of stackable building blocks.

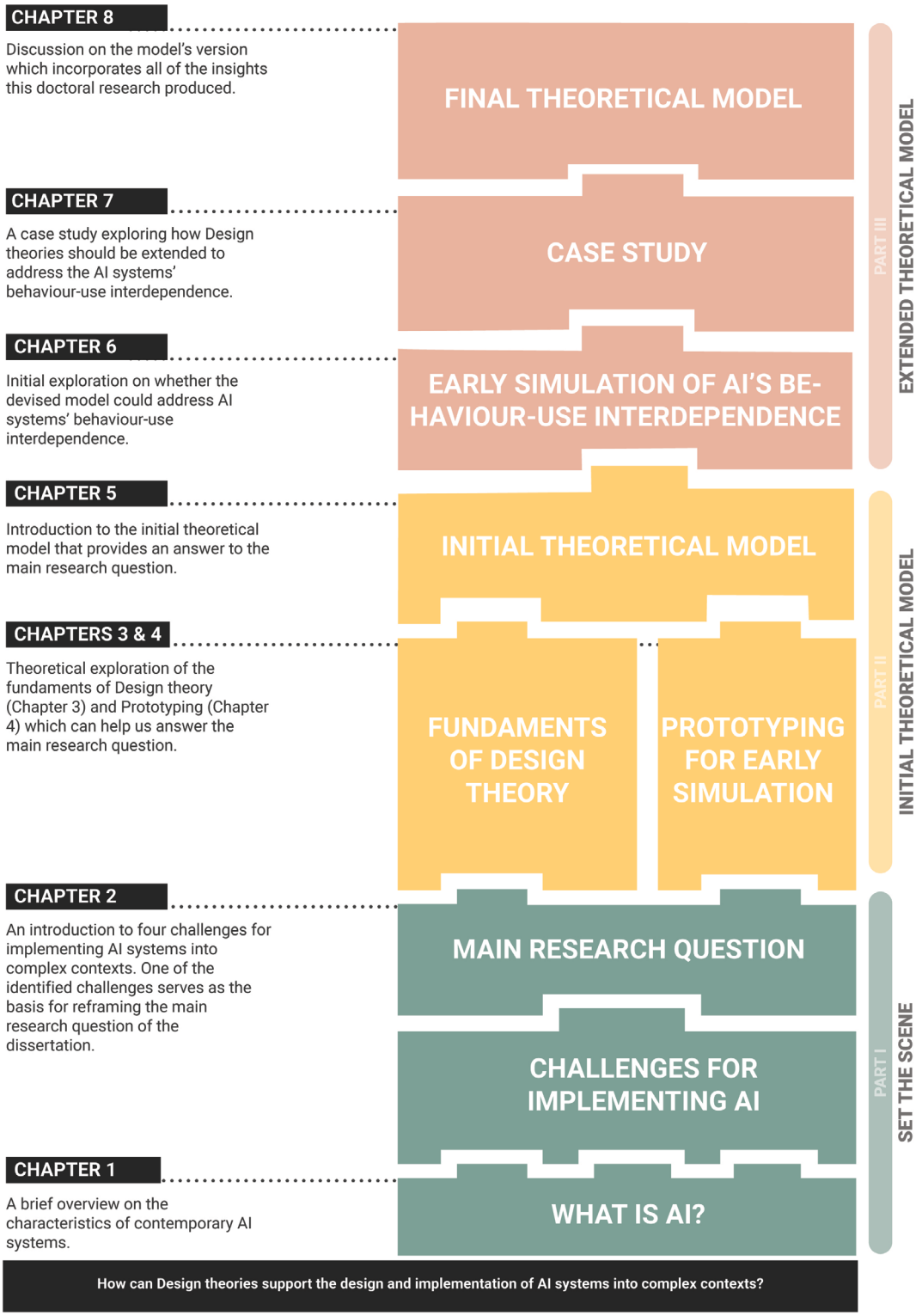


Figure 2 A visual overview on the dissertation's outline (to be read from the bottom).

Part I: Set the scene

Chapter 1: Characteristics of contemporary AI systems

AI systems are self-learning and come with their own unique set of characteristics. Thus, if we are to understand how to support the design and implementation of AI systems into complex contexts, we first need to gain a deeper level of understanding on these characteristics. To do so, in this chapter, we first present a historical overview on the developments in the AI field and its most widely utilised approach to designing AI in both industry and academia – the rational agent approach. We then pair this with an overview of AI's sub field – Machine Learning – which is, at the time of writing, the most widely applied branch of the field. As a result, we identify three characteristics of contemporary AI systems: (1) AI systems' complex functional processes, (2) the high dependency on human-generated data and (3) the continuous process of learning and adapting the system's behaviour to fit to the manner in which humans use it.

Chapter 2: Challenges for implementing AI systems into complex contexts

Chapter 2 focuses on the discussion of the main challenges that emerge when embedding AI systems into complex contexts. We start by outlining four principles an AI system must abide by to warrant ethical implementation: (1) transparency, (2) fairness, (3) responsibility, and (4) non-maleficence (i.e., causing no harm). We then pair these with the three characteristics expounded upon in Chapter 1. The combination results in four challenges: (1) **ensuring transparency** of the innerworkings of AI systems, (2) **mitigating implicit biases** in the data used to train and retrain these systems, (3) the difficulty to **align** the behaviour of AI systems **with human values**, and (4) **addressing the behaviour-use interdependence** of these systems. There are research approaches for all challenges, except for the fourth one. We contend that Design theories are particularly well-positioned to tackle it. This gap in the literature presents us with a clear opportunity to contribute to the efforts of designing ethical AI systems. It also results in the reframing of the initial research question we

presented in this Introduction. As such, the main research question of the dissertation becomes:

"How can a theoretical model be designed that supports the early simulation of AI systems' behaviour-use interdependence by utilising Design theories?"

Part II: Initial theoretical model

Chapter 3: The fundaments of Design Theory

This chapter provides an overview of the ways Design theories allow us to address the behaviour-use interdependence challenge identified in Chapter 2. As such, it consists of two sections. In Section I, we present and compare the two main paradigms of Design Theory – Rational Problem Solving and Reflective Practice. Three important insights emerge from this: (1) the need to establish a feedback mechanism to better understand a problem; (2) the influence of a designer over the design process; and (3) unintended consequences propel the design process forward. In Section II, we introduce abduction[10] as the key reasoning mode of Design, defined by two models. These models provide insight into the manner in which the behaviour-use interdependence can be formalised when designing new solutions. As such, the chapter provides us with a much-needed foundation for further exploring the claim we made in Chapter 2 that the field of Design is well positioned to tackle the behaviour-use interdependence challenge.

Chapter 4: Prototyping for early simulation of behaviour and use

Building upon the previously presented overview on Design Theory, in Chapter 4, we discuss the manners in which the design practice of prototyping can support the early simulation of the behaviour and use of new solutions. As such, we first present an overview on existing literature. We then complement it with two examples from my design practice. Three insights emerge from this overview. Firstly, prototypes support us in observing the different types of outcomes and uses a solution can uncover. Secondly, they serve as a bridge

10 Abductive reasoning is seen as the only logical inference that allows us to generate new hypotheses (see Chapter 3).

between behaviour and use. Thirdly, they enable a different type of abduction which aids us to formulaically build the relationship between behaviour and use in complex contexts. Therefore, prototyping marks another puzzle piece of Design theories we can use to answer the main research question of this dissertation.

Chapter 5: Initial theoretical model

In this chapter, we theoretically examine the suitability of existing theories for the design and implementation of AI systems into complex contexts. As a result, we introduce an initial version of a theoretical model, which proposes a set of relationships among variables found in Design theories. As such, the model could aid us in simulating the behaviour-use interdependence of AI systems so that we can adapt the system's behaviour to trigger the desired use and outcomes. Consequently, we contend that the model is best suited for the early stages of conceptual design when neither the behaviour nor the desired use (or outcomes) are clear. Hence, it provides a preliminary answer to the main research question of the dissertation.

Part III: Extended theoretical model

Chapter 6: Early simulation of AI system's behaviour-use interdependence

In this chapter we present an example of a student team that successfully simulated the behaviour-use interdependence of an AI-powered fitness app by applying the initial theoretical model. We use the example as a means to explore the manners, in which the model can inform and shape real-world decision-making and design processes. Consequently, the example allows us to suggest an extended formulation of the initial theoretical model. It also highlights the need for further research around the manner in which: (1) designers apply the three different types of abductive reasoning; (2) the three abductive types influence each other; and (3) the role non-abductive activities play. These three insights also highlight the need to extend the existing Design theories to address the new context of designing AI systems.

Chapter 7: Case study – Designing a smart in-car system concept for a large automaker

Chapter 7 builds upon the insights and research directions presented in Chapter 6. As such, we report on a five-month long design project conducted in collaboration with a large automaker. During the project, we apply the extended theoretical model to the conceptual design of an in-car AI system. The study yields two main conclusions. Firstly, to support the early simulation of AI systems' behaviour-use interdependence, all three types of abduction discussed in the dissertation - explanatory, innovative, and manipulative - need to be applied. Secondly, five insights can guide such extension: (1) explanatory abduction is usually followed by innovative abduction; (2) the inductive generation of new values and requirements informs the formulation of every variable of the model; (3) visuals generated as a result of inductive reasoning (e.g., data visualisations) facilitate explanatory abduction; (4) the deductive evaluation of each row's result against requirements and values supports the move from one abduction type to another; and (5) manipulative abduction plays a facilitative role while carrying out innovative abduction. These insights form the basis of the final theoretical model and provide us with a manner to continuously support ***the early simulation of AI systems' behaviour-use interdependence***.

Chapter 8: Final theoretical model

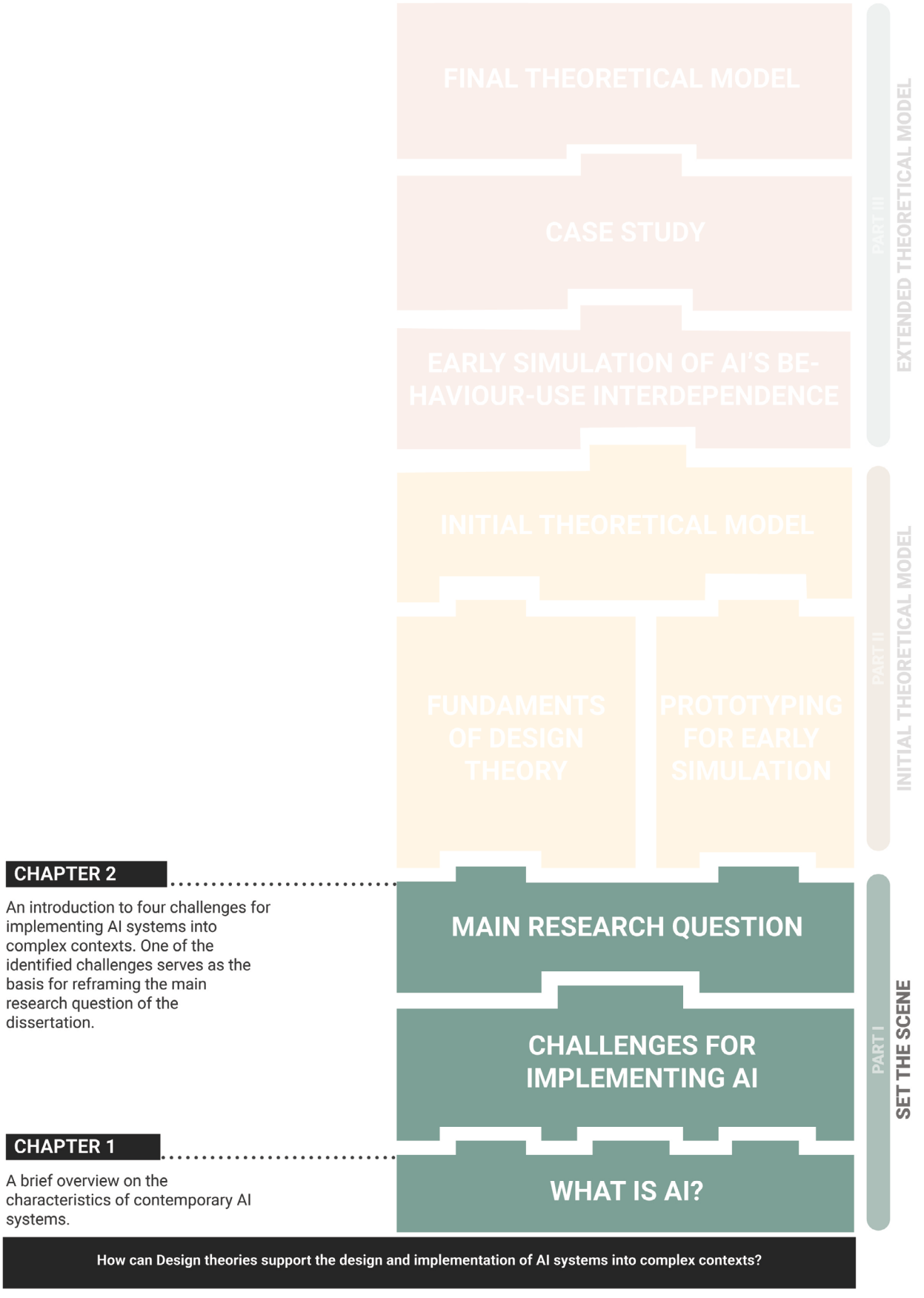
In Chapter 8, we reflect the findings generated thus far into the final version of the theoretical model we term **Theoretical model for Prototyping AI** or the **PAI model**. The model is presented by detailing its domain (i.e., *early simulation of AI systems' behaviour-use interdependence during the conceptual design stage of their development*), variables (and their definitions), and the relationships among these variables presented both formulaically and schematically. Finally, we introduce three predictions about the PAI model. These are presented around the topics of (1) early simulation of behaviour-use interdependence of AI systems; (2) the influence of different data types on the AI system's behaviour; and (3) the manners in which different developers will apply the model. These can serve as the starting point for the theory-testing part of

the cycle Cash (2018) introduced. The predictions are then followed by a discussion on the manners in which the proposed theoretical model fulfils the requirements of good theory Wacker (1998, 2008) put forward. Finally, we conclude the chapter with a brief discussion on the implications of the introduced theoretical model for Design theory, education, and practice. These also suggest possible avenues for further development of the Design field so that it can be prepared for the new domain of designing AI systems. Namely, the need for designers to (1) have a basic understanding of the new material with which they will be designing (i.e., AI systems), (2) utilise methods and tools designed for the transient nature of AI systems, and (3) learn to collaborate with data scientists, engineers, ethicists, and individuals who understand the regulations on AI systems.

PART I

SET THE SCENE




FINAL THEORETICAL MODEL
CASE STUDY
EARLY SIMULATION OF AI'S BE-
HAVIOUR-USE INTERDEPENDENCE
PART III
EXTENDED THEORETICAL MODEL
INITIAL THEORETICAL MODEL
FUNDAMENTS
OF DESIGN
THEORY
PROTOTYPING
FOR EARLY
SIMULATION
PART II
INITIAL THEORETICAL MODEL
CHAPTER 2
An introduction to four challenges for implementing AI systems into complex contexts. One of the identified challenges serves as the basis for reframing the main research question of the dissertation.
MAIN RESEARCH QUESTION
CHALLENGES FOR
IMPLEMENTING AI
CHAPTER 1
A brief overview on the characteristics of contemporary AI systems.
WHAT IS AI?
PART I
SET THE SCENE
How can Design theories support the design and implementation of AI systems into complex contexts?

CHAPTER 1

CHARACTERISTICS OF CONTEMPORARY AI SYSTEMS

AI systems are defined by their continuously self-learning nature that presents its own unique set of characteristics. Therefore, establishing a basic level of understanding on the fundaments of this technology is imperative if we are to successfully design and implement AI systems into complex contexts. To do so, this chapter provides a brief overview on AI's history and the current developments in the field, guided by the following research question: *"What are the characteristics of contemporary AI systems?"*.

The chapter is structured around two topics. First, we present the developments in the field of Artificial Intelligence by sketching out the four dominant views in the field on what constitutes intelligence. This overview then serves as the foundation to the introduction of different types of Machine Learning (ML) approaches. We specifically focus on this sub field of AI since ML has become the most widely applied branch of the field in the 2010s and 2020s. In fact, almost all of the examples used in this dissertation are of ML systems. The overview provided in this chapter is by no means exhaustive. For a more comprehensive one about AI/ML in general, please refer to the textbook on AI

written by Russell and Norvig (2021). Finally, the chapter is concluded with an overview on the high-level characteristics of contemporary AI systems.

Four research directions to intelligence

The field of Artificial Intelligence has its roots in a plethora of disciplines such as Philosophy, Mathematics, Economics, Neuroscience, Psychology, Computer Engineering, C`vvvontrol theory and Cybernetics, and Linguistics (Russell & Norvig, 2021). Consequently, over the years, the views on what constitutes "intelligence" came to be different and sometimes even contradictory. Russell and Norvig (ibid) categorise these different interpretations of "intelligence" along two dichotomies: (1) *human* versus *rational* and (2) *thought* versus *behaviour*. The first dichotomy stems from the different definitions of intelligence – either as the degree to which an agent's actions resemble that of a *human*, or the degree to which these actions can be defined as *rational* (ibid, p. 1). The second dichotomy juxtaposes the views that intelligence can ither be seen as an internal *(*i.e., *thought*) or as an external (i.e., *behaviour*) phenomenon (ibid).

Over time, these dichotomies, and the underlying premises they entail resulted in four main research approaches towards developing artificial intelligence (see Figure 3). Firstly, we have the research approach based on the premise that AI should be seen as something that ***acts humanly***. According to this view, intelligence will be achieved when a machine's actions cannot be distinguished from the actions of a human being. It is widely considered that a machine will be accepted to do so when it manages to pass the so-called Turing test. The test involves a human interrogator posing a few written questions to a computer. To pass the test, the computer should return answers that cannot be distinguished from answers a human would give. Hence, a model stemming from this approach should make use of six main capabilities: (1) *natural language processing* (so that it can communicate with the interrogator), (2) *knowledge representation* (to be able to store and access what it learns), (3) *automated reasoning* (to be able to answer questions and draw conclusions), and (4) *machine learning* (to adapt to new situations and be able to recognise patterns). Other researchers have proposed a total Turing test, which involves

real-world interactions with people and objects. To be able to pass such a test, the computer should also be able to perform (5) *computer vision and speech recognition* (to perceive the world) and (6) *robotics* (to move around and manipulate objects). These six capabilities represent the main branches of the AI field (Russell & Norvig, 2021).

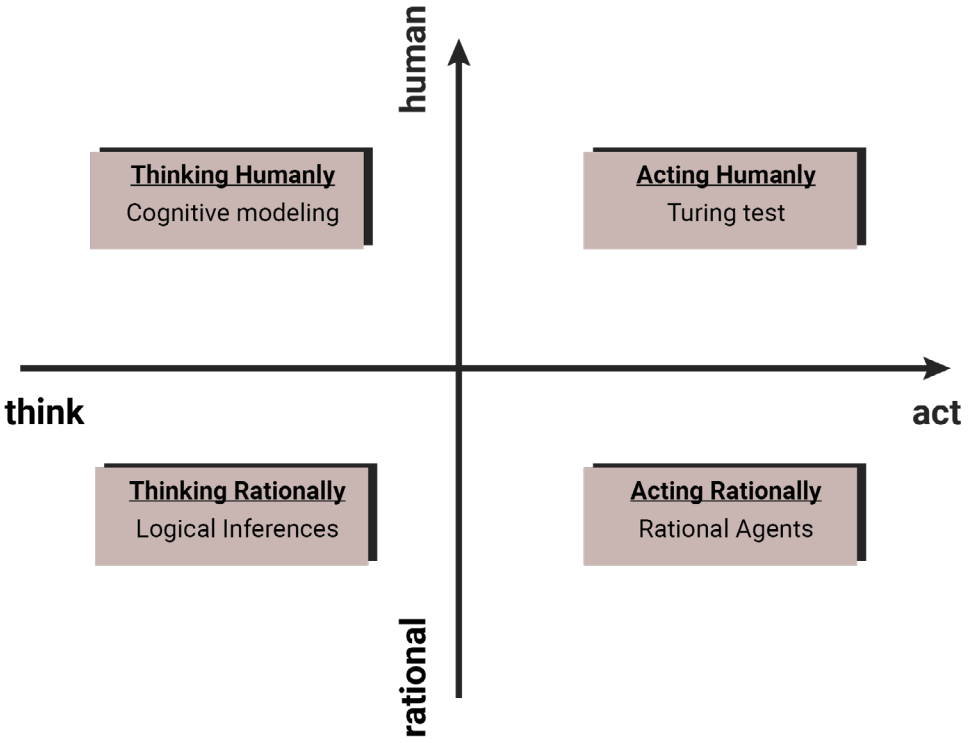


Figure 3 An overview of the four research directions in intelligence research derived from the work of Russell & Norvig (2021)

The second research approach to AI is grounded in the assertion that to achieve intelligence, a machine should be able to ***think like a human***. This view posits that to be able to devise such machines, one needs to develop accurate cognitive models akin to the ones humans use (Russel & Norvig, 2021). However, researchers must first gain a deeper understanding of the processes that drive human cognition. It is considered that once there is a sufficiently precise theory of the mind, this can be translated to machine theory. Researchers usually obtain the needed knowledge by employing the

methods of introspection, psychological experiments, and brain imaging. A field that addresses AI from this perspective is Cognitive Science. It combines experimental methods from Psychology and computer models from AI so that precise and testable theories of the human mind can be devised.

The third research approach to AI is founded within the notion that a machine should be able to ***think rationally.*** This approach has largely been defined by the field of Logic and uses logical principles that provide precise notation for statements and objects in the real world (Russell & Norvig, 2021, p. 5). As such, by 1965 programmes were able to solve any "solvable problem defined in logical notation" (ibid, p. 5). This launched the so-called logicist tradition in AI that hoped to build programs defined by logical principles that can create intelligent systems. A few notable examples of such rule-based systems are Terry Winograd's natural language understanding systems called SHRDLU (1968), Stanford's infectious disease diagnosis system MYCIN (1976), and XCON (also known as R1), which was used to configure computer systems (1978). However, to be able to apply these logical principles, one requires knowledge about the world that is certain – a requirement that is difficult to achieve in practice. There are no clear-cut rules of how society works. The theory of probability aims to address this gap by providing tools and methods that support one in reasoning about uncertain information. However, while the theory allows for the construction of comprehensive model of rational thought, it does not generate intelligent behaviour (ibid).

Finally, the fourth research approach to AI is based on the premise that a machine can be considered intelligent only when it ***acts rationally*** (i.e., acting "so as to achieve the best outcome or, when there is uncertainty, the best expected outcome" (ibid, p. 6)). Although it is a distinct manner of developing AI, the acting rationally approach builds upon principles and capabilities developed under the other three approaches discussed thus far. For instance, in order to ensure that the agent is rational, in some cases it needs to be able to apply logical inferences (i.e., from the thinking rationally quadrant) so that it can deduce the best (expected) outcome. However, there are situations in which rationality cannot be expressed in inferences defined by formal logic

(i.e., the example Russell and Norvig give is of recoiling from a hot stove – a reflex that requires immediate action rather than careful deliberation). In such cases, to build a rational agent, the six capabilities that are necessary to pass the Turing test are needed, too (i.e., natural language processing, knowledge representation, automated reasoning, machine learning, computer vision, speech recognition, and robotics).

The four research approaches have garnered their own supporters and research streams and continue to do so. However, in recent years, the rational agent approach has become the most widely utilised one in both industry and academia since it provides a clear and concise way to model intelligent systems (Russell & Norvig, 2021). Furthermore, it has been shown to be scalable and versatile, especially when dealing with complex systems. Hence, making it easy to apply to a variety of real-world applications like decision-making and planning tasks (ibid). Therefore, in the past decade increasingly more AI systems employing this approach have been embedded in complex contexts. This shift was further supported by the increasing computational power and the creation of the World Wide Web. The combination of both provided the means and the infrastructure for the creation of massive amounts of information generated every day
. This newfound resource led to the creation of new learning algorithms that were designed to work with the large amounts of data and benefit from them. An AI branch that emerged as uniquely positioned to take advantage of the large amounts of both labelled and unlabelled data is Machine Learning (ML).

Machine Learning

The premise behind ML is straightforward: rather than manually programming a system to anticipate the desired results for every conceivable input, an ML model is trained by being shown examples of desired input-output behaviour (Jordan & Mitchell, 2015). This continuous automated improvement leads to better decision-making mechanisms based on vast amounts of collected data. For instance, ML applications have achieved highly accurate results in the fields of computer vision, speech recognition, natural language processing and robot control (e.g., Rahwan et al., 2019). A large array of ML algorithms has been

employed to cover the wide variety of data and problem types exhibited across different fields (De Choudhury et al., 2014), such as linear regression, logistic regression, decision trees, and deep neural networks. Despite their differences, however, they can be grouped in four general types: supervised, unsupervised, reinforcement and deep learning.

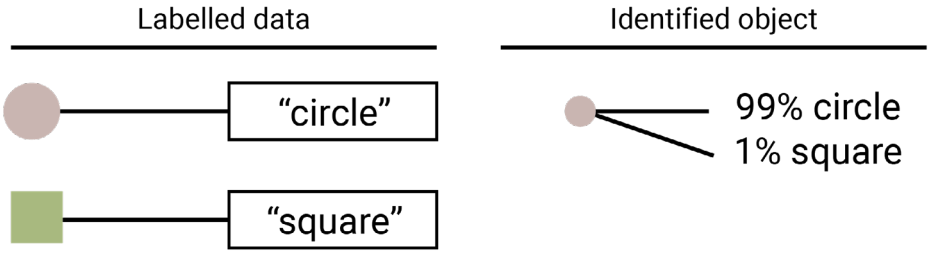


Figure 4: Supervised learning

The most widely applied type of ML is that of **supervised learning** (Figure 4). Such requires a labelled set of training data in which both input and output variables are given. With the correct answer known (the output) for each situation (input), the algorithm is trained to iteratively make predictions on the training data until an acceptable level of performance is achieved. For instance, supervised learning is widely used in email clients in order to identify spam email. To do so, the agent is trained by being shown multiple examples of spam emails (labelled data). During its training, it is shown examples of emails (input) that are then connected to a clearly defined output (either *spam* or *not spam*). As such, once the agent is deployed in real-life settings, it can confidently assign the label of either spam or not spam to emails it has never "seen" before.

The second type is that of **unsupervised learning** (Figure 5). Unsupervised learning analyses unlabelled data in order to extract information from them. As such, the system develops and organises the data, and searches common characteristics among them. Therefore, it often uncovers previously unknown patterns (Russell & Norvig, 2021). The most common unsupervised learning task is to detect common characteristics among the input data and create clusters around them (i.e., clustering). Clustering is used when we do not know

enough information (or any information) about the data we have collected, and we are looking to create groups from them. For instance, when fed with a large number of unlabelled photos from the internet, an AI model can start creating clusters such as “dogs” or “babies”. This type of learning is widely used in recommendations systems that group together users with similar viewing patterns (e.g., the claim one might find on different sites: “people who liked this, also liked these”).

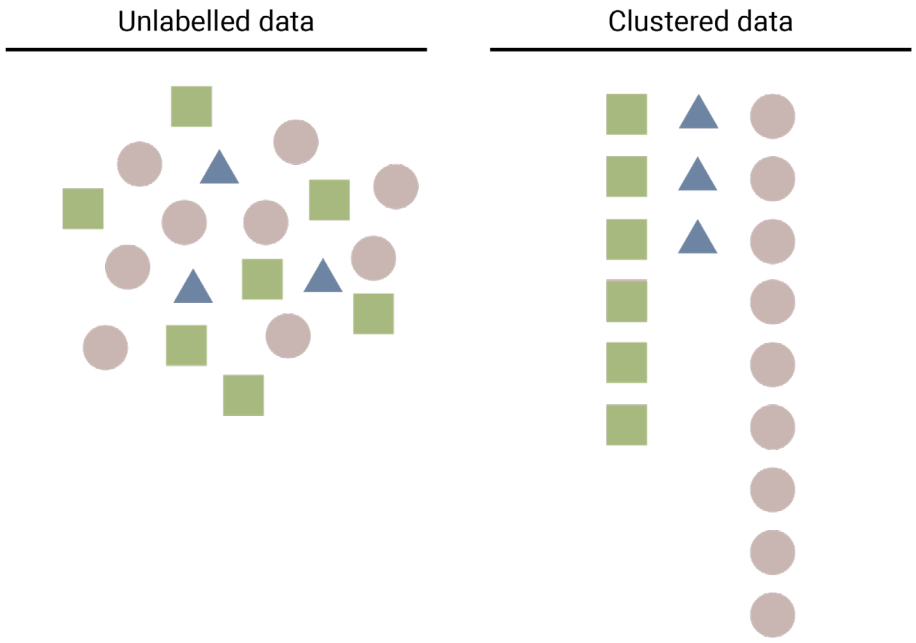


Figure 5: Unsupervised learning

The third main type of ML is **reinforcement learning** (Figure 6). Utilising this type of learning allows the agent to learn from a series of rewards and punishments (i.e., reinforcements) so that it can optimise the sum of future rewards. For instance, if the agent is to play a game of chess (i.e., act), at the end of the game the agent is told the outcome (i.e., the agent’s current state). If the agent has won, it gets rewarded by receiving more points. If, however, it has lost, a punishment is administered (i.e., by deducting points). Then,

given this output, the agent needs to decide which of its actions (prior to the reinforcement) were responsible for the outcome (i.e., win or lose). Once it has done so, it can alter its course of action so that it can maximise its reward and minimise punishment (Russell & Norvig, 2021). An example application for this type of learning is recommendation systems where the agent is rewarded based on whether the user likes its suggestion and punished if they do not. Services that use reinforcement learning to form their recommendations are TikTok, Facebook, Netflix, and Spotify. Therefore, if the user liked the suggestion, the agents gets "rewarded" to show more similar types of recommendations.



Figure 6: Reinforcement learning

Finally, we have **deep learning** (Figure 7). Deep learning is widely considered as the catalyst for the wide-spread renewed interest in AI we are seeing today due to its impressive results in variety of domains (LeCun et al., 2015; Ching et al., 2018). Further, it is making considerable advances in solving problems that have resisted the best attempts of the AI community for many years (Ching et al., 2018), especially in the fields of machine translation, visual object and speech recognition, and speech and image synthesis (Russell & Norvig, 2021). Deep learning models are used across different disciplines and applications (e.g., image recognition and autonomous driving). Partially mimicking the way

biological neurons exchange information, deep learning models have three main elements: an *input layer* which receives the data, an *output layer*, returning the analysed data and multiple *hidden layers*[1] that perform mathematical computations on the input data. Each of the layers contains multiple nodes (i.e., neurons) through which data and computation flow. As such, each node of the network performs some form of computation which is then passed onto another node in a different layer. This transfer of information from one layer to the other is made possible through the so-called synapses that have different weights. This weight determines how impactful a node is in the entire neural network. Most deep neural nets have multiple hidden layers with a large number of nodes in each of the layers. Therefore, the weight assigned to each synapse is done by the neural net itself during its training. Despite having a simple structure, however, the exact functional processes that generate these outputs are complex and hard to interpret even by the very scientists who designed them (Rahwan et al., 2019).

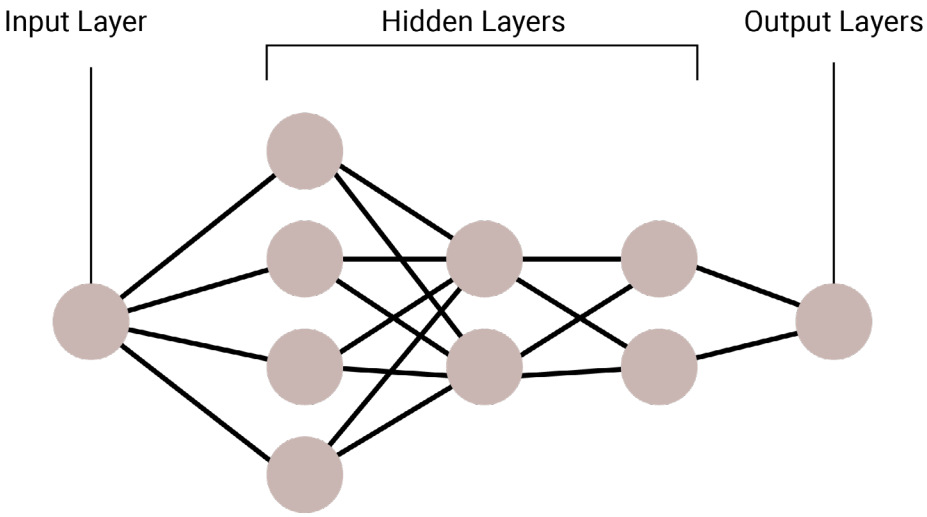


Figure 7: Deep learning

1 These layers are sometimes called "hidden" because they are not directly observable by the end user.

The sub field of Machine Learning is still relatively young and continues to evolve quickly with new approaches and methods emerging on monthly basis. Therefore, the overview presented in this section is by no means exhaustive. Still, the four types of ML presented here can provide us with a foundational understanding on the continuous stream of new approaches (Jordan & Mitchell, 2015).

Characteristics of contemporary AI systems

From the overviews presented in this chapter, three general characteristics of contemporary AI systems emerge. First, although one can clearly distinguish between the four main types of ML, in practice few AI systems use only one ML type (Jordan & Mitchell, 2015). For instance, a notable example of an AI system that makes use of at least two types of learning simultaneously is AlphaGo[2]. It used deep supervised learning to learn how to play the game of Go from human players and then deep reinforcement learning that allowed the agent to learn the game based on the sessions of playing the game itself (Silver et al., 2017). Further, even though the majority of contemporary AI systems are based on the rational agent approach to intelligence, they also use a combination of skills typical to the other approaches (i.e., think and act humanly and think rationally). Thus, the manner in which contemporary AI systems are built is multi-layered and their functional processes are complex (i.e., it is not always clear how the model makes its predictions and inferences).

Second, the majority of contemporary models rely on large troves of both labelled and unlabelled data, which are used to continuously make inferences and predictions. Take for instance one of the currently most advanced natural language processing AI systems - GPT-3[3] (e.g., see Piper, 2020). Released in 2020 by the AI research lab Open AI, GPT-3[4] was trained on data found on the web. The model showed impressive results in a wide range of applications. It was able to generate essays and poetry and write new computer code (e.g., see

2 The AlphaGo attracted a lot of attention when it won against the world's Go champion in 2016.
3 GPT-3 is also the model behind ChatGPT, which managed to reach 1 million users in just 5 days in 2022.
4 GPT stands for Generative Pre-trained Transformer.

Heaven (2020; 2022) for an overview). To achieve its impressive performance, the model was trained on 45TB data and 175 billion parameters[5] (Brown et al., 2020). Due to the central part which data plays in the performance of similar ML models, the outcomes of these systems are directly dependent on patterns that can be deduced from human-generated data.

Third, most AI systems today are also continuously learning and adapting their behaviours to the data they are given or has been generated by their users. Consumer-facing recommendation models like the ones Spotify, TikTok, Netflix, YouTube and many more companies use exemplify this characteristic well. These models have collected enough data on each of their users' preferences so that they can accurately recommend the next video, movie, song, or book one should check out. For example, if one is to watch a YouTube or TikTok video on how to cook a "fancy vegan three-course dinner", chances are that the recommendation models these companies use will start recommending to the user to continue watching videos on how to cook vegan dishes. It will also show the user other types of videos that are closely related to the activity of cooking a fancy vegan meal, such as "how to set a formal dining table" or "how to store produce" or even "how to host a great dinner party". Hence, this characteristic makes these models susceptible to changes in the manner in which people use them (e.g., Microsoft's Tay discussed in the Introduction of the dissertation).

Conclusion

The purpose of this chapter was to provide an answer to the following research question: ***"What are the characteristics of contemporary AI systems?".*** In order to do so, we first presented a brief background on the four research directions to intelligence that have propelled forward the field of Artificial Intelligence: (1) thinking humanly, (2) acting humanly, (3) thinking rationally, and (4) acting rationally (i.e., the rational agent approach). From this overview became clear that most contemporary AI systems are manifestations of the rational agent approach to intelligence. We then provided a brief introduction of the four main

5 Parameters define how the model input is to be transformed to the desired output. As such, they are learnt from the data on which the model is trained.

types of Machine Learning: (1) supervised, (2) unsupervised, (3) reinforcement, and (4) deep learning. Although the clear categorisation of four distinct types is conducive to ease of explanation, most contemporary AI systems are usually a combination of two or more of these types of ML.

Finally, the chapter elaborated upon the three characteristics of AI systems that emerged from the two overviews: (1) AI systems' complex functional processes, (2) the high dependency on human-generated data and (3) the continuous process of learning and adapting the system's behaviour to fit to the manner in which humans use it. These characteristics serve as the starting point for Chapter 2, which elaborates upon the challenges that emerge once AI systems are implemented within a complex context. As such, these insights help us to further our efforts of answering the initial research question of this dissertation: ***"How can Design theories support the design and implementation of AI systems into complex contexts?"***

CHAPTER 2

CHALLENGES FOR IMPLEMENTING AI SYSTEMS INTO COMPLEX CONTEXTS

In 2015 Google's photo recognition model started labelling black people as "gorillas" (BBC, 2015). The case sparked multiple controversies and Google pledged to do its best to prevent this from happening again. However, after almost three years of working on the problem, the only solution Google was able to come up with to prevent the event from repeating was to remove the label "gorillas" from their datasets. Hence, even photos of gorillas would not be labelled as gorillas (Vincent, 2018). Then, in September 2020, a similar controversy emerged when it became evident that Twitter's photo cropping feature powered by an image recognition AI would always favour white people over people of colour and women over men (Hern, 2020). Twitter admitted to the mistake a few months later (BBC, 2021a). This pattern was reaffirmed when, in September 2021, Facebook's algorithm labelled a black man as a "primate" (BBC, 2021b).

As these examples show, the devising and maintaining of AI systems that are to be embedded into complex contexts is an inherently wicked problem (Stoimenova & Price, 2020). The solution to such problems is always a "one-shot operation" (i.e., every implemented solution can cause a number of unintended consequences that cannot be undone (Rittel & Webber, 1973, p. 163)), and the person who devises them has "no right to be wrong" as they are liable for the outcomes they create (p. 167). Further, the solution can *be* neither true nor false. It can only *be classified* as good or bad (p. 162).

The aim of this chapter is to deepen our understanding on why it continues to be challenging to ensure AI systems remain "good" after they are embedded into complex contexts. In order to do so, we are guided by the following research questions:

What are the challenges to implement AI systems into complex contexts?

What research approaches exist to address the identified challenges?

As such, the chapter is structured as follows. We start by introducing four principles that can guide the development of "good" AI systems. The introduction is followed by four sections – one for each of the principles. Each section is structured identically. Firstly, we describe a challenge stemming from the characteristics of AI models identified in Chapter 1: (1) AI systems' complex functional processes, (2) the high dependency on human-generated data and (3) the continuous process of learning and adapting the system's behaviour to fit to the manner in which humans (like to) use it. Once the challenge is established, an example is introduced to showcase how the challenge has manifested in the real world[1] and the impact it had on the complex context it was part of. Then, each of the chapter's four sections is concluded with a brief overview on existing research approaches from across different academic fields that address the identified challenge. The overview on research directions

1 It is worth noting that the manner in which we mapped each of the examples to only one of the principles Jobin et al. (2019) elaborated upon is relatively reductionist. The implementation of an AI system into a complex context is often accompanied by consequences that stem from all of the principles discussed in this chapter. Still, for ease of explanation, we have focused on one ethical principle per example.

is not complete since new research is published on a daily basis[2]. Still, it allows us to identify a clear gap in the literature which leads to the reframing of the dissertation's main research question.

The four principles for "good" AI

The discipline "concerned with what is morally good and bad" is Ethics (Britannica, n.d.). In fact, the predominant view on how AI systems should be implemented into a complex context is that it needs to be done in an ethical way (e.g., Rahwan et al., 2019). Over the past several years, a plethora of public and private initiatives have arisen globally to define values, principles and models for the ethical development and deployment of AI (Mittelstadt, 2019). All of them are aimed at regulating possible negative impacts and stopping potentially harmful technologies from entering the market.

In 2023, the advances of large language models such as the one used in ChatGPT[3] made the importance of developing good AI systems a topic widely discussed both by experts and laypeople alike. For instance, a letter signed by more than 1,800 people including Elon Musk, Steve Wozniak and engineers from Amazon, Meta, DeepMind, Google, and Microsoft urges for a six-month ban on AI more powerful than OpenAI's model GPT-4. The temporary ban, the letter stipulates, should be used so that AI labs and independent researchers can create and implement safety protocols. Such, arguably, could ensure AI systems that are aligned with societal objectives (Paul, 2023). While the letter garnered a lot of attention, no big technological company has announced they will introduce a temporary ban. In fact, prominent names in the field of AI have argued against the ban (see LeCun & Ng, 2023).

The importance of ensuring the AI systems we develop and deploy are *ethical*, has never been more pressing and well-understood. Therefore, in the remainder of the chapter, we discuss the four principles of ethical AI Jobin and colleagues

2 For example, the number of articles containing the key phrase "ethical AI" published from 2017 to 2022 is 300 000. Similarly, there are 869,000 papers that contain the key phrase "explainable AI" for the same period of time (according to a Semantic Scholar search carried out in September 2022).
3 ChatGPT is the fastest-growing web platform in history (e.g., see Hu, 2023),

(2019) put forward: (1) ***transparency***, (2) ***fairness***, (3) ***responsibility***, and (4) ***non-maleficence*** (i.e., causing no harm). These four principles emerged as a result of a rigorous scoping review on 84 different ethical guidelines from around the globe, published in both academic and grey literature (Jobin et al., 2019).

Transparency principle

Challenges for achieving transparent AI systems

While different guidelines impose various meanings onto the word "transparency", they often refer to efforts to increase the explainability and interpretability of the algorithmic model (Jobin et al., 2019). However, one of the main characteristics of most contemporary AI systems (e.g., deep learning) is that they have functional processes that generate complex and hard to interpret outputs (Rudin, 2019). The inability to explain and to understand these processes has resulted in the widely adopted term "**black boxes**", signifying the limited ability of humans to understand why an algorithmic model has produced a specific output. Since the nature of such algorithms is highly recursive and nonlinear, it does not lend itself easily to an explanation a human can understand (Li et al., 2018). Further, the reasoning behind why a certain output was produced is largely opaque (Rahwan et al., 2019). Therefore, when AI systems act in unintended manners, our ability to explain why (and consequently, prevent it from happening again) is limited due to the black-box nature of such algorithms.

Example of non-transparent AI systems' negative consequences

Ever since its inception in 2005, YouTube, like many other social media platforms, has increasingly relied on recommendation algorithms. These are optimised to ensure people will stay as long as possible on the platform by providing them with a selection of videos, personalised to their taste and preferences. The videos are supplied by the millions of creators on the platform. And for some of them, YouTube has become their primary means to earn a living, leading to the creation of an entirely new job title (i.e., "YouTuber"). This new job title can be very lucrative. For instance, the highest earning creator on

YouTube, MrBeast, made more than $54 million in 2021 alone (Spangler, 2022). However, in an attempt to continuously optimise their recommendation system, over the years YouTube kept tweaking its algorithm leaving all of its creators to guess how the system works and what kind of content it optimises for. This has led to numerous videos and articles giving potential explanations on how the algorithm behind the recommendation system might work. There are even multiple conventions where successful YouTubers share their experience with optimising their content for the algorithm (e.g., VidSummit in Los Angeles, the US). This never-ending race to figure out the objectives of the non-transparent model behind YouTube's recommendation system has real-world consequences for YouTubers. For instance, in recent years, many creators have come forward to discuss the depression and burnouts they have experienced while trying to understand how to optimise their content for the ever-changing model (e.g., Parkin, 2018).

Existing approaches to achieving transparent AI systems

It is important to ensure that the AI models we use are explainable and transparent. Not only in the case when we try to guess what the algorithm optimises for (like many YouTubers do), but also in order to understand the reasoning behind why a certain diagnose has been given to a patient, for instance. Scholars believe such explainability to be a way in which oversight can be enacted, behaviour can be anticipated and influenced, and negative unintended consequences can be mitigated (Jobin et al., 2019; Rahwan et al., 2019).

There are three main approaches to achieve that (e.g., Rudin, 2019; Murdoch et al., 2019): (1) by creating a *separate relatively simple algorithmic model* (e.g., linear regression) that takes a black-box model as its input and tries to explain it post-hoc (Samek et al., 2017); (2) by using only *inherently interpretable algorithmic* models when high-stake decisions are required (e.g., whether a person could be granted a parole or a loan) (e.g., Rudin, 2019); and (3) by *continuing to use black-box models* because they work differently than humans (Robbins, 2019). For example, while the classification of a mole as malignant is important since it "significantly affects the patient" (Vollmer, 2018), the

AI algorithm, which makes such a classification, works well because it does not use "human articulable" reasons for its classification (Robbins, 2019). However, according to some scholars (e.g., Floridi, 2011), such unexplainable algorithms should operate only within very well-defined boundaries so that their consequences can be managed.

Fairness principle

Challenges for achieving fair AI systems

The concept of fairness in AI is multifaceted and hence, invites diverse perspectives and interpretations, depending on the field to which an AI fairness scholar looks for a contribution. One dominant idea to ensure such fairness is the notion that for a solution to be fair the data it is trained on should not contain harmful biases (Jobin et al., 2019). However, AI models are typically built and trained on human-generated data. Consequently, these models more often than not reflect and/or amplify pre-existing human and structural **biases** (either as a result of the training data or the personal biases of the model's developers). These hidden biases can also introduce complex biases of their own once they start interacting with their contexts, usually at some point after their deployment (Dixon et al., 2018). Therefore, it is important to identify and eliminate harmful biases, such as gender or racial bias, upfront so that negative unintended consequences can be mitigated or even avoided. However, implicit biases in both the training data and the real time data fed to a model are difficult to identify in the type of non-binary data the majority of systems use today (Bellamy et al., 2018).

Example of unfair AI systems' negative consequences

In 2013, the Dutch Tax Office (i.e., Belastingdients) started using an algorithmic decision-making model that detects potential childcare benefits fraud at an early stage. The model used automated profile generation of potential fraudsters based on racial profiling (Amnesty International, 2021). The generated profiles supported the authorities' decision to stop the alleged perpetrators' child support and request the immediate return of the subsidies (in full). Soon, it became evident that the model extrapolated

the institutionalised racism on which it was trained. As a result, tens of thousands of families (most of whom from a minority background) were falsely persecuted. Many of them lost their homes, went through a divorce, and their children had to be placed in host families (Kleinnijenhuis, 2018). The revelations became known as the "Toeslagenaffaire" (i.e., The Benefits Scandal). These insights were met with a slew of resignations – first of ministers and then of the entire cabinet (Hanley, 2021). Further, the Tax Office was fined with € 2.75 million (ANP, 2021) and the victims of the biased algorithm were promised to get awarded € 30,000 each (Nu, 2022). An official investigation was launched by the government to look further into the matter (Kleinnijenhuis, 2022).

Existing approaches to achieving fair AI systems

The implicit biases in AI models' training and subsequently, real time data used after their deployment continue to be notoriously difficult to identify beforehand (Bellamy et al., 2018). However, as we saw from the example, once they get introduced in complex systems, the consequences can be devastating. Therefore, a burgeoning body of literature has emerged on devising ways to ensure algorithmic fairness. A large number of articles, for instance, focuses on quantifying the deviation between an AI model's predictions and a formalised metric of equality (e.g., men and women should be treated equally[4]) (Lee et al., 2021). According to them, algorithmic fairness should address the implicit bias that can be found in the data used to train and retrain AI models. In order to do so, multiple statistical metrics have been devised such as: (1) anti-classification, entailing the exclusion of protected characteristics like race, gender, or their proxies when deriving estimates; (2) classification parity, aiming for equivalent predictive performance across groups defined by protected attributes; (3) calibration, guaranteeing outcomes' independence from protected attributes after controlling for estimated risk; (4) equalised odds, gauging the ratio of false positive and negative rates; and (5) individual fairness,

4 Such all-encompassing notion of equality can also be damaging. For instance, in the criminal justice system, women are typically less likely to commit a future violent crime than men with similar criminal histories. While, generally, it is unfair to treat men and women differently, gender-neutral risk scores can systematically overestimate a woman's recidivism risk and can, in turn, encourage unnecessarily harsh judicial decisions (Rudin, 2019).

demanding equitable treatment of akin individuals (see Corbett-Davies and Goel (2018) and Lee et al. (2021). These (and other similar types) metrics are loosely derived from the concept of egalitarianism (ibid).

This statistical interpretation of fairness is appealing due to its simplicity and ease of application. Hence, it is not surprising that a plethora of open-source tools like Microsoft's Fairlearn and IBM's AI Fairness 360 have been released to aid AI developers in achieving such statistical fairness. However, different statistical measures are often at odds with one another. For instance, equalising false positive rates, false negative rates, and positive predictive values across protected groups simultaneously is unattainable (Chouldechova, 2017; Buijsman, 2023).

Therefore, a growing group of scholars argues for complementing such statistical measures with a more contextually aware and ethically informed approaches to AI fairness. For instance, we can look at the work of Buijsman (e.g., 2023), Ruf and Detyniecki (2021) and Lee and colleagues (2021) for initial frameworks on how philosophical approaches can be used to make the trade-offs between algorithmic fairness measures (e.g., for which measure to optimise), as well as how to balance model's accuracy and fairness.

Responsibility principle

Challenges for achieving responsible AI systems

The responsibility principle most often signifies acting with integrity (Jobin et al., 2019). This is in line with a standpoint held by many scholars across different academic fields that an ethical/responsible AI is the AI system that is **aligned with human values** (e.g., Santoni de Sio & van den Hoven, 2018)[5]. Yet, defining a set of values that are globally applicable continues to be difficult, if not impossible (Awad et al., 2018). Further, when overarching value definitions are attempted, they usually assume a deterministic worldview (i.e., AI is seen as a force of change to which humans must adapt) (Greene et al., 2019). Consequently, the guidelines and values become conductive to decision-

5 See also Mittelstadt (2019), Jobin et al. (2019), Floridi & Cowls (2019), Rahwan e al. (2019).

making power being delegated to a narrow circle of experts (ibid). Further, it is challenging to ensure the accommodation of emerging values, especially the ones that cannot be anticipated during the design stage of a solution (de Reuver et al., 2020). Finally, because of the complex structure of most AI models, it continues to be difficult to embed the desired human values and ethics within the development cycle and the AI system itself (Mittelstadt, 2019).

Example of non-responsible AI systems' negative consequences

At the beginning of the Covid-19 Pandemic, the UK government decided that the exams of high school children in the country would be cancelled (Bedingfield, 2020). Instead of exams, the government decided to use a combination of teachers' predictions on what they thought their students might score on the exam and an algorithm that was designed to bring an objective perspective (ibid). The algorithm was based on multiple different data sources and created three sets of grades: (1) a distribution of the grades a student received on the subject in previous years, (2) a predicted distribution of exam grades for students from previous years (which were then compared to the actual exam grades these students received to verify the accuracy of the model) and (3) a predicted distribution of grades for the current students. The algorithm then gauged the difference between the predicted distribution of current students with the one of past students and used this to adjust the prediction. Finally, each student was assigned their grade based on a ranking their teacher had provided to the algorithm (i.e., the student ranked as number 1 will get the highest possible grade in the distribution). Once the results of the algorithm came out, it turned out that the model had downgraded 39% of the high grades initially estimated by the teachers. Further, on average, students in state schools received lower grades than those in private ones, further perpetuating pre-existing socioeconomic inequality (Porter, 2020) and causing a grading crisis in the UK (Shead, 2020). In response, the government scrapped the algorithm and decided to go with the subjective prediction teachers made. According to experts, the manner in which this algorithm was developed failed to consider the views and values of its stakeholders and, as such, was applied irresponsibly. They claim that many of the observed problems could have been

predicted and resolved beforehand if the algorithm developers had carried out proper consultations with ethics, education, and statistics experts, who could have pointed at potential issues (Bedingfield, 2020). In fact, the Royal Statistical Society had warned the government about potential problems with the AI system months in advance (Satariano, 2020).

Existing approaches to achieving responsible AI systems

There are three prominent approaches attempting to solve this challenge: (1) *human-centered AI/ML (HCML)*[6] (e.g., Lovejoy, 2018; Riedl, 2019), aimed at making the output of an AI model easier to understand by its users, ultimately ensuring a seamless user experience and personalisation (Chancellor et al., 2019); (2) *Value Sensitive Design* (VSD)[7], aimed at intentionally embedding desirable values[8] in the design of artefacts; and (3) *ethical guidelines*, aimed at providing guidance to developers and companies[9] on which values to consider[10] when developing AI systems. Such ethical guidelines also play a crucial role in both Value Sensitive Design and Human-Centered Machine Learning. For instance, an important part of the VSD approach is the so-called conceptual inquiry into the ethical and philosophical issues, which include the consideration of ethical guidelines (Friedman & Nissenbaum, 1997). Similarly,

6 This research approach also can be found under the names of Augmented Intelligence (Brown, 2017), human-AI interaction (Amershi et al., 2019) or interactive ML (Fails & Olsen, 2003), and sometimes Useful AI (Mosqueira-Rey et al., 2022).

7 This approach also manifests itself under different names that include but are not limited to: Value Sensitive Design (VSD) (Friedman & Kahn, 2003; van Wynsberghe, 2013; Davis & Nathan, 2015; van den Hoven et al., 2015), Values in Design (Nissenbaum, 2001), Design for Values (van den Hoven et al., 2015), and Worth-Centred Design (Cockton, 2009). All of them share at least four key claims: (1) values can be expressed and embedded in technology, (2) technologies have real and sometimes non-obvious impacts on those who are directly and indirectly affected, (3) explicit thinking about the values that are imparted in technical design is morally significant, and (4) value considerations should be surfaced early in the technical design process (van den Hoven, 2015).

8 Values here are defined as entities that appear in technologies, built consciously or unconsciously by developers, and made material through a technology's features (Friedman & Nissenbaum, 1997; Flanagan et al., 2008).

9 Many such guidelines (in academia and (non-profit) organizations) have been developed in the past few years (for a comprehensive review on these see Mittelstadt (2019)).

10 In May 2021, the European Union became the world's first governmental body to release a draft proposal for high-risk AI system regulation (Benjamin et al., 2021).

ethical guidelines can inform the initial stages of the HCML approach by informing designers on which ethical values they need to consider[11].

Non-maleficence principle

Challenges for achieving non-maleficent AI systems

The ethical principle of non-maleficence pertains to the idea that AI systems should never cause foreseeable or unintentional harm (Jobin et al., 2019). However, this is difficult to achieve since contemporary AI systems are continuously learning either from the data they have been fed during training or by the manners people use them. As such, they are defined by a strong **behaviour-use interdependence** (i.e., the way an AI system is used heavily influences the way it behaves and vice versa). This challenge is especially visible in recommendation systems relying (at least partially) on reinforcement learning. For instance, the type of songs one listens to on Spotify heavily influences the songs its algorithm will suggest in the future. The same goes for the type of videos one watches on TikTok or YouTube, or the type of news one reads on Facebook. However, it can also be observed in simpler application like the Microsoft's Twitter bot Tay discussed in Chapter 1 (p. 5). Further still, AI models can be easily influenced (even duped) once their input data differs too much from the data they were trained on. For example, the unusual online behaviour of millions of people during and due to the Covid-19 pandemic confused the predictive algorithms that run behind the scenes in inventory management, fraud detection, marketing and product or movies recommendation (Heaven, 2020). As a result, for instance, the sales forecast on which companies relied to reorder stock no longer matched the reality (ibid). Moreover, researchers have already demonstrated how to deceive facial-recognition systems by sticking a printed pattern on glasses or hats (Komkov & Petiushko, 2019), "make" speech-recognition systems hear phantom phrases by inserting patterns of white noise in the audio (Cisse et al., 2017)

11 Next to these approaches, another emerging area of research is the so-called human-in-the-loop approach where human knowledge is incorporated into the model-building process (e.g., during data processing, data annotation and iterative labelling) (Wu et al., 2022). An extensive overview on the manners in which humans can be involved in the training process of AI systems can be found in Mosqueira-Rey et al. (2022) and Wu et al. (2022)

and trick personal assistants like Google Home or Alexa into "hearing" voice commands simply by pointing an infrared laser at them (Greenberg, 2019). This strong interdependence between an AI system's behaviour and its use makes it increasingly difficult to anticipate and mitigate the wide array of potentially harmful consequences such systems could produce.

Example of AI systems' negative consequences

Social media platforms like Facebook and YouTube rely heavily on recommendation systems which are a combination of different types of machine learning (e.g., deep learning and reinforcement learning). Thus, they try to continuously learn what their users want to see and provide them with a personalised feed. The continuous adjusting of the behaviour of these systems to adapt to the way they were being used was one of the causes for both platforms to come into the crossfire. In 2016, their algorithms became a means for mass misinformation and manipulation, resulting in the now infamous term "fake news." (Sydell, 2017). Despite the slew of public debates, initial legislation and the pledge of both Google and Facebook to prevent misinformation, the spread of fake news after the mass shooting in Las Vegas in October 2017 proved very difficult to prevent (ibid). Further still, in the midst of the Covid-19 pandemic in 2020 and 2021, despite all efforts of both companies, the "fake news" phenomenon continued to reign, often leading to cases such as people lighting 5G network antennae on fire (Satariano & Alba, 2020), the rise of conspiracy movements such as QAnon (e.g., Nagesh, 2021), fights over toilet paper (Mao, 2020), and the Capitol Riot on January 6th, 2021 in the US (Timberg et al., 2021). Arguably, none of the aforementioned consequences were intended by Google or Facebook. Yet, despite their efforts, it proved to be difficult to prevent their systems from spreading misinformation.

Existing approaches to achieving non-maleficent AI systems

To our knowledge, no existing approaches explicitly address the challenge of behaviour-use interdependence. Although Rahwan and colleagues, in their seminal review for Nature in 2019, argue for establishing a new discipline that could systematically study machine's behaviour and use (i.e., akin to "how

ethology and behavioural ecology study animal behaviour" (p. 477)), such discipline has not been clearly established yet. Further, some types of the human-in-the-loop approach like interactive machine learning[12] or useful AI[13] can appear to be addressing the behaviour-use interdependence challenge. Namely, they do focus on detailing how human feedback can be used to train the model before it is deployed. However, none of them specifically addresses the interdependence between behaviour and potential use once the model has been deployed and how this can lead to a host of unintended outcomes (see Mosqueira-Rey et al., 2022 for a comprehensive overview). It is precisely during the interaction between the AI system and the variety and deep complexity of human contexts in which it may be applied that many harmful outcomes start to emerge. In fact, it is usually indeterminate how others will use an AI system and what the impact of that use would be until the solution becomes part of a broader context. This makes it difficult to anticipate which consequences need to be mitigated prior to implementing the solution. Consequently, it makes it challenging to ensure an AI system we design will remain non-maleficent throughout its entire lifecycle.

Main research question of the dissertation

So far, we introduced four challenges to implementing AI systems into complex contexts: (1) ensuring transparency, (2) mitigating implicit biases, (3) aligning with human values, and (4) addressing the behaviour-use interdependence. While multiple scholars across different academic fields address the first three, no current approaches, to our knowledge, explicitly address the fourth one. Yet, as it could be seen from the examples discussed so far, it is important to ensure that we design the behaviour of the AI system in such a way that it will trigger its intended use once it is deployed. Further, we also need to ensure that the AI will continue behaving in an intended way even after different types of unexpected uses are performed by humans or other AI systems. Therefore, it

12 This is a type of approach where the human iteratively teaches the AI model so that its learning behaviour can be optimised (Mosqueira-Rey et al., 2022).
13 Useful AI refers to the approach where once deployed, AI systems can receive corrections from their users that can be used as additional training data (e.g., CAPTCHAs) (Mosqueira-Rey et al., 2022).

becomes imperative to develop theoretical models and methods that support us in simulating (potential) uses as early as possible so that we can fine-tune the behaviour of the system to trigger the desired use. This, in fact, becomes the main premise of this doctoral dissertation.

The field of Design is well-positioned to address the challenge of behaviour-use interdependence. Namely, the notion of two interdependent entities parallels one of the field's most influential ideas: ill-defined problems become better understood through the iterative process of designing a solution and vice versa. Consequently, they are never fully defined, and evolve over time. This interdependence can be found at the core of some of the most influential Design theories such as the theory of co-evolution (Maher, 2000; Dorst & Cross, 2001), Simon's Rational Problem Solving (1996) and Schön's Reflective Practice (1983) (see Crilly, 2021). Furthermore, both *"behaviour"* and *"use"* are central to the cognitive act of designing. Both can be found in what many design scholars consider the core of design – ***abductive reasoning*** (e.g., March, 1984)[14]. For instance, according to Roozenburg[15] (1993), the variables of *behaviour* and *use* are instrumental to our ability to generate new solutions. Further, behaviour can be easily likened to what Roozenburg (ibid) defines as the **mode of action** or the "(functional) behaviour of the artefact in response to influences exerted on it from its environment" (p. 12). On the other hand, the manner in which an AI system is used coincides with the definition of **actuation** Roozenburg (ibid) uses: the action that the user applies onto the artefact that allows it to function and be "connected" (p. 13) to its immediate environment. As such, the field of Design Theory could provide us with insights on how the behaviour and use influence each other.

14 See also Zeng and Cheng (1991); Roozenburg (1993); Takeda (1994); Dorst (2011); Dong et al. (2015); Kroll & Koskela (2015); and Hatchuel et al. (2018).
15 Roozenburg's work on abductive reasoning in Design is widely considered as seminal (e.g., Kroll & Koskela, 2015).

Table 1: an overview of existing challenges and research approaches (and the field to which these approaches belong) related to the four ethical principles Jobin et al. (2019) put forward and the AI characteristics identified in Chapter 1.

Principle	AI characteristic	Challenge	Approaches	Aim	Authors	Field
Transparency	Complex model structures	Black box	A post-hoc model to explain the black box model	Create a relatively simple algorithmic model that takes a black-box model as its input and tries to explain it post-hoc.	e.g., Gunning (2017); Murdoch et al. (2019); Samek et al. (2017); Hagras (2018).	Computer Science
			Inherently interpretable models	Create algorithmic models that are easy to understand and interpret, yet, sophisticated enough to properly fit the data and provide an accurate result.	e.g., Rudin (2019), Lipton et al. (2018).	
			No need to understand the model	Create black box models that operate within boundary conditions set by humans.	e.g., Robbins (2019), Floridi (2011).	
Fairness	High dependency on human-generated data	Bias	Mathematical models (e.g., anti-classification; classification parity; calibration)	Use formal mathematical definitions of fairness to evaluate existing systems and when engineering new algorithms.	e.g., Dixon et al. (2018); Bellamy et al. (2018); Holstein (2019); Recasens et al. (2013).	Computer Science
			Contextual approach	Identify existing biases by considering fairness as a contextual construct.	e.g., Recasens et al., 2013; Rudin, 2019; Holstein, 2019; Lee et al., 2021.	Ethics

Responsibility	Complex model structures	Human values	Ethical guidelines	Devise a set of ethical guidelines that support AI development in ensuring human values are embedded into the AI system.	e.g., Mittelstadt (2019); Google (2019); Microsoft (2019); Greene et al. (2019); Smallman (2019); Awad et al. (2018); Benkler (2019).	Ethics
			Human-centered AI	Ensure AI systems are aligned with human values and remain human-centred.	e.g., Riedl (2019); Chancellor et al. (2019); Google's PAIR (2019); Lovejoy (2018); Brown (2017); Amershi et al. (2019); Yang et al. (2020).	Human Computer Interaction (HCI)
			Value Sensitive Design	Intentionally embed desirable (ethical) values into technical artefacts	e.g., Flanagan et al. (2008); Friedman & Kahn (2003); Davis & Nathan (2015); van Wynsberghe (2013); Nissenbaum (2001); Cockton (2009); Santoni de Sio & van den Hoven (2018); de Reuver et al. (2020).	
Non-maleficence	Continuous learning cycle	Behaviour-use interdependence	No existing approaches	n/a	n/a	n/a

Taking into consideration the aforementioned reasons, we adopt the behaviour-use interdependence challenge as the vantage point for the rest of the dissertation. This allows us to formulate the main research question of the dissertation. Namely:

"How can a theoretical model be designed that supports the early simulation of AI systems' behaviour-use interdependence by utilising Design theories?"

This new, more detailed, research question builds upon the initial one we presented in the dissertation's Introduction (i.e., ***"How can Design theories support the design and implementation of AI systems into complex contexts**?")*.

Consequently, this dissertation encompasses three distinct contributions. First, to the field of AI systems design, the aim of this dissertation is to utilise Design theories in order to address the behaviour-use interdependence challenge we outlined in the chapter. Second, the use of these theories also necessitates an exploration on whether and how they should be extended to address the design of the new material (AI systems). Finally, the dissertation will discuss the implications of the other two contributions to the existing body of research in Design Theory.

Conclusion

This chapter was based on the premise that in order to implement an AI system into a larger context, it should be ethical. Stemming from this, we adopted the four ethical principles Jobin and colleagues (2019) discussed as point for convergence among the different global ethical guidelines: **transparency**, **fairness**, **responsibility**, and **non-maleficence**. These four principles then served as foundation to answer the research questions of the chapter:

What are the challenges to implement AI systems into complex contexts?

What research approaches exist to address the identified challenges?

The answer to the first question came from the combination of the ethical principles Jobin et al. (2019) put forward and the characteristics of AI systems identified in Chapter 1: AI's **complex functional processes**, the **high**

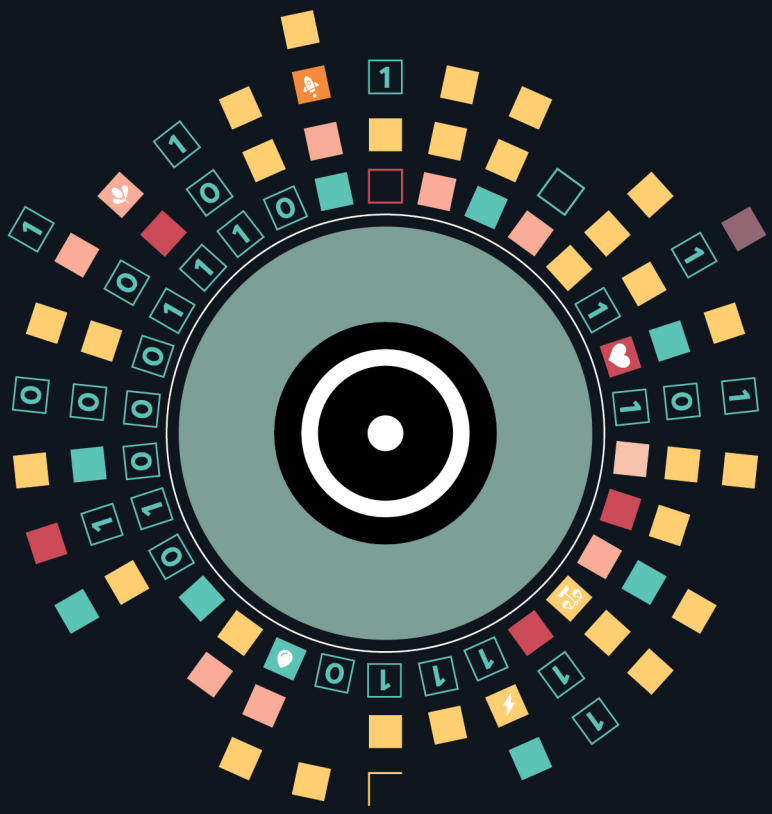
dependency on human-generated data, and AI's **continuous learning cycle**. By combining the ethical principles and AI's characteristics, we arrived at the following four challenges: (1) **ensuring transparency**, (2) **mitigating implicit biases**, (3) **aligning with human values**, and (4) **addressing the behaviour-use interdependence**. These four challenges are interrelated and heavily influence each other. For instance, if we are to create an algorithm that can explain every black box model there is, we would also be able to identify where the potential biases are and thus, make the solution fairer. Therefore, the advancements made on achieving all four challenges are equally important if we are to ensure AI systems will not create far-reaching harmful outcomes.

The answer to the second research question is directly related to that of the first one. For each challenge, we identified existing research approaches from across different academic fields trying to resolve it. However, as it can be seen from the overview in Table 1, there are research approaches for all challenges, except for the fourth one – **behaviour-use interdependence**. This gap in the literature presents us with a clear opportunity to contribute to the efforts of designing ethical AI systems that are to be implemented into complex contexts. We then pointed to the central role in the **cognitive act of designing both behaviour and use play**.

Finally, these new insights served as a foundation for the reformulation of the dissertation's main research question. Namely: ***"How can a theoretical model be designed that supports the early simulation of AI systems' behaviour-use interdependence by utilising Design theories?".*** This new question also brings us to the close of Part I (i.e., *Set the scene*) of the dissertation. In Part II, we explore Design theories that allow us to provide an initial answer to the main question.

PART II

INITIAL THEORETICAL MODEL



FINAL THEORETICAL MODEL

CASE STUDY

EARLY SIMULATION OF AI'S BE-HAVIOUR-USE INTERDEPENDENCE

PART III EXTENDED THEORETICAL MODEL

CHAPTER 5

Introduction to the initial theoretical model that provides an answer to the main research question.

INITIAL THEORETICAL MODEL

CHAPTERS 3 & 4

Theoretical exploration of the fundaments of Design theory (Chapter 3) and Prototyping (Chapter 4) which can help us answer the main research question.

FUNDAMENTS OF DESIGN THEORY

PROTOTYPING FOR EARLY SIMULATION

PART II INITIAL THEORETICAL MODEL

INSIGHTS

Characteristics of AI systems: (1) complex functional processes, (2) high dependency on human-generated data and (3) continuous process of learning and adapting the system's behaviour to fit the way humans use it.

The challenges to implement AI systems in existing contexts: (1) ensuring transparency, (2) mitigating implicit biases, (3) aligning with human values, and (4) addressing the behaviour-use interdependence.

MAIN RESEARCH QUESTION

CHALLENGES FOR IMPLEMENTING AI

WHAT IS AI?

PART I SET THE SCENE

How can Design theories support the design and implementation of AI systems into complex contexts?

CHAPTER 3

THE FUNDAMENTS OF DESIGN THEORY

In the previous part of the dissertation, we sketched out the issues a new theory can address as well as the potential importance the research of these issues carries. This background understanding allowed us to define the main research question of the dissertation: *"How can a theoretical model be designed that supports the early simulation of AI systems' behaviour-use interdependence by utilising Design theories?"***.** In this chapter, we start to unpack the question by addressing its second part: "***behaviour-use interdependence"*** and ***"Design theories***". Therefore, the following research question guides this chapter:

"How do Design theories address the behaviour and use of solutions?"

The chapter is divided in two sections. First, Section I, presents and compares the two paradigms of Design Theory that have influenced the developments in the field since its inception: the Rational Problem Solving (Simon, 1969, 1981, 1996) and the Reflective Practice (Schön, 1983). Such insight is needed to provide a foundational understanding of the Design Theory field. Using this basis, Section II then elaborates upon the type of reasoning that has been widely considered as the core of design – abductive reasoning. As we already

contended in Chapter 2, the theory of abductive reasoning in Design is well-positioned to provide us with insights on the relationship between behaviour and use. Such insights will also further strengthen our understanding on how Design theories (as well as the concepts of behaviour and use) support the design of a new solution. The chapter is concluded with an overview on the insights that emerged from both of its parts.

Section I: The two defining paradigms of Design Theory

The developments in the Design Theory field have historically been defined by two theoretical paradigms: Rational Problem Solving (Simon, 1969, 1981, 1996) and Reflective Practice (Schön, 1983). The two paradigms have often been seen as diametrically opposite by a number of design scholars (e.g., Schön, 1983; Cross, 2007). Yet, they do share some similarities, especially if we are to consider the second and third edition of Simon's seminal work, The Sciences of the Artificial (as we do in this chapter) (Meng, 2009). The degree of similarity or differentiation of the two paradigms falls outside the scope of this dissertation. Still, when designing new theoretical models which can support the design process, we need to consider both paradigms (Roozenburg & Dorst, 1998). Both have influenced the developments in the field since its inception and provided the theoretical background against which different design practices, procedures and principles have been developed. Therefore, they provide a fruitful basis for understanding the existing approaches to designing new solutions.

Rational Problem Solving

This paradigm was mainly developed by Herbert Simon (1969, 1981, 1996) and particularly the ideas he put forward in his seminal book "The Sciences of the Artificial". When first published in 1969, his work built upon views put forward by Alexander (1964) and the prevailing positivistic idea at the time that the design process can be rationally and systematically analysed and described (e.g., Zeng & Cheng, 1991; Cross, 1993). He elaborated upon them by using insights from psychology and the field of AI. In 1981, in the second edition of his book, these ideas evolved to include the notion of the continuous "open search for new goals whilst designing" (Meng, 2009, p. 65) – an idea

resembling the constructivist theory of Schön (1983). As such, Simon's theory introduced a manner for describing design within the paradigm of technical rationality. It also provided rigorous foundations for much of the existing knowledge in Design Methodology (Dorst, 1997). In this chapter, we refer to the third edition of his book published in 1996 as it provides an overview of Simon's most recent ideas and theories.

To Simon, the design process is best represented as a parallel search process. When designing one needs to simultaneously explore multiple combination of components (which Simon terms "assemblies" (p. 124)) while searching for a solution that can satisfice the given goal (i.e., purpose) and requirements. The latter are provided by the client. When designing artefacts, the clients are usually easily discernible (i.e., they are the ones who initiated or paid for the project). When designing complex systems such as "rebuilding the center of a city" (p. 163), however, there are multiple clients, and it is up to the designer to decide which priorities and interests she should try to satisfice. Simon draws similar distinctions when it comes to the goals the clients provide. The design of an artefact is accomplished by trying to satisfice a goal that remains the same throughout the entire design process. When designing complex systems, however, the goal is clear and stable, but only at the beginning. It can change over the course of the design process. This happens since every interaction with the problem and its potential assemblies allows the designer to reach her initial goal, but also helps new goals to emerge. Simon equates this process of emergence to "painting in oil" (p. 163).

The search process and consequently, the emergence of new goals, is propelled forward by the identification of (unintended) consequences that emerge as a result of following each alternative. These consequences lie well into the future and therefore, they cannot be easily predicted upfront. Hence, according to Simon, one needs to look for ways to receive immediate feedback on the consequences of a solution. Thus, establishing robust feedback mechanisms is important. Such could allow the artefact to continuously respond to discrepancies between the system's actual and desired states, and then adapt

to fluctuations in the environment. Simon, however, does not elaborate on how such feedback mechanism ought to function and be established.

Finally, according to Simon, the design process ends when the designer has managed to find a solution (i.e., "an assembly") that satisfices the given goal and requirements. In order to decide whether the found assembly could satisfice these, one needs to consider only three variables: the goal or purpose of the artefact, its internal structure, and the environment in which it will be placed. For instance, if a clock is to fulfil its purpose ("*tell time*"), its structure (i.e., "*arrangements of gears and the application of forces*") needs to be well-suited to its environment (i.e., "*a clock meant for ships requires properties that are irrelevant for the landlubber's clock"*) (Simon, 1996, p. 6). The resulting solution, however, can never represent the optimal outcome since one can never know enough about a situation to assert with certainty that a solution provides the best fit. Hence, the designed solution can be assessed not as "the best", but as "better" or "worse" than the other proposed alternatives.

Reflective Practice

This paradigm was introduced by Donald Schön (1983) in his seminal book *"The Reflective Practitioner"* as what he claimed to be a "diametrically opposed" view on design to that of Simon
. The Reflective Practice provides a different, a constructivist and intuitive way of looking at designing. One that accounts for the iterative process typically employed while designing and that is more representative of the ambiguities and complexities of the everyday design situations (Roozenburg & Dorst, 1998).

According to Schön, design is a reflective conversation between the situation at hand and the designer where the former talks back to the latter. The design process, therefore, commences with the deliberate act of problem setting. While the design problem is specific and preliminary set (e.g., by the client) in Rational Problem Solving, within the paradigm of Reflective Practice, design problems are seen as situated, unique and never fully known. Hence, it is the designer's job to understand the problem better. According to Schön, she can do so by naming the things she will pay attention to in the problem situation and

then impose a suitable frame onto them. The latter is rooted in the knowledge a designer already has (i.e., what Schön terms a designer's underlying theory (p. 153)) but also corresponds with the designer's own goals and her view from which she can approach the problem. Thus, the designer needs to deliberately construct, shape, and change the problem setting. With that, she becomes part of the problem setting, too.

Once the designer has selected a frame, it becomes the starting point for the initiation of a move (i.e., the deliberate action the designer takes to both understand the situation and change it, so it fits better the proposed frame). She enacts such moves through experiments that could anticipate/simulate "what consequences and implications can be made to follow from it" (p. 131). According to Schön, three distinct types of experiments happen during a move: 1) exploratory experiment (when action is taken to see what its consequences might be without having a preconceived predictions or expectations); 2) move-testing experiments (when action is taken to produce an intended change); and 3) hypothesis testing experiments (when action is taken to confirm an existing hypothesis). Despite the different purposes these types of experiment serve, however, according to Schön, they occur simultaneously with each enactment of a move.

The unintended consequences or side effects that emerge after a move, Schön terms as "surprises" (p. 153). According to him, these surprises propel the act of designing forward and trigger the process of reflection-in-action. During the design process the designer actively tries to make sense of the situation at hand and reflects on the knowing that has been implicit in her actions . To Schön, typical questions designers ask themselves during such reflection are focused on elaborating upon the features they notice, the criteria they use to select these features, and the manner in which they are framing the problem they are trying to solve. During this, the designer not only asks herself "*Do you get what you intend?*" but also, "*Do you like what you get?*" (p. 146).

If the elicited surprises are positive and desirable, then the designer affirms the move. However, when the move and consequently the experiment have produced a negative surprise – the designer negates the move and selects

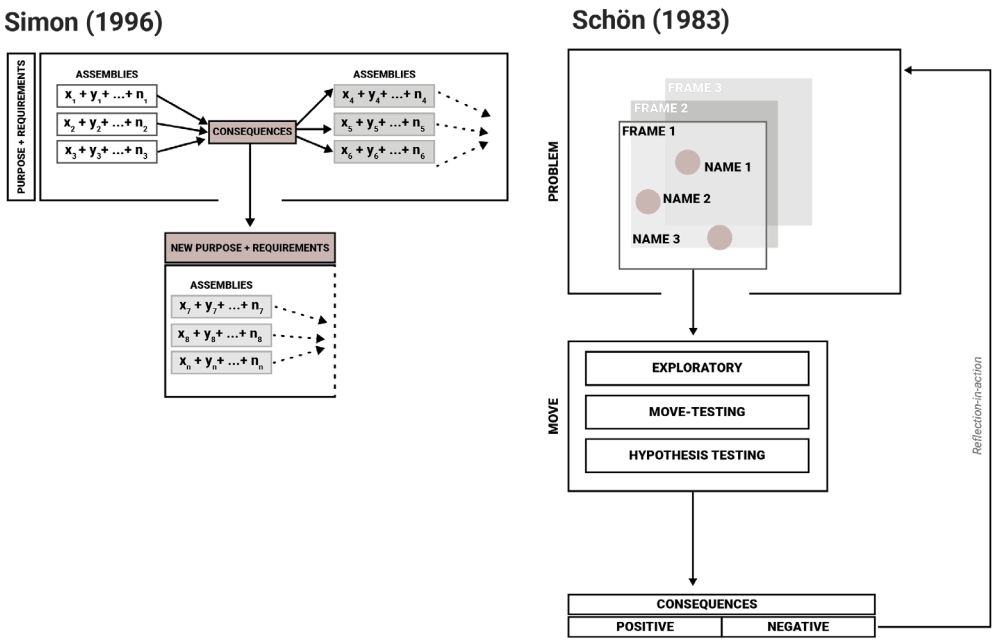
a new frame. Therefore, the process of framing, moving, and reflecting commences again. During it, the designer tries to change the situation so that it fits the new frame she selected. As such, the iterative design process unfolds gradually through the clarification of both the problem and solution spaces (akin to the theory of co-evolution of problem-solution space (Dorst & Cross, 2001)).

Overview of the paradigms

The paradigms of Rational Problem Solving and Reflective Practice have largely defined the theoretical developments in the field of Design. As such, when designing new theoretical models which can support the design process, we need to consider both paradigms. It is important to do so since they can, jointly, lead to the better understanding of design procedures, principles, and practices (Dorst & Dijkhuis, 1995; Roozenburg & Dorst, 1998). In Figure 8, a visual overview can be found on the manners in which both Simon and Schön see the design process.

There are three insights that emerged from the overview presented thus far that can aid us when designing a new theoretical model. First, to both, Schön and Simon, establishing a feedback mechanism of how the solution performs is important so that its goal can be continuously updated (for Simon) and so that we can have a better understanding of the problem (for Schön). Simon does not elaborate on how such feedback mechanism ought to function and be established. However, as it can be seen from Figure 8, it is the consequences of the solution that could be seen as providing such feedback and trigger another search process for different assemblies. Similarly, for Schön, it is the surprises (both positive and negative) that trigger the process of reflection-in-action where the situation “talks back” to the designer. This allows the designer to better understand the problem and create new frame on how the problem might be approached. In the process of identifying these surprises, Schön also puts central the notion of experiments (as part of what he terms a “move”).

Figure 8: Visual representations on Simon (left) and Schön's (right) paradigms.



Second, both scholars consider the role of the designer. For Simon, the designer plays a role only when she needs to choose who the client is, and which priorities and interests she should try to satisfice during the design of a complex system. For Schön, the manner in which a designer chooses to approach a problematic situation is always rooted in her own goals. Thus, she becomes part of the problem setting. This different view on the role a designer plays is also reflected in how the two scholars determine when the design process ends. To Simon, it ends when the designer has managed to find a solution that satisfices the goal and requirements. According to Schön, however, it is the designer who actively decides whether to end the design process by looking at the surprises her moves have created and reflect on them.

Finally, both scholars see the unintended consequences a solution has produced (i.e., surprises for Schön), as a natural and needed outcome of the design process. In fact, to Schön, such consequences are central to the process of design. They are the ones that allow the designer to gradually construct, shape and understand the situation better so that she can change it to fit the selected frame or find a better frame that suits the situation.

Section II: The core of design

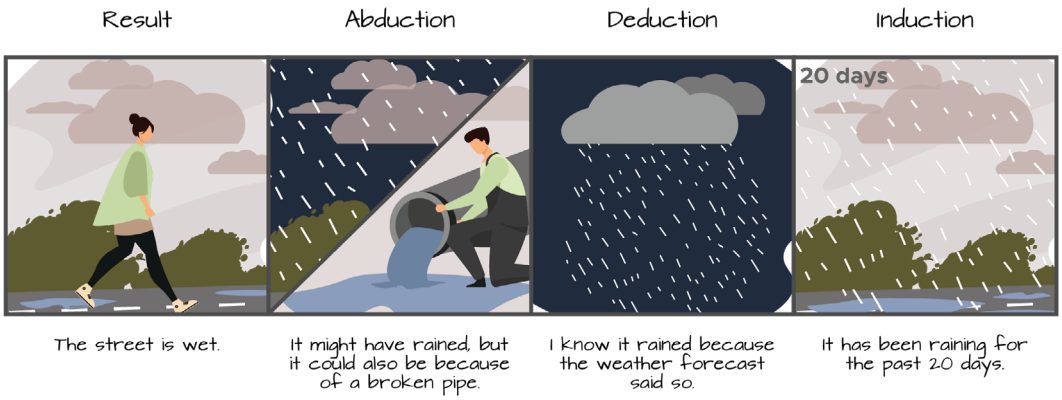
The research question this chapter aims to answer is: ***"How do Design theories address the behaviour and use of solutions?".*** The two paradigms presented in Section I provide a general understanding on the two main approaches to designing. Yet, they do not indicate how the behaviour and use of a solution are related to its design. In order to answer this, we need an in-depth knowledge on the core of design – abduction, as briefly discussed in Chapter 2.

It is widely agreed upon that the reasoning mode that defines the act of designing a new solution is abduction (e.g., March, 1984)[1]. Abduction was first introduced as a distinct type of reasoning by Peirce (1994) to denote the synthetic thinking employed in producing new insights. As such, according to Peirce, abduction[2] represents the "act of insight" that "comes to us like a flash" (5.181), "the process of forming an explanatory hypothesis" (5.171) and "the only kind of argument which starts a new idea" (2.96). Hence, abduction is distinctly different from the other type of synthetic reasoning (i.e., induction). Namely, abduction allows us to suppose that given an existing general rule or theory, a phenomenon exists that could explain our observations. Induction is the inference that allows us to arrive at a general law that can account for our observations. Both abduction and induction are distinctly different from the analytical type of reasoning – deduction – which supports us in applying a general rule to a specific case to reach a specific result (see Image 2).

1 Other authors that agree on the importance of abduction in Design include Zeng & Cheng (1991); Roozenburg (1993); Takeda (1994; 2001); Liedtka et al. (2007); Kolko (2010); Dorst (2011); Kroll & Koskela (2015); Dong et al. (2016); and Verganti et al. (2020).

2 The definition of abduction Peirce introduced is also close to the creative psychology constructs of sudden insight and aha moment (e.g., Akin & Akin, 1996)

Image 2 An explanation on how one can use the different types of reasoning to reach the conclusion that it rained when they see a wet street.



According to Peirce, the relationship among the three can be defined in the following manner: "Deduction proves that something ***must*** be; Induction shows that something ***actually is*** operative; Abduction merely suggests that something ***may be***" (5.171).

The notion that abduction is the key reasoning mode of design was first introduced by March[3] (e.g., Roozenburg, 1993; Kroll & Koskela, 2015). The claim was subsequently adopted and discussed by numerous design scholars[4]. According to them (e.g., Zeng & Cheng, 1991), however, while the abduction Peirce describes is invaluable in traditional problem solving, it fails to fully capture the manner in which hypotheses[5] are formulated as

3 It is important to note that abductive reasoning continues to stir an ongoing debate around its properties and nature, unlike the other, better-known, reasoning patterns (i.e., deduction and induction) (Kroll & Koskela, 2015). As such, especially in Philosophy of Science, abduction is widely discussed, and different types of abduction have been introduced that provide (richer) insights to the notion of abduction in general. A noteworthy example of this is the work of Magnani (e.g., 2007).

4 See also Zeng and Cheng (1991), Roozenburg (1993), Takeda (1994), Dorst (2011), Dong et al. (2015), Kroll & Koskela (2015) and Hatchuel et al. (2018).

5 It is important to note that the type of hypotheses we are referring to here are different than the ones employed in science. March (1984), for instance, claims that scientific hypotheses are distinctly different than the ones employed in Design. While the former represents a "general principle induced from particular events and observations" (p. 268), the latter are used as a "particular instance produced from a general notion and specific data" (p. 269).

"*tentative descriptions for solutions to design problems*" (Roozenburg, 1993, p. 4). Therefore, multiple Design scholars presented different classifications of the manner, in which design abduction is different than the one introduced by Peirce.

A well-accepted differentiation between the type of abduction Peirce introduced and the one employed in design is given by Roozenburg (Kroll & Koskela, 2015; Dong et al., 2015). In his seminal paper from 1993, Roozenburg likens Peircean abduction to a term Habermas (1978) introduced – ***explanatory abduction***. Explanatory abduction, Roozenburg claims, can be formalised as a logical explanation of an observation used to identify the use of a known to be true principle. This can be a law or theory such as, "*If a motor has no gas, then it does not start*". Therefore, it allows us to reason from the observed effect ("*The car doesn't start*") to a possible cause ("*The tank is empty, I guess*") (Roozenburg, 1993, p. 10). As such, explanatory abduction works only in situations when the law (i.e., working principle) and the observed effect are known. Thus, according to Roozenburg, this type of abduction is not about "discovery" but about "diagnosis" and "troubleshooting" (ibid, p. 10).

Roozenburg (1993) terms the type of abduction he sees as central in design as ***innovative abduction*** (again, adopting the term introduced by Habermas (1978)). According to Roozenburg (1993), this type of abduction best represents the reasoning in Design and allows designers to conceive of new solutions only by being given a desired purpose. Innovative abduction is a reasoning mode starting from a surprising fact, which is yet to be explained (the result). This is followed by the conception of a new rule (a working principle, law, or theory) which allows for inferring the cause (the case). The rule itself, therefore, is not yet assumed to be true but it is seen as a new hypothesis that needs to be tested by deduction and induction (ibid). Roozenburg then goes on to designate innovative abduction as the kernel of design and consequently, the key mode of reasoning in design synthesis. Such is found throughout the entire design process, he claims, but it is present most prominently during the creation of what he terms a principal solution: "an idealized representation (a

scheme) of the structure of a system, that defines those characteristics of the system that are essential for its functioning." (p. 12).

The distinction between explanatory and innovative abduction presented above is not exhaustive of the history of abduction in the field of Design. For instance, what Roozenburg terms innovative abduction is given different names by different scholars. March (1984), for example, terms it as "productive reasoning[6]" (p. 267), Takeda et al. (2001) denote it as the third type of abduction (p. 3) and Dorst (2011) calls it abduction-2 (p. 524). Regardless of the different names they give it, the descriptions they arrive at are comparable to that of Roozenburg's (Kroll & Koskela, 2016). Therefore, in this chapter, we will not recount the different terms Design scholars use to denote abduction in further detail[7]. Instead, we proceed by identifying and elaborating upon two models[8] of innovative abduction used in Design[9] that can be seen as representatives of the two paradigms of Design Theory: Rational Problem Solving and Reflective Practice[10].

6 The reason March introduced the term of productive reasoning to denote abduction is as follows: "Peirce did not use the term productive: he used at different times the terms abductive, retroductive, presumptive, hypothetic. In the design context our choice of term seems more telling and natural" (March, 1984, p. 274).

7 The introduction and implications of abduction to the field of Design are already well-described by a plethora of authors among whom March (1984), Roozenburg (1993), Takeda (1994), Takeda et al. (2001), Dorst (e.g., 2011), Dong et al. (e.g., 2016) and Kroll and Koskela (e.g., 2014, 2015, 2016).

8 We refer only to these core papers of Roozenburg (1993) and Dorst (2011) because they are very explicit in their definition of abductive reasoning in the field of Design while maintaining a broader orientation (they adopted theories from e.g., March (1984), Habermas (2015) and Schön (1983)). The most important conclusions we draw here have been checked with other (later) papers by Roozenburg, Dorst and other authors (e.g., Takeda (1994, 2001); Dong and MacDonald (2016); Dong et al. (2016); Kroll and Koskela (2017)) and the work on Function-Behavior-Structure ontology of Gero (e.g., 2007).

9 It is worth noting that apart from the field of Design, abductive reasoning has also received attention in the fields of Philosophy of Science (most notably by Magnani (e.g., 2007, 2011) and in the field of Artificial Intelligence under the name of Inference to the Best Explanation (IBE) (however, IBE, although confused by many authors to be equal to Peircean abduction, is distinctly different from it – for an in-depth discussion on the topic, please, refer to Campos (2011)).

10 As stated in Section I, it is important to consider developments in Design pertaining to both paradigms of Design Theory since doing so could lead to the better understanding of design procedures, principles, and practices (Dorst & Dijkhuis, 1995; Dorst, 1997; Roozenburg & Dorst, 1998).

Abduction in the paradigm of Rational Problem Solving

The first model describing the manner in which abduction is applied to the process of designing a new solution originates from principles of the Rational Problem Solving paradigm (Roozenburg, 1993). According to him, the design process always starts with a **purpose** (what Simon also sometimes terms *a goal*). For instance, when designing a kettle, the purpose would be to "*be able to boil the poured-in water*". To fulfil its purpose, the kettle needs to *behave* in a certain way (e.g., *the bottom needs to heat up, so it can transform the heat to the water inside*). Roozenburg terms this **mode of action** (i.e., the solution's behaviour)**.** Introducing this term is a deliberate choice that allows him "to avoid the ambiguous term 'function'"[11] (Roozenburg, 1993, p. 12). Thus, according to him, the mode of action signifies the "(functional) behaviour of the artefact in response to influences exerted on it from its environment" (p. 12). The mode also serves as the bridge between the artefact and its immediate environment.

To account for the manner in which a human is to *use* the product (i.e., *putting the kettle on the burner*), Roozenburg introduces the notion of **actuation**. While the mode of action refers to how the artefact behaves in relation to the situation, actuation is the action that the user applies onto the artefact that allows it to function and be "connected" (p. 13) to its immediate environment. The outcome of the design process Roozenburg terms as **form**. (i.e., *the material and the geometrical shape of the product*). Therefore, the form of the kettle and the way it is used (actuated) causes it to behave in a certain way (mode of action), and therefore, by this behaviour, it can fulfil its purpose. The model of innovative abduction Roozenburg suggested is as follows: ***((form ^ actuation) → mode of action) → purpose***[12]

11 In the paradigm of Rational Problem Solving, the term "function" has been used in different ways by different scholars. For instance, in some instances, it can denote the purpose of a solution and in others – its behaviour (e.g., Rosenman & Gero, 1998). On the other hand, the term ***behaviour*** has many different meanings and connotations and can be misleading when discussing unintended consequences (see Merton, 1936).

12 In his final formulation, Roozenburg subsumes ***mode of action*** into the combination of ***actuation and form***. Therefore, he arrives at the following final formulation for innovative abduction: **form and actuation → purpose**. This action is not surprising considering the prevailing idea at the time that

Abduction in the paradigm of Reflective Practice

Another model of abduction in Design is rooted in the Reflective Practice paradigm. It was introduced by Dorst in 2011. He, as already explained, termed the core abductive reasoning in Design as "abduction-2". To explain the manner in which abduction-2 works, Dorst (ibid) presents the example of designers who are given a task to redesign a metropolitan entertainment quarter. The designers were asked to do so in such a way that their redesign can solve the habitual accidents of "drunkenness, petty theft, drug dealing, and, later in the night, sporadic violence" (p. 528) that was happening in the entertainment area. Presented with such a problem, the first thing a designer had to do was to define (an aspired) **value** the potential solution should achieve (e.g., "*people having good time*"). To achieve this value, the designers had to think about how the problem can be approached (e.g., "*this problem could be approached AS IF they were dealing with organizing a good-sized music festival*" (p. 529)). Dorst terms this action as a **frame**[13] and uses it to denote a cognitive act of looking at a problem situation from a specific viewpoint that informs how the problem can be solved (Dorst, 2015)[14.] Last but not least, in accordance with the Reflective Practice paradigm, Dorst also defines framing as the variable that allows the designer to gradually build her understanding of the problem and design a solution that approaches the problem in an original manner (ibid).

Once the frame was established, the designers then proceeded by defining a clear **how** the aspired result could be achieved. For example, they *provided an overview on the waiting times to enter each of the clubs in the quarter*. Then, once the value, the frame and the how were clear, the designers created the solution, which Dorst terms as **what** (i.e., a service, system, solution). The

---

the behaviour of an artefact can be fully defined by its purpose, structure (i.e., Roozenburg's form) and environment (i.e., Roozenburg's actuation) (see Simon, 1996).

13 According to Dorst (2011), the formulation of frames follows the format: "*IF we look at the problem situation from this viewpoint, and adopt the working principle associated with that position, THEN we will create the value we are striving for.*" (Dorst, 2011, p. 525).

14 As already discussed, framing is a practice that is central to the paradigm of Reflective Practice (Schön, 1983). However, since Schön's theory as presented in the "Reflective Practitioner" is "weak and fuzzy" (Roozenburg & Dorst, 1998, p. 40), Dorst introduced his own definition.

solution Dorst gives as an example is *an app that provides an overview of waiting times for each club*. Finally, according to Dorst, this sequence of actions can be represented as the following expression, where the frame facilitates the move from Value to How:

$$\underset{\text{FRAME}}{\textbf{WHAT + HOW} \rightarrow \textbf{VALUE}}$$

Overview of abduction defined by the paradigms of Design Theory

The two models clearly represent the two different paradigms of Design Theory. For Roozenburg, the purpose and the problem setting are clear and provided by the client and the design process is seen as a search. For Dorst, the subjective interpretation of a designer (and the situatedness of a design), housed in the variable of frame, are central. Despite these differences, however, one can clearly see the similarities between the two models (Table 2).

Table 2 An overview of the variables of Roozenburg and Dorst's abduction models

Rational Problem Solving (Roozenburg, 1993)	**Reflective Practice (Dorst, 2011)**	**Comparison**
purpose	value	Both serve as the desired outcome of the design process. However, the purpose is static while the value is constantly evolving.
mode of action actuation	how	In his final model, Roozenburg puts an emphasis on the actuation while to Dorst, the working mechanism is important.
form	what	Both represent the manner in which a solution can be manifested – either as a form, or as a what.

n/a	frame	The cognitive device Dorst uses to allow the designer to better represent and construct the problem setting.

First, both models commence with a pre-defined desired outcome: purpose for Roozenburg and value for Dorst. However, reflecting the paradigms to which they belong, the purpose remains stable throughout the design process (in accordance with Simon's early view on the process of design); while the value is continuously evolving to reflect the better understanding of the problem the designer gradually acquires (following Schön's prescription). Furthermore, Dorst's focus on value also represents a larger shift in the field of Design, which sees designing as a value-producing activity (e.g., Friedman, 1996; Calabretta & Kleinsmann, 2017)[15]. This is distinctly different from the previously dominant view on design as an activity of designing a product that fulfils a given purpose (as was the case for e.g., Roozenburg (1993)).

Second, both authors claim that the design process will eventually lead to a tangible outcome – either a *form* (Roozenburg) or a *solution/what* (Dorst). However, the underlying principles according to which the "form" or the "what" will be reached are different, reflecting the paradigm on which they are based. For instance, Roozenburg's model reflects the process of product design. Thus, to him the design process is a search for a satisficing option. Dorst's model, on the other hand, addresses the shift in Design from designing products to designing services and systems. Further still, fully in line with Schön's paradigm, Dorst's model introduces the variable of frame as the cognitive device that allows the designers to better represent and construct their ever-evolving understanding of the problem setting.

Third, both scholars introduce variables to signify the way a solution has to behave (i.e., Roozenburg's *mode of action* and Dorst's *how*). Roozenburg also introduces a variable to denote the manner in which a solution is to be used (i.e., *actuation*). As such, the formulaic expressions of abduction both scholars introduce provide us with clues on how to design the desired behaviour of a

15 See also Rosenman & Gero (1998); Cockton (2006); JafariNaimi et al. (2015); Dong et al. (2017).

solution. Further, the unabridged version of Roozenburg's model also provides a clear relationship between the desired behaviour of a solution (i.e., *mode of action*) and its use (i.e., *actuation*). As such, both the mode and actuation should be considered as stand-alone variables due to the fact that the AI system's future behaviour is highly dependent on the way it is going to be used, as discussed in the previous chapters. This contention is supported by Dong et al. (2016b). According to them, both mode and actuation have to be explicitly considered when designing digital artefacts since the actuation of a solution is directly dependent on its mode of action. These provide us with invaluable insights on how new solutions can be designed by taking into consideration both the behaviour of a solution and its use. Thus, bringing us one step closer to providing an answer on how Design theories could enable us to simulate the behaviour-use interdependence.

Conclusion

This chapter started by stating the need to answer the following research question:

"How do Design theories address the behaviour and use of solutions?"

The exploration was carried out in two stages. First, in Section I, we provided an overview on the two main paradigms that have been defining for the field of Design Theory. Then, in Section II, we delved deeper into the core of design and discussed the reasoning that provides models for the relationship between behaviour and use and how these two variables can be used to design new solutions. An overview on the insights that emerged from both parts can be found below.

As stated in Section I, if we are to devise a theory representative of the design process, we need to consider the paradigms of Rational Problem Solving and Reflective Practice (Dorst, 1997; Roozenburg & Dorst, 1998). From the overview of Simon (1996) and Schön (1983), three important insights emerged: (1) establishing a feedback mechanism is important so that we can have a better understanding of the problem; (2) the designer has influence over the design

process (only a limited one for Simon and an instrumental one for Schön); and (3) unintended consequences (i.e., surprises for Schön) are a natural and needed outcome of the design process as they are the ones that propel the design process forward by, for instance, causing a reframe/iteration.

Section II introduced abduction in its capacity as the key reasoning mode of design, followed by the models introduced by Roozenburg (1993) and Dorst (2011) (each one related to one of the paradigms in Section I). Despite the differences the two abduction models have, the reasoning mode they represent is comparable (Kroll & Koskela, 2016). They both start from an initially agreed-upon point (*purpose* and *value*), then they use that to define either implicitly or explicitly the behaviour and use of the potential solution (*mode* and *actuation*/*how*). The combination of these consequently leads to a tangible solution – either a *form* or an *object, service, or a system*. Therefore, both models provide an important piece for our understanding on the relationship between behaviour and use and how they both contribute to the design of new solutions. Thus, it provides new building blocks that can support our attempt to answer the main research question of this dissertation: ***"How can a theoretical model be designed that supports the early simulation of AI systems' behaviour-use interdependence by utilising Design theories?".***

CHAPTER 4

PROTOTYPING FOR EARLY SIMULATION OF BEHAVIOUR AND USE

In the previous chapter, we presented an overview on Design theories that can provide us with insights on how to address the behaviour-use interdependence of a solution. As such, we elaborated upon the two main paradigms defining the field and the models that formalise the relationship between behaviour and use (i.e., mode of action and actuation). In this chapter, we address the first part of the research question (in bold): *"How can a theoretical model be designed that* ***supports the early simulation*** *of AI systems' behaviour-use interdependence by utilising Design theories?"***.** This exploration is guided by the following question:

"***What Design theories support the early simulation of a solution's behaviour and use?".***

In order to provide an answer to the question, the chapter first introduces an overview on the characteristics and benefits of a design practice that has been widely used as a means to simulate potential futures – prototyping. This overview is then complemented with two examples from my design practice. Each of them showcases the manners in which prototypes can be used to

simulate the behaviour and use of new solutions. The chapter is concluded with a discussion on the relationship between prototyping and abductive reasoning.

The design practice that supports early simulation

Prototyping is a widely acknowledged practice (across all sub fields of Design) as an important means for early simulation of potential solutions. As such, it enables designers to explore and communicate what it will be like to interact with future products, systems, and services (Buxton, 2007; Lim, et al., 2008; Stappers, 2010). According to Lim et al. (2008, p. 8), prototypes are a "tangible attempt to view a design's future impact so that we can predict and evaluate certain effects before we unleash it on the world". They generate deep level of understanding about novel contexts (ibid) and have shown promise to identify uncertainty or the so-called unknown unknowns early in the product development process (Jensen et al., 2017). For instance, according to de Reuver et al. (2020), prototyping can also serve as a "mechanism to uncover value implications of a novel platform in an early stage, in a controlled environment" (p. 6).

Many definitions of a *prototype* exist, depending on the field they originate from and the purpose they serve, complemented by a plethora of methods and tools[1]. The term can be used for artefacts ranging from simple visual mock-ups or sketches that represent design thinking and doing (Ullman et al., 1990; Suwa & Tversky, 1997) through experience prototypes such as the "Wizard of Oz" (Buchenau & Suri, 2000), conscription devices (Henderson, 1991), minimal viable products (MVP)[2] (Ries, 2011), provotypes (Mogensen, 1992), prototrialing (Jensen, 2017), and boundary object (Star & Griesemer, 1989) to highly functional sophisticated pre-production prototypes that are seen as "representations of a design made before the final artifacts exist" (Buchenau

1 See Kleinsmann and Ten Bhömer (2020) for a comprehensive overview.
2 MVP is an initial product iteration that contains essential features which are sufficient for initial customers to use and offer feedback on. It serves as a way to inform the development of future iterations of the product.

& Suri, 2000, p. 424) and can serve as "aids for imagination" (Kurvinen et al., 2008, p. 47).

Reflecting the multiplicity of terms, the prototyping process has been known to provide multiple diverse advantages. For instance, it can stimulate experimenting and reflection-in-action by framing, and discovering possibilities in a design space (e.g., Schön, 1983; Cockton, 2006). Further, it can serve as a robust feedback mechanism[3] (Buxton, 2007; Lim, et al., 2008), improve a design team's understanding about the design itself (Vetterli et al., 2012) and in some cases, the creativity levels and divergence in ideation (Gerber, 2009). Prototypes can be used to explore the socio-material relationships and issues of the context (e.g., Gill et al., 2011; Elverum & Welo, 2016; Tironi, 2018), inform decision-making on desirability, viability, and feasibility (Menold et al., 2017) and serve as tools for communication and involvement of both internal and external to the organisation stakeholders (Jensen et al., 2017). They also evoke a focused discussion in teams, confront theories and allow users to experience their world differently (Sanders & Stappers, 2014). Prototypes can also be used as tangible rapid learning cycles (Leifer & Steinert, 2011; Haines-Gadd et al., 2015) that enable direct access to challenges and potential solutions (Kurvinen et al., 2008).

Across this multiplicity of purposes, a clear pattern can be discerned as to using prototypes for their ability to support both discovery and evaluation (e.g., Lim et al., 2008). First, prototyping, when used during discovery, supports the incorporation of the situation's back talk (Schön, 1983; Lim et al., 2008). Doing so allows for flexibility and quick adaptation to the unique design situation. It also supports designers to learn, discover, generate, and refine their designs (Buxton, 2007) by stimulating framing, and discovering possibilities in a design space (Lim et al., 2008). On the other hand, prototyping can also be seen as a means to evaluate design's failure or success (Lim et al., 2008) and can

3 It is important to note that when Schön refers to experiments, he discusses the activity of sketching. This view of sketching being able to support the back talk of a situation is shared by other scholars as well (e.g., Goldschmidt, 2003). However, sketching can also be seen as a type of prototyping (e.g. (Ullman et al., 1990; Suwa & Tversky, 1997; Jensen et al.., 2017). Therefore, in this dissertation, we discuss the wider practice of prototyping instead of focusing only on sketching.

be utilised once the "design opportunity has been established" (Sanders & Stappers, 2014, p. 10). Therefore, prototypes evaluate whether the manner in which the solution's version behaves, it is being used and the outcomes it produces are similar to the ones intended by the designer (e.g., Otto & Wood, 2001; Ulrich and Eppinger, 2012).

Given this brief overview on the characteristics and benefits of prototypes, one can contend that they are conductive to the simulation of the solution's behaviour and use. In order to further support this contention, we provide two examples of prototypes that enabled a multidisciplinary team to simulate the behaviour and use of their solutions. The examples come from a project I carried out between November 2015 and September 2016 for a large European airline. The project's aim was to support a multidisciplinary team I was part of in their efforts to create and implement a new design-led innovation process. The work around the innovation process has been discussed in previous publications (i.e., Stoimenova et al. (2016) and Stoimenova & de Lille (2017)). The process and the project itself are outside the scope of this dissertation.

Using prototypes to simulate behaviour and use

This section presents an overview on two of the devised prototypes during the project. These were chosen since they clearly exemplify the manners in which prototypes facilitate the early simulation of behaviour and use. They also provide insights into how the outcomes (both intended and unintended) can be used to further detail the behaviour of a solution. Therefore, we discuss each one of them around the notions of *behaviour*, *use* and *prototypes* to reflect our contention that prototypes can be used to simulate both the behaviour and use of new solutions. We also discuss the *purpose* of the prototype to provide context. Finally, we discuss both the *intended* and *unintended outcomes* their use triggered, since the elicited outcomes serve as a trigger for reflection-in-action or a new search process (following Schön (1983) and Simon (1996)).

Example 1: prototype user involvement

Table 3 An overview on Example 1, along the lines of its purpose, prototypes, behaviour, use and outcomes.

What	Description
Purpose	Help the team understand the implicit regulations in place and gauge to what extent they can involve passengers waiting to board their plane in the innovation process.
Prototypes	Templates for ideation methods like Lego Serious Play, Brainwriting and Context mapping (Visser et al., 2005), signs, and a configuration for the setup of the area (i.e., put a table and a sign near three of the gates at the local airport and brought along paper, scissors, Lego bricks, markers, sticky notes, and pre-cut photos).
Behaviour	Asking passengers to follow the steps of each method while trying to generate novel ideas on how to improve the time they spend waiting at the airport.
Use	Passengers went through the predefined steps and discussed with the team their ideas on how to improve their journey. They noticed the signs, but the team still needed to go and ask every passenger around whether they would like to join.
Intended outcomes	A set-up with A2 signs established the team's credibility; They elicited the implicit regulations at the gates on what materials can be used and how the area can be set up; Twenty passengers joined the team.
Unintended outcomes	The majority of the generated ideas were similar (e.g., "more comfortable chairs" and "more power outlets"); The used methods could not elicit any tacit knowledge about the passengers.



Image 3: A photo of a passenger session carried out near one of the local airport's gates.

The first prototype (Table 3 and Image 3) was used to gauge the manners in which the team can involve passengers in the innovation process. To do so, they prototyped templates for existing methods by adapting them to their context, as well as different test setups near three of the gates at the local airport (e.g., a sign, a table, a Lego set, sticky notes, markers, sheets of paper and pre-cut photos). Doing so allowed the team to understand the implicit regulations at the airport when it comes to setting up a testing area. They also managed to gauge to what extent they can involve passengers waiting to board their plane in the innovation process (e.g., although 20 passengers joined us, the ideas they generated were very similar (e.g., "more comfortable chairs" and "more power outlets")).

The elicited outcomes served as a trigger for reflection-in-action. For instance, an important insight was the fact that novel ideas were difficult to elicit. This led to the team's resolve to involve passengers only as data input (e.g., use the context mapping method to ask passengers to map their own journeys). This way they would have enough time to ask more detailed questions about the passenger's experience. The insight was later applied to other prototypes dealing with the topic of passenger involvement. For instance, in subsequent prototypes, the team made a specific template for context mapping[4] and arrived at a selection of photos that works well with international passengers and at the airport. Furthermore, they established general guidelines on how such context mapping can be carried out at the gates. Finally, they also developed a simple guideline on the types of questions that can be asked during such passenger involvement.

4 Context mapping is a method which involves creating a visual representation of the context a designer is trying to understand and analyse. It can also be used in co-design activities where users and stakeholders map out the context together with the designer (Visser et al., 2005)

Example 2: test set-ups in live environments

Table 4 An overview on Example 2, along the lines of its purpose, prototypes, behaviour, use and outcomes.

What	Description
Purpose	Understand how to set up MVP tests in a live environment with real employees and passengers without disrupting the operations of Y, the airport, or other airlines.
Proto-types	A step-by-step set up on how to test an MVP of a boarding procedure, paired with initial guidelines for the team on how to conduct short inter-views.
Behaviour	Set up and tested the MVP during the boarding procedure of an existing flight, created signs on how to board and where the passengers need to queue. The team conducted multiple short interviews with employees and passengers. The MVP had to be quickly changed during the test.
Use	Passengers and employees were confused whether the boarding call is real. They queued in the wrong lane.
Intended outcomes	All passengers boarded their flight on time.
Unin-tended outcomes	The employees were confused but had to reassure passengers to follow instructions; The team did not follow the previously agreed-upon inter-view guide; The low-level fidelity MVP was confusing for the passengers and employees; The initial MVP had to be changed in the middle of the test to alleviate the created confusion; The pre-made templates for the MVP were difficult to update on the spot.

The second example represents a prototype the team used to understand how to set up MVP tests with real passengers and employees without disrupting the operations of the airline, other airlines, or the airport (Table 4 and Image 4). To gauge how this can be achieved, they started with an MVP for a new boarding procedure (i.e., set up different boarding lanes and reminders for passengers so that they can board the plane faster). Typically, the boarding of passengers involves many rigid carefully orchestrated protocols and when testing new ones, such regulations have to be considered. Hence, the team used this MVP as a starting point for the design of a new testing process and setup the team could use for future MVPs.

The team also wanted to check whether the MVP format and fidelity they have chosen is a good fit for the environment and whether the initial interview guidelines allow them to collect authentic reactions from passengers. Based

on the observed use, they concluded that (1) MVP tests could not be carried out with little guidance; (2) low-fidelity MVPs do not work in their environment since both passengers and employees were confused by the instructions; and (3) the team still had trouble executing ad-hoc brief interviews with passengers and employees.

Image 4 An MVP test of a new boarding procedure



The elicited outcomes were addressed in subsequent prototypes. For instance, the team further developed the interview guideline by providing examples of specific situations and how to formulate good situation-specific questions. They also provided general guidelines on what constitutes a good and a bad question. Second, due to the confusion the low-level fidelity MVP caused, in each subsequent prototype the team made use only of mid-level fidelity MVPs that are easy to update on the spot without the knowledge of specific software programmes. Third, because of the ad-hoc actions of the team and the confusion of both passengers and employees, the clear need emerged for a well-defined testing protocol (e.g., employees that will be involved in the

test need to understand why they are carrying out tests in such a way). Last but not least, in each subsequent prototype, the team always used a clear designation of a test area. Doing so made the passengers aware that there might be an ongoing test and reassure them that they will manage to board their flight on time.

Discussion

The aforementioned examples clearly showcase how prototyping can be used to simulate the behaviour of a potential solution in order to elicit different uses and outcomes. As such, prototypes played the role of a *bridge* between the two variables and supported the team to deepen their understanding of the context for which they had to design, triggering reflection-in-action. Further still, we observed how prototyping enabled the team to swiftly change the solution's behaviour in order to respond to the elicited uses and outcomes. In particular, unintended outcomes were the ones that impacted the changes of behaviour the most and moved the design process forward. It was these that were usually addressed in subsequent prototypes. They also supported the team in gradually building up a solution that fits with the context and generates intended outcomes. For instance, in Example 2, when the MVP was not clear and passengers were getting confused, the team needed to adjust their behaviour (i.e., update the MVP on the spot) to alleviate passengers and employees' confusion. The new behaviour was then housed in the updated MVP. This observation is in line with the paradigms of both Rational Problem Solving and Reflective practice, where it is the unintended outcomes that propel the design process forward (Schön, 1983; Simon, 1996).

Utilising prototypes in their capacity to simulate behaviour and use can be seen in the practice of other designers, too. An example of such comes from the work of ten Bhömer, who designed a shirt with sensors that can support the physical rehabilitation of the elderly (see e.g., ten Bhömer et al. (2013, p. 37)). The designer created a highly interactive prototype with sensors to be placed on specific body parts, capable of measuring arm and lower back movements, and recovery progress, supported by sound feedback. By intentionally building a functional prototype and defining its behaviour, ten Bhömer was able to

observe how his stakeholders such as therapists, caretakers, and patients used the prototype. According to him, being able to observe the manners in which the prototype was used allowed him to identify the requirements the concept should fulfil, the scenarios the shirt could be worn in, and develop a list of steps for further improvement (e.g., increasing the sensitivity of the used sensors (p. 38)).

Prototypes and abduction

The presented examples clearly show how the use of prototypes can be instrumental in the generation of new (design) hypotheses about the complex context for which one has to design. This ability to generate hypotheses, as we discussed in Chapter 3, is one of the hallmarks of abductive reasoning. Yet, neither the model of Roozenburg (1993), nor that of Dorst (2011) explicitly discusses the role prototypes play. Further still, to our knowledge, no Design scholar has discussed prototypes as potential enablers of abductive reasoning. Still, what we saw in the examples above is indicative of abduction. The use of prototypes triggered the generation of hypotheses that would possibly not have been generated were it not for these prototypes. Furthermore, these prototypes also enabled the team to directly address the outcomes they observed.

To find an explanation on how a prototype can trigger abduction, we turn to the field of Philosophy of Science and in particular to the work of Magnani (e.g., 2007). Both Magnani and Nersessian (e.g., 2002) have written extensively about a type of reasoning they call *model based*. This term is used to denote the type of thinking that happens when various types of representations (i.e., both internal and external models) are constructed and manipulated to support one's thinking process (Magnani, 2007). For instance, in order to solve a geometrical problem, one often needs to draw or at least imagine the geometrical shape. Within this broader topic of model-based reasoning resides the type of abduction Magnani terms *manipulative*. This is a type of productive reasoning that supports "thinking through doing" (Magnani, 2007, p. 7). Thus, it allows us to elicit new and still unexpressed information codified within the context. One can do so by building various *external* epistemic mediators that function as a

new source of information and knowledge[5]. In fact, according to Magnani (ibid), this type of abduction is core to scientific discovery and as such, it is aimed at "creating communicable accounts of new experiences to integrate them into previously existing systems of experimental and linguistic (theoretical) practices" (p. 4).

The prototypes discussed above and the manner in which the team used them fit Magnani's description of manipulative abduction well. They served the role of external epistemic mediators that enabled the team to uncover knowledge that was implicitly present in their context. Therefore, when dealing with complex systems, it is our contention that using prototypes in their capacity to generate hypotheses becomes an invaluable part of the design process. It also gives us a mechanism to elevate the procedural outcomes of prototyping to the level of abstraction of the models Roozenburg and Dorst introduced (i.e., both are representative of abductive reasoning). Given all these insights, we arrive at the following working definition of an AI system prototype: *"an externalised representation of a potential AI system"*[6]such as sketches, mock-ups, pieces of software and in some cases, hardware.

Conclusion

The research question guiding this chapter was: ***What Design theories support the early simulation of a solution's behaviour and use?".*** In order to answer it, we presented an overview on the characteristics and benefits of a Design theory that is well-known for its ability to support simulation of future solution states – prototyping. We then postulated that prototypes are also conductive to the simulation of the solution's behaviour and use. In order to support this statement, we re-interpreted previously carried out empirical research by showcasing two examples of prototypes that were used to design and implement a new innovation process into a complex context.

5 For a comprehensive overview on the different types of abduction and the specifics of model-based reasoning, please, refer to the work of Nersessian (e.g., 2002) and Magnani (e.g., 2007).
6 This definition combines my own work in for a large European airline (as discussed above) and the work of Schön and Wiggins (1992), Suwa et al. (2000) and Stappers and Giaccardi (2017).

From the presented examples, three additional insights emerged that help us understand how prototypes can support the continuous simulation of the behaviour and use of a solution. First, the prototypes discussed in the two examples allowed the team to externalise the intended behaviour into the context. Thus, they supported the team in observing the different types of outcomes and uses the behaviour, housed in a prototype, can elicit. Second, prototypes served as a bridge between behaviour and use. As such, they were also conductive and instrumental to the generation of hypotheses that allowed the team to further develop their solution and address the unintended outcomes they uncovered. And thus, third, prototypes were conductive to what Magnani terms manipulative abduction. In fact, their abductive capabilities also provide us with a solid foundation on which we can formulaically build the relationship between behaviour and use in complex contexts. Therefore, prototyping marks the third piece of the puzzle of Design theories we can use to answer the main research question of this dissertation: ***"How can a theoretical model be designed that supports the early simulation of AI systems' behaviour-use interdependence by utilising Design theories?".*** The next chapter provides an initial answer to it.

CHAPTER 5

INITIAL THEORETICAL MODEL

The main premise of the dissertation is as follows: if we are to ensure that AI systems are non-maleficent, we need to address the behaviour-use interdependence that defines these contemporary systems. Therefore, we claimed, it becomes imperative to develop theoretical models and methods that support us in simulating potential uses as early as possible. Doing so would allow us to fine-tune the behaviour of the system to trigger the desired use. We then postulated that Design theories could help us address this challenge, resulting in the main research question of the dissertation:

"How can a theoretical model be designed that supports the early simulation of AI systems' behaviour-use interdependence by utilising Design theories?"

Then, in Chapters 3 and 4, we explored Design theories that could help us understand the manners in which behaviour and use of a solution are connected (Chapter 3) and how their early simulation can be executed (Chapter 4). This exploration resulted in the following insights. First, Design theories have largely been defined by one of the two main paradigms in Design – Rational Problem Solving (Simon, 1996) or Reflective Practice (Schön, 1983). Despite their differences, they share some similarities: (1) establishing a feedback

mechanism is instrumental to achieving a better understanding of the problem; (2) the designer has influence over the design process; and (3) unintended outcomes are a natural and needed stage of the design process since they propel it forward.

Second, the Design theory that provides us with formally defined relationships between the variables of behaviour and use is that of innovative abduction. This type of abduction allows us to not only identify the relationship between the variables of behaviour and use, but also, introduces the manners in which the two can be used to design new solutions. There are two models in Design that define the relationship – one introduced by Roozenburg (1993) and one by Dorst (2011). Despite the differences the two abduction models have, they both start from (1) an initially agreed-upon starting point (*purpose* and *value*), (2) which they use to define either implicitly or explicitly the behaviour and use of the potential solution (*mode* and *actuation*/*how*), and (3) the combination of these leads to a solution – either a *form* or an *object, service, or a system*.

Third, prototypes can support the continuous simulation of the behaviour and use of a solution that is to be implemented into a complex context. In fact, they (1) support us in observing the different types of outcomes and uses the behaviour can uncover; (2) serve as a bridge between behaviour and use; and (3) enable what Magnani terms manipulative abduction.

In this final for Part II chapter, we position these theories in the context of designing AI systems to provide an initial answer to the main research question. As such, the chapter is structured as follows. First, we introduce an initial theoretical model built around the identified variables thus far by using a fictional example. Through the fictional example we also introduce the order in which the model is to be applied and the domain to which it should be applied. We then discuss the manner in which the model supports an iterative design process. The chapter is concluded with a short note on the role values play in the introduced theoretical model.

Theoretical model

To explain how we could design a theoretical model that supports the simulation of a behaviour-use interdependence, we use a fictional example of a developer[1] who wants to devise an AI system that reduces the burden people with chronic kidney problems experience[2]. A potential starting point might be the insight that currently, kidney patients undergo long periods of dialysis, without an overview of their daily process and not knowing whether they are experiencing life-threatening symptoms. Given this knowledge, she (or the client) might decide that these patients should feel informed and in control (e.g., *"people with chronic kidney conditions are in control of their health").* This can become our **purpose**[3]**.**

Next, the developer needs to understand the problem better. She can do so by collecting **data** (both qualitative and quantitative). While data has always played an important role in the design process (e.g., when collecting user insights), the development of contemporary AI systems (especially the ones relying on ML) is heavily dependent on large quantities of data (i.e., big data) (e.g., Russell & Norvig, 2021). To define the variable, we adopt the definition of Woodward (2011), formalised by Horvath (2016): "*records produced by experiments and measurements that serve as evidence for the existence or features of a*

1 Developer in this context means anyone who is involved in the design and development of AI systems.

2 Although this is a fictional example, to provide the necessary level of detail, we use the case of the company AliveCor. The start-up produces cell-phone cases and AppleWatch wristbands that can perform electrocardiograms (ECGs) (Topol, 2019). Based on them, potassium levels in "near real time" can be detected without drawing blood (Dillon & Friedman, 2018). This case was deliberately chosen as it addresses the complex context of healthcare in which multiple stakeholders (e.g., patient, nephrologist, GP, hospitals, hospital staff, medical device systems manufacturers, insurance companies) come into play. They all expect the solution to deliver values tailored to them, while the AI system continuously learns from its users and consequently could exhibit novel behaviours. As such, this case provides a wide range of challenges and exemplifies the type of contexts and domains for which the theoretical model is developed.

3 As already explained in Chapter 2, according to e.g., Roozenburg (1993) and Simon (1996), the design process always starts with a clearly defined purpose (i.e., the goal an artefact is designed for). The design process of developing AI systems also starts with a clearly defined purpose a development team could work towards. Therefore, both the variable "purpose" and its definition provided by Roozenburg (1993) continues to be suitable for the context of designing AI systems.

phenomenon" (p. 214). In the early stages of the traditional design process, these data tend to be primarily qualitative so that the problem space can be well-understood and defined. However, quantitative data from e.g., sensors can be utilised as well to provide another level of insights.

Going back to our example, the data collection could go as follows: the developer would carefully study her patient's context through regular interviews and observations, and then map their day-to-day journey. These provide contextual understanding of the problem. They also help her discover, for instance, that physicians are oftentimes unsure about the precise dosage of medications since each one of them can directly influence the blood potassium levels of the patient. Higher levels than recommended can be fatal and require immediate treatment. Yet, the only way to identify blood potassium levels is for the patient to undergo an invasive test performed in a laboratory. This produces high levels of uncertainty and stress for the patient and all stakeholders involved.

The combination of **purpose** and **data** leads us to a vantage point from which she can approach the problem at hand. Namely, "*if potassium levels are detected regularly, changes in the dosage of a medication can be easily administered in the comfort of the user's home (no need for blood tests)*"[4]. This vantage point can be likened to the variable of **frame**[5] (Schön, 1983; Dorst, 2011) discussed in Chapter 3. Further, as it can be seen from the formulation our fictional developer used, the frame already suggests the intended outcome: *administer a new medication in the comfort of the user's home*. If we are to represent formulaically (akin to Roozenburg and Dorst) this progression from purpose and data to a frame, we can adopt the following expression:

4 This formulation of the frame follows closely the manner in which Dorst (2011) suggests a frame ought to be formulated: "*IF we look at the problem situation from this viewpoint, and adopt the working principle associated with that position, THEN we will create the value we are striving for.*" (p. 525).

5 According to Dorst (2011), frames are a cognitive act of looking at a problem situation from a specific viewpoint that informs how the problem can be solved.

purpose + data → frame[6]

The identified frame provides us with an indication on how the problem can be solved i.e., *detect potassium levels regularly*. There are multiple potential directions in which one can go from here. For instance, during her data collection, our developer might find out that deep neural networks can detect blood potassium levels in an electrocardiogram (ECG) (Topol, 2019). Therefore, a plausible way in which the potential solution could behave is: "a *deep neural network detects potassium levels in ECGs and informs the patient and their physician about them when needed"*. As already discussed in Chapters 2 and 3, the behaviour of an AI system can be equated to the variable of **mode of action**[7] (Roozenburg, 1993).

Going back to the example, the combination of the frame and mode of action can be manifested as a simple piece of hardware[8] (e.g., a wristband) equipped with electrodes that can measure its user's pulse and hence provide us with the data to produce an ECG. We need a variable that can represent both the pulse-measuring strip and reflect the transient nature of AI systems. A design concept that is well-suited for this role is that of a **prototype**. Viewing each of the states of the AI system as a prototype will promote the notion of designing for something transient (i.e., a solution that can always change the outcomes it delivers). In effect, this will enable developers to learn continuously as the solution evolves. Moreover, as discussed in the previous chapter, the act of prototyping is what supports the early simulation of a new solution's behaviour and use. The definition of prototype we adopt, as discussed in Chapter 4, is *an externalised representation of a potential AI system* such as sketches, mock-ups, pieces of software and in some cases, hardware.

6 This expression can also be seen as representative of explanatory abduction (e.g., see Dong and MacDonald (2016)). The relationship between explanatory and innovative abduction is discussed in Chapter 6.

7 "(Functional) behaviour of an artefact" (Roonzeburg, 1993, p. 12).

8 Although the example used here is of a "piece of hardware", oftentimes the AI model is embedded into existing experiences, processes, or software.

This process of combining the frame and mode of action in order to develop a prototype can be represented formulaically as:

frame + mode of action → prototype

Since the goal of this prototype is to better understand the context and generate hypotheses about it (see Chapter 4), a good starting point would be to provide the users with the wristband and without much guidance to observe the way they use it over the course of a month. A variable that can denote the "*the action a user applies*" (Roozenburg, 1993) onto a prototype is **actuation**.

Observing the manner in which our prototype is used will generate multiple insights on the context, time of day, and expectations users have while wearing the wristband, as well as on the way it impacts their daily routine and those of the stakeholders. Therefore, it is through this first prototype that the developer understands the ways in which her AI system can be actuated and ultimately observe the different outcomes it can create. In effect, the prototype serves as a bridge between the mode of action and actuation (as discussed in Chapter 4). If we are to put the act of utilising actuation to elicit the potential outcomes of an AI system in a formulaic expression, we arrive to:

prototype + actuation → outcomes

These outcomes can be intended: e.g., our user feels like she has better control over her daily routine. They can also be unintended (both positive and negative) and can stem from the fact that the wristband creates new dynamics in our user's life. For example, the prototype can introduce a lot of uncertainty and tension between the patient and her partner as they begin to obsess over insignificant changes registered by the wristband. On the other hand, the unintended outcomes can also be positive. For instance, the prototype can help the user to keep better track of how her health data is being used and shared. Such unintended outcomes, in theory, could also support us in defining

a different purpose. Thus, simulating the behaviour-use interdependence of AI systems could go through numerous iterations through this theoretical model:

purpose + data → frame

frame + mode of action → prototype

prototype + actuation → outcomes

The elicited outcomes (especially the unintended ones) trigger another iteration through the model[9]. The developer starts collecting new data (both qualitative and quantitative) on how the patient's daily routine changed, interviews with the patient and her spouse, as well as with the involved stakeholders. But also, reviewing the ECG and potassium levels and map them to events of what was observed and communicated in the previous iteration. Such additional data collection will help in refining our frame and add another dimension to the mode of action by adding behaviours that could address the identified tension. These are reflected in a new, more detailed, prototype. Consequently, new intended and unintended outcomes are uncovered and addressed in later iterations. As such, also the prototypes she designs could become more detailed or help her to understand a different part of the solution such as the way to deliver information to nephrologist, GPs, and patients. Therefore, the impact these delivered outcomes have not only on the patient, but also on the community and the other stakeholders (i.e., the GP, nurses, family) can be simulated.

The continuous simulation of both the mode and actuation of a potential AI system the theoretical model is conductive to, allows us to gain knowledge about their interdependence and continuously adapt them to respond to the unintended outcomes the system creates. As such, this theoretical model outlines the steps that can be used to simulate the interdependence and thus, gradually build AI systems that can be implemented into complex contexts. Doing so will also ensure that the developer can achieve a deeper level of understanding about the interaction between the unintended outcomes and the

9 The identification of unintended outcomes is especially important since they are the ones that propel the design process forward, according to both Simon (1996) and Schön (1983) (see Chapter 3).

context they operate in. An overview on the variables of the model as well as on their definitions can be found in Table 4.

Table 4 A comparison between reviewed and adapted definitions to each of the variables of innovative abduction. The changes are marked with bold.

Variable	Reviewed definitions	Adapted definitions
purpose	The goal of an artefact, e.g., "boil water" (Roozenburg, 1993, p. 12)	The goal of an **AI system**
data	Not addressed by Roozenburg or Dorst	Records produced by experiments and measurements that serve as evidence for the existence or features of a phenomenon (Woodward (2011), formalized by Horvath (2016)).
frame	A cognitive act of looking at a problem situation from a specific viewpoint that informs how the problem can be solved (Dorst, 2011)	Same
mode of action	"(Functional) behaviour of the artefact in response to influences exerted on it from its environment" (Roozenburg, p. 12)	(Functional) behaviour of **the AI system** in response to influences exerted on it from its environment.
prototype	"A tangible attempt to view a design's future impact so that we can predict and evaluate certain effects before we unleash it on the world" (Lim et al., 2008, p. 8)	**An externalised representation of a potential AI system.**
actuation	The action that the user applies onto the artefact that allows it to function and be "connected" (Roozenburg, 1993, p. 13) to its immediate environment.	The action that the user applies onto **the AI system** that allows it to function and be "connected" to its immediate environment.
outcomes	Not mentioned	The intended and unintended results of the AI system's actuation

A note on values and prototypes

As it can be seen from the introduction of the initial theoretical model, its design was informed by the variables and the relationship among these identified in the previous two chapters. Amid this construction, we also introduced two modifications to these Design theories: (1) our treatment of

values and (2) the cognitive role we assigned to prototypes. In the remainder of this section, we discuss each one of them.

Values

We made the decision to term the results of the abductive process as *outcomes* instead of *values* (unlike Dorst (2011)). There are two reasons for that. First, to Dorst (ibid) values are the "*outcomes one wishes to create*". However, as it can be seen from the introduction of the model, and from the examples of prototypes given in Chapter 4, outcomes can be both intended and unintended. In fact, as discussed in Chapter 3 of the dissertation, it is the unintended outcomes that propel the continuous exploration of the behaviour-use interdependence. Therefore, choosing the term "value" as Dorst terms it omits an important part of the outcomes the actuation of a prototype creates – the unintended ones.

Second, in the context of developing AI systems, the term *value* has come to be synonymous with ethical values (e.g., non-maleficence, fairness, equality) (see e.g., Santoni de Sio & van den Hoven, 2018; Mittelstadt, 2019; Jobin et al., 2019; Floridi & Cowls, 2019; Rahwan et al., 2019). In fact, the majority of introduced guidelines on how AI systems ought to be developed discuss the importance of embedding the right ethical values into the solution (see Chapter 2). Further, multiple prominent approaches exist to ensure that AI system's behaviour is aligned with human values such as Value Sensitive Design (e.g., Friedman & Kahn 2003), Values in Design (Nissenbaum, 2001), Design for Values (van den Hoven et al., 2015), and Worth-Centred Design (Cockton, 2009). Therefore, the use of term "values" could bring unnecessary confusion.

Prototypes

As already explained, we made the decision to substitute the terms "form" (Roozenburg, 1993) "WHAT" (Dorst, 2011) with "prototype". Prototyping is widely acknowledged across Design's subfields as a crucial means for early simulation of potential solutions (see Chapter 4). It empowers designers to envision and communicate future product, system, and service interactions (Buxton, 2007; Lim et al., 2008; Stappers, 2010). It also plays a pivotal role in

generating new design hypotheses within complex contexts (i.e., facilitating manipulative abduction). Yet, neither the model of Roozenburg (1993), nor that of Dorst (2011) explicitly discusses the role prototypes play. Further still, to our knowledge, no Design scholar has discussed prototypes as potential enablers of abductive reasoning.

Therefore, our decision to formally define prototypes' role as the bridge and facilitator of the relationship between mode of action (i.e., behaviour) and actuation (i.e., use) allows us to provide a mechanism to formally elevate the procedural outcomes of prototyping to the level of abstraction of abductive reasoning. As such, it provides us with a frame of reference to position the cognitive significance of prototyping in the process of reasoning about a new design solution. Hence, this new formulation can serve as a starting point for further research that can expand our understanding of (cognitive) design processes.

Conclusion

This chapter presented a theoretical model that can aid us to simulate the behaviour-use interdependence of AI systems. Doing so, we posited, allows us to continuously adapt the system's behaviours (i.e., mode of action) so that they will trigger the desired use (i.e., actuation). Hence, providing a preliminary answer to the main research question of the dissertation: *"How can a theoretical model be designed that supports the early simulation of AI systems' behaviour-use interdependence by utilising Design theories?"*. We presented the model formulaically, following the format used by Roozenburg (1993) and Dorst (2011).

In order to exemplify the relationships among the model's variables and the manner in which the model could be used, we employed the fictional case of reducing the burden people with chronic kidney problems feel. As such, the case allowed us to also introduce the updated definitions of each variable, the order in which each of them is to be addressed, and the domain to which the model should be applied. We then discussed the manner in which the model supports an iterative design process by continuously addressing the unintended

outcomes the actuation of the devised prototypes elicits. However, as already explained, this chapter presented only an initial version of the theoretical model. Hence, the described relationships among the identified variables are only conceptual. In the next part of the dissertation, we will further explore and define these relationships.

PART III

EXTENDED THEORETICAL MODEL



CHAPTER 8

Discussion on the model's version which incorporates all of the insights this doctoral research produced.

CHAPTER 7

A case study exploring how Design theories should be extended to address the AI systems' behaviour-use interdependence.

CHAPTER 6

Initial exploration on whether the devised model could address AI systems' behaviour-use interdependence.

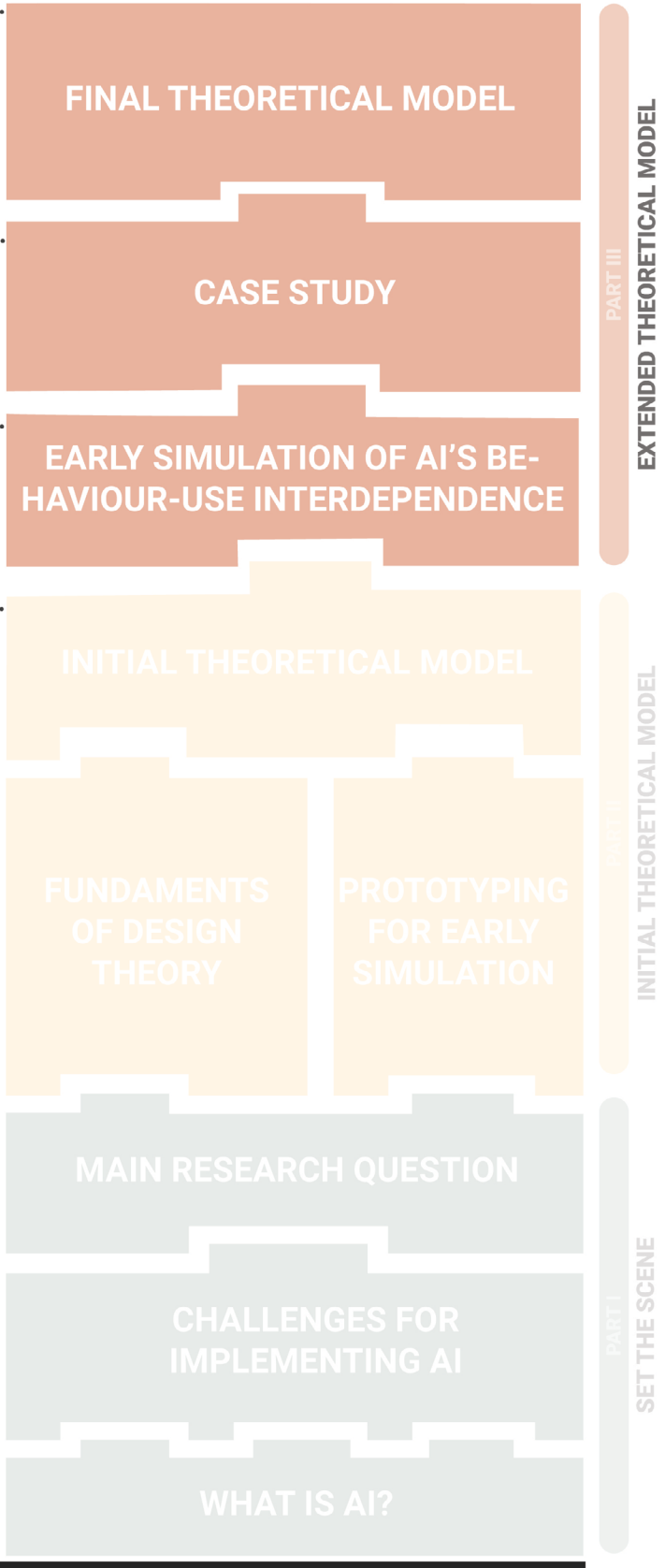
INSIGHTS

We introduced an initial theoretical model based on Design theories. The model is based on the insights generated in Chapters 3 & 4.

First, the similarities of the two main paradigms of Design: (1) a feedback mechanism enables us to better understand the problem; (2) the designer is part of the design process; and (3) unintended outcomes propel the design process forward.

Second, innovative abduction formalises the relationships between behaviour and use and showcases how these can be used to design a new solution.
We reviewed two formulations:
Roozenburg: *((form ^ actuation) -->mode of action) -->purpose*
Dorst: *WHAT + HOW -->VALUE*

Third, prototyping supports us in observing the different types of outcomes and uses the behaviour elicits. It serves as a bridge between behaviour and use; and is conductive to what Magnani (2008) terms manipulative abduction.



How can Design theories support the design and implementation of AI systems into complex contexts?

CHAPTER 6

EARLY SIMULATION OF AI SYSTEM'S BEHAVIOUR-USE INTERDEPENDENCE

At the beginning of the dissertation, we set out to explore how to devise a theoretical model that supports the early simulation of AI systems' behaviour-use interdependence. We then postulated that Design theories are well-equipped to address the interdependence due to the core of design reasoning – innovative abduction. This type of abduction explicitly formalises the relationship among the behaviour and use (termed *mode of action* and *actuation* by Roozenburg (1993), respectively). Using these as a starting point, in Part II, we explored Design theories and the manners in which they could aid us in simulating behaviour-use interdependence of AI systems. The part was concluded in Chapter 5 where we introduced a theoretical model based on these theories that can support us in answering the main research question of the dissertation: *"How can a theoretical model be designed that supports the early simulation of AI systems' behaviour-use interdependence by utilising Design theories?"*.

The Design theories we utilised to provide an answer, however, have been developed for the design of products and services (see Chapters 3 and 4). The

purpose of this dissertation is to define how the behaviour-use interdependence of ***AI systems*** can be simulated. Unlike products or services, the behaviour of an AI system is continuously adapting to the data its users generate. For instance, a reinforcement learning model trained to maximise long-term profit can learn short-term trading strategies based on its own past actions and the manner in which it is being used (Rahwan et al., 2019). This continuous automated improvement leads to better decision-making mechanisms. However, it also gives rise to a multitude of unintended behaviours and uses. Such are numerous and continue to be hard to influence (ibid) due to the unparallel level of scale and personalisation AI systems support (Amodei et al., 2016). Therefore, in Part III of the dissertation, to which this chapter serves as a start, we explore further how to simulate the behaviour-use interdependence in the context of designing AI systems.

In order to concretely contextualise the theoretical model within the realm of designing AI systems, we begin our exploration with an empirical example that serves as a platform for the model's application and conceptual evolution. The example allows us to explore the manners in which the model can inform and shape real-world decision-making and AI development processes. It also serves as an epistemic mediator, unveiling avenues for refining and augmenting the theoretical model's conceptual framing. Thereby, facilitating the initial model's adaptation to the novel context. As such, it enables us to systematically identify the nuances that necessitate model refinement and pinpoint the potential insights that merit further exploration.

The chapter is structured as follows. First, we provide a brief description of the example's background and the design process the team followed. This is complemented with an overview on how their design process maps onto the variables of the theoretical model introduced in Chapter 5. The empirical example overview is followed by a discussion section where we elaborate upon the insights that can be discerned by the example and use them as a starting point for the further conceptual development of the theoretical model. Finally, we also position these developments within the broader landscape of existing

Design Theory literature, underscoring potential contributions to the evolving discourse in the field.

Example's background

The empirical example presented in this chapter comes from my teaching practice. The controlled environment of an educational setting makes it easy to isolate and study the different manners in which the theoretical model could be applied. Moreover, the problem-solving and reasoning of students on the *advanced beginner level* (Lawson & Dorst, 2009) can be explicitly seen in their work (Dorst & Reymen, 2004). As such, the example showcases how a three-person student team, each with an average of four years of experience, successfully managed to simulate the behaviour-use interdependence of an AI system four times. This resulted in a conceptual design for an AI-powered fitness app that supports its users to establish healthier diet and exercise routines by provoking them to reflect on their choices. The team achieved this by utilising simple non-functional prototypes, which resulted in multiple elicited (intended and unintended) uses and outcomes. As such, this case presents us with the type of complex system for which we devised the theoretical model. It is defined by a collection of interconnected and interdependent social (i.e., *users who want to establish healthier habits*), physical (i.e., *the user's environment in terms of access to exercise tools and healthy food/meal choices*), and technical (i.e., *the AI system to be designed*) elements.

The design project was carried out during one of the first-year obligatory 20-week long courses of a master's programme in the faculty of Industrial Design Engineering at TU Delft from February to May 2020. At the beginning of the course, the students were presented with an explanation of the theoretical model during a two-hour session where they could also ask questions. They were then provided with a conference paper (Stoimenova & Kleinsmann, 2020) further explicating the model. At the end of the session, they were given the following task: *design a concept for an AI system by applying the theoretical model*.

During the course, I assumed the role of a teacher and supervisor. As such, I supported them throughout their design process. I also regularly answered their questions about applying the theoretical model to simulate the behaviour-use interdependence of AI systems. Hence, we had regular bi-weekly one-hour meetings, which provided me with an in-depth understanding of the team's design process. Furthermore, I also had access to their data, designs, and conclusions.

Design process

The team began their design process (see Figure 9) by defining a purpose: *support people in establishing and sustaining healthy routines*. To identify ways in which they can achieve it, they conducted initial research on the subject and interviewed five participants they recruited to understand their existing exercise and dietary habits and goals. The insights the team elicited led them to conclude that if they are to achieve their initial purpose, their AI system should give advice *both in alignment with the user's preferences and in deviation from them.* This triggered the team to start ideating on how they could achieve that. Thus, they iteratively designed the first version of their concept by alternating between detailing how their concept should behave (i.e., *provide personalised advice that intentionally suggests new activities that are not aligned with user's preferences*) and how it should be used (e.g., *reflect on the contradictory suggestions*). These, the team then materialised in their first non-functional prototype: a *mock-up of an AI-powered fitness app that guides people through their attempts to lead a healthier lifestyle*.

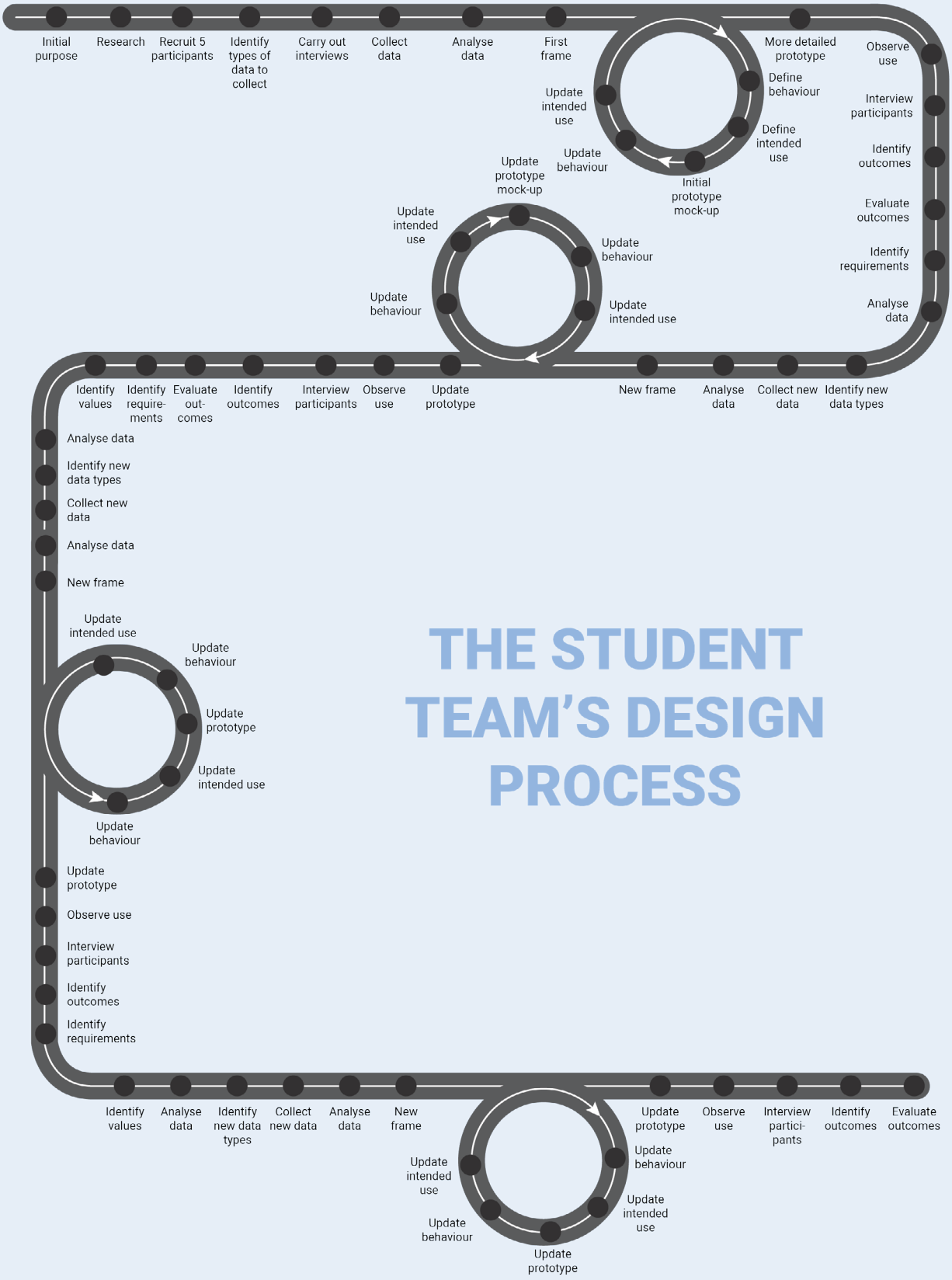
The team then set up the prototype using the "Wizard of Oz" technique and observed that even though their participants had no problem using the concept, they did not engage with all of its features. To better understand why this happened, the team conducted feedback interviews with their participants. This allowed the team to discern their prototype's direct outcomes (e.g., *participants did not find the suggestions personalised enough, but felt in control of their routines*). The team then carefully evaluated the outcomes against the initial purpose and classified them as intended (i.e., *participants feel in control*) and unintended (i.e., *suggestions that were meant to be personalised were not*

seen as such by the participants). Finally, they identified the requirements their concept ought to fulfil (e.g., *app's suggestions should feel personalised*).

The unintended outcomes and the requirements served as the starting point for the team's second iteration. They started by evaluating the data they had gathered thus far from their observations on the prototype's use and the feedback interviews. This resulted in identifying two new types of data the team needed to collect*:* (1) *a collection of relevant meals* and (2) *users' weekly schedule*. The analysis of the existing and newly gathered data led the team to reason that *"If the AI intentionally makes mistakes in its advice, the app can elicit the personal preferences of its users more easily".* This served as the starting point for a new ideation step where the team iterated over the manner in which their concept ought to behave (e.g., *some of the provided advice is contradictory to user's preferences*) and be used (e.g., *provide feedback to the system*). This resulted in an addition to their first non-functional prototype, which provides workout schedule and suggests meals to help users achieve their goal. Similar to their first iteration, the team gave their updated prototype to their participants and observed how they used it. As expected, some participants were reluctant to follow the proposed schedule. Others found the suggestions not personalised enough and refused to teach the system. There were also unexpected uses. Some participants expected to find recipes for the suggested meals. Others changed their preferences to align with the suggestions of the app (e.g., *"Again, seafood. I'm not a big fan of seafood. But, it kind of looks very good in this picture. So uhm, let's say okay, not so bad as before, so I can try it").* The team then evaluated the outcomes against their purpose and identified another set of requirements their concept ought to fulfil (e.g., *simplify the meal prepping process*). They also explicitly discussed the values their solution should embody (e.g., *users should not follow the AI's advice blindly*).

The team used the elicited unintended outcomes as a starting point for their third and subsequently fourth iteration. In both iterations, they followed an identical process to the one thus far. They first analysed the data collected from the previous iterations. They then added more data types they needed to collect such as (1) *more exercise options*, (2) *detailed meal suggestions,* (3) *data*

Figure 9 An overview on the design process followed by the student team



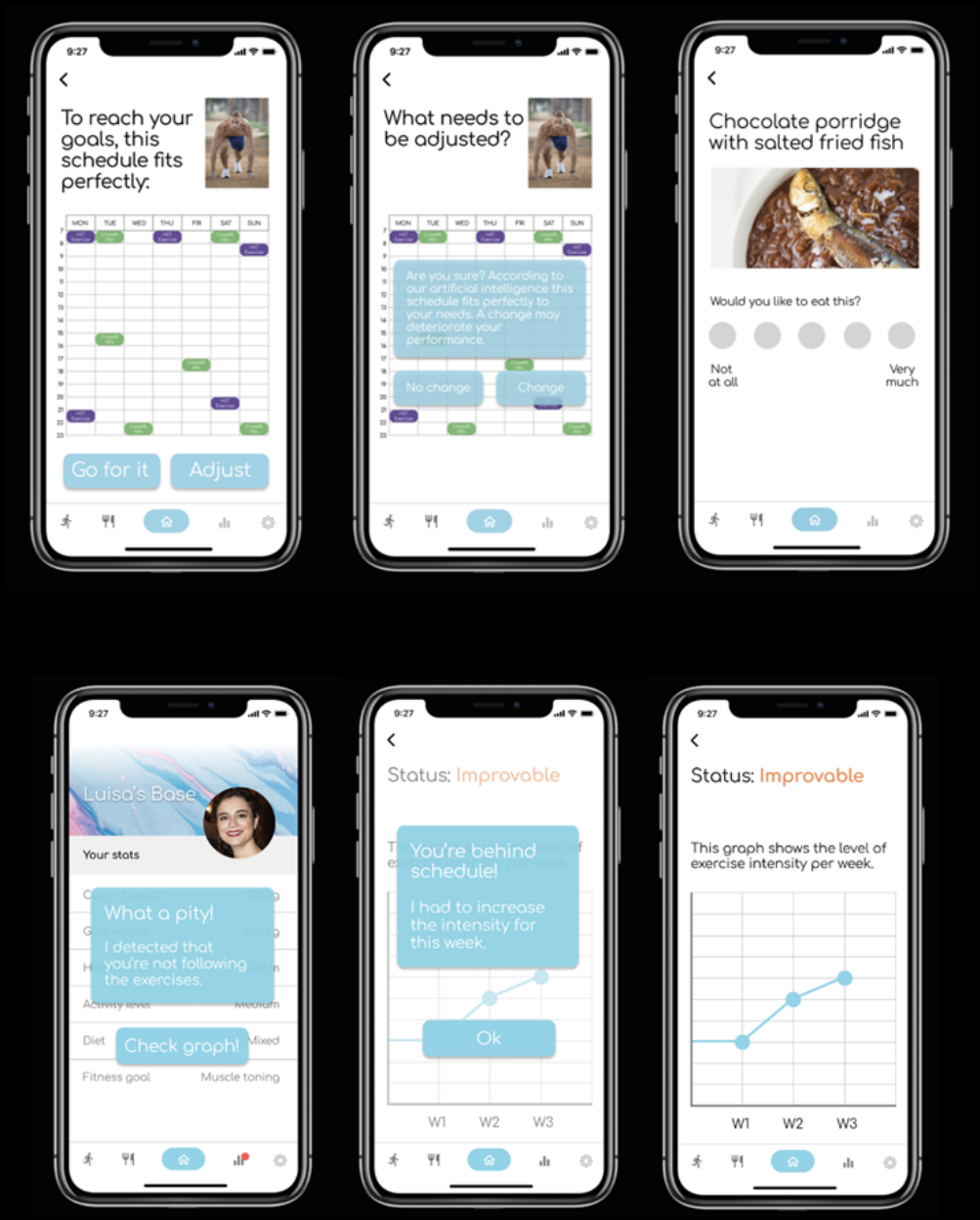


Image 5 Example screens from the app prototype of AI system the team designed.

on the exercises participants performed during the previous week, (4) *a list of products for the meal suggestions*, and (5) *users' photos*. This step was always concluded with a new suggestion on how to approach the situation at hand, which was infallibly followed by an ideation step. The ideation was focused on how their concept's behaviour and intended use should be updated and how these can be reflected into their update to the non-functional prototype used in the previous iterations.

As a result, the team iteratively updated their prototype to include (1) *dish recommendations*, (2) *food delivery option*, (3) *detailed recipes*, (4) *exercise sequence images*, (5) *an AI trainer persona*, (6) *performed exercise intensity*, and (7) *graphs showcasing the user's lack of commitment to their goals*. An overview on the prototypes the team created can be found in Image 5. These prototypes were then given to the participants. By doing so the team managed to elicit a number of both intended and unintended outcomes which supported them in moving forward with the design process, as well as to uncover the requirements and values their concept should embody. This gave them a means to reflect onto their own values, as well as to detail the behaviour of the concept in such a way so that it can address the uses and outcomes they elicited.

Mapping the design process onto the theoretical model

As previously discussed, at the beginning of the design project, the team was given the theoretical model introduced in Chapter 5. As such, they were aware of its variables, used it to support their design process and to simulate the behaviour-use interdependence of their concept. In this section, we discuss the design process the team followed along the variables of the theoretical model, starting with its first row:

purpose + data → frame

At the beginning of their project, the team defined the purpose they wanted to achieve. Doing so gave them the foundation on which they could detail the types of data they needed to collect. The data, analysed in the scope of the purpose, enabled the team to define their frame (e.g., *If the AI intentionally*

makes mistakes in its advice, the app can elicit the personal preferences of its users more easily"). After the first iteration, the team kept updating the types of data they needed and the frames that could enable them to achieve the purpose and address the unintended outcomes they had elicited previously. This new frame then became the starting point for the next row of the theoretical model:

frame + mode of action → prototype

The new (updated) frame served as a starting point for their ideation process. To do so, they first defined the mode of action (e.g., *include deviations in the workouts such as extreme exercises*) which became more detailed with every iteration (e.g., *show user's lack of commitment to the proposed exercises*). Although not prescribed by the model, the team also explicitly defined the manners in which they wanted their prototype to be used (e.g., *reflect on the contradictory suggestions*). These too became more detailed with every iteration through the model (e.g., *rate the proposed suggestions*). The combination of the mode and the intended use led the team to design a (non-functional) prototype. Once the initial version of the prototype was devised, the team also updated the mode and intended use, which led to a new, more detailed version of their prototype. Such iterative behaviour is central to the process of design. For example, in creativity research and especially around the notion of co-evolution of problem and solution space (e.g., Dorst & Cross, 2001; Crilly, 2021), it is well-researched that (parts of) the already formulated solution are habitually used as a starting point to generate new ideas[1].

The resulting prototype then served as the starting point for the third row of the theoretical model:

prototype + actuation → outcomes

In each iteration, once the team had their (updated) prototype, they always gave it to their participants and observed the manners in which they used it

1 These are called solution to solution space transitions and are characterised by synthesis and extension of solution ideas (Cash & Gonçalves, 2017). Further, this iterative behaviour can also be seen as a manifestation of what Schön (1983) terms the backtalk (i.e., design is a reflective conversation between the situation and the designers where the former talks back to the latter).

(i.e., actuation). The combination of the prototype and its actuation resulted in outcomes – both intended (e.g., *participants feel in control of their routines*) and unintended (e.g., *compliance with app suggestions that were contradictory to participant's explicit preferences*). Next to the outcomes, the team also identified requirements (e.g., *provide users with different exercise options*) and made explicit the values they wanted to embed into their concept (e.g., *users should not feel like the design of the app is prejudiced against them*).

The outcomes, requirements and values were then reflected into the subsequent iterations' frames. For instance, one of Iteration 3's outcomes was: *compliance with suggestions of the app even when the suggestion was meant as something the participant will consider a mistake*. The subsequent frame (Iteration 4) became: *If the app provides provocative advice to its users, they will reflect on their choices*. Each frame was then infallibly reflected into the corresponding mode (e.g., *blame the users in lack of commitment*). Hence, with every iteration the mode, intended use and prototype were becoming more detailed. These, in turn, allowed the team to elicit an ever-growing number of both intended and unintended outcomes that informed the manner in which they designed their concept. Similarly, Iteration 1's requirement *apps suggestions should feel personalised*, was followed by Iteration 2's frame: *If the AI intentionally makes mistakes in its advice, the app can elicit the personal preferences of its users more easily.* Further, one of Iteration 2's value was *users should not follow the AI's advice blindly*. The susceptibility of users to the suggestions of the AI was not something the team wanted their concept to support as they saw it as a means to manipulate their users. Therefore, this was reflected in the formulation of Iteration 3 and 4's frames respectively (bolded):

> *If the app provides more nuanced deviations from the preferences of the user,* **it won't be as easy for people to comply with them***, and if*

Undergoing smaller iterations causes designers to "appreciate things in the situation that go beyond their initial perceptions of the problem" (p. 148). Therefore, although not prescribed by the theoretical model, this is an indispensable part of the design process.

the app provides provocative advice to its users, **they will reflect on their choices**.

Using the guidance of the theoretical model, the team managed to simulate the behaviour-use interdependence of their conceptual AI system four times and elicit 11 different (intended and unintended) outcomes by making use of simple prototypes. These iterations allowed them to not only explore the ranges of potential uses and outcomes, but also to gradually build-upon and adjust their mode (i.e., behaviour) so that it can address the elicited unintended outcomes. It also enabled them to identify and elaborate upon the requirements and values their concept ought to fulfil, as well as the data they need to collect. Thus, supporting the team to iteratively design a concept for an AI system that can fulfil the purpose they identified at the beginning of their project. An example on how the team addressed each variable of the model can be found in Table 5.

Table 5 An overview on the formulation of each variable during Iteration 3[2].

Variable	**Performed activities**
Data	As previously + more exercise options, detailed meal suggestions (including ingredients, instructions, nutritional value, time) and food delivery options
Frame	If the app provides more nuanced deviations from the preferences of the user, they won't readily comply with them.
Mode of action	As before + including deviations in the workouts such as extreme exercises (too hard and high intensity, as well as extremely short exercises) and introduce difficult meal plans (e.g., very long preparation times and exotic ingredients).
Intended actuation	Carefully read and reflect on the suggested exercises and meals.
Prototype	As before + update of the app interface to include dish recommendation (and detailed recipes) and food delivery.
Mode of action	As before + assign labels to each user that can showcase them how they performed during the past week.

2 We chose to exemplify Iteration 3 since during this iteration the team already had a more detailed concept. In addition, they managed to elicit the most unintended outcomes during Iteration 3. An overview on all iterations can be found in the Appendix.

Intended actuation	As before + notice the labels the algorithm has assigned to them.
Prototype	As before + image sequences showing different exercises and times.
Actuation	E.g., refusal to perform more challenging exercises; tried to change the proposed schedule; wondered how to provide feedback to the AI so that it would not suggest such exercises; wondered how to make the AI show easier recipes.
Outcomes	Participants are puzzled over the choices the AI made (e.g., one user tried to figure out what he might have said in the first interaction that made the AI think that his diet should be vegetarian: *"well, there's a question I have here... I don't know if the current diet is something that I've said that I have or it's the diet my coach suggested to me..."*); a photo of the female yoga instructor was considered too prejudiced and not inclusive enough for men who like to practice yoga by a male participant); too complex suggestions (e.g., *"40 min? 40 min for a sandwich?"*); compliance with suggestions of the app even though the suggestion was meant as something the participant will consider a mistake (*"Preparation time 8h, okay. Yeah, I mean that also looks very yummy. But I again would not like to cook during the week, something that takes eight hours. Then I would try it out on the weekends one time"*).
Require-ments	Meals should not require too much effort from the user; Introduce an option to shop the ingredients needed for the meal online.
Values	Users should not feel like the design of the app is prejudiced against them

Discussion

As it can be seen from the example, the team used all of the variables of our initial theoretical model while designing their AI-powered fitness app concept. While they made use of all model's variables, they also used variables that were not prescribed: *intended use*, *requirements*, and *values*. In this section, we discuss each of the identified new variables in light of extant literature and the theoretical model.

The intended use

In each iteration the team explicitly defined the manners in which they intended their concept to be used. Let us call this action *intended actuation* to reflect the already existing variable of actuation that signifies use (see Chapter 5). As we saw from the example, the intended actuation always came after the

mode has been defined and before a prototype is devised. Hence, it resides between the variables of mode of action and prototype (i.e., *frame + mode of action + intended actuation → prototype*). This observation is in line with existing theories of design practice. For instance, according to Roozenburg (1993), the prescribed way in which a designed solution is to act is integral to the process of design. Well-designed products always provide a (visual) cue on how they are to be handled (Norman, 2013). For instance, a door has a handlebar, digital solutions come with a pre-catered menu of options and services are guided by their service blueprints. Therefore, if we are to create a model that is an accurate representation of a design process, we need to have a variable to represent the manner in which the designers intend their solution to be actuated.

However, a clear distinction is needed between *intended actuation* and *actuation*. To do so, we propose to rename the latter to **observed actuation** (i.e., *the actions a user performs in order to allow the AI system to function and be "connected" to its immediate environment).* Such an update will allow us to clearly differentiate between the use for which the AI system's behaviour (i.e., mode of action) was designed and the observed types of use that happen once the AI system gets deployed. Doing so also ensures that the theoretical model reflects the ever-evolving nature of AI systems.

Introducing two different types of actuations also has implications for the manner in which we formulate the theoretical model. Namely:

purpose + data → frame

frame + mode of action + ***intended actuation*** *→ prototype*

prototype + ***observed actuation*** *→ outcomes*

Given these changes and considering the different types of abductions introduced in Chapter 3, each row of the model becomes representative of a different type of abduction. The first row (i.e., *purpose + data → frame*) is a clear example of explanatory abduction. This type of abduction allows us to generate a logical explanation for an observation (Roozenburg, 1993) (Chapter 3). For

instance, the team found that their participants find the AI system's suggestions not personalised enough (i.e., an observation). This observation became part of the data they had collected thus far. The team then explained this away by deciding to make their users take a more pro-active role in explicitly teaching the system. The explanation resulted in a new frame: *"If the AI intentionally makes mistakes in its advice, the users will teach the system how to attune to their preferences (hence, feel personalised)"*.

The second updated row of the theoretical model (i.e., *frame + mode of action + intended actuation → prototype*) describes innovative abduction. Innovative abduction is central to conceptual design and allows us to devise innovative solutions to ill-defined problems (e.g., Roozenburg, 1993; Dorst, 2011; 2015) (Chapter 3). In our case, it is also the type of abduction that allows us to design a prototype which embodies the identified frame (which Dong and MacDonald (2016) see as the starting point for innovative abduction), mode of action and intended actuation. As such, the revised formulation of the second row follows the same logic as the formulation of innovative abduction both Roozenburg (1993) and Dorst (2011) provide. For instance, according to Roozenburg, innovative abduction can be represented as:

$$((form \wedge actuation) \rightarrow mode\ of\ action) \rightarrow purpose$$

(i.e., the form and the actuation of the solution directly influence its mode and the combination of these ensures that the initially defined purpose can be achieved). If we are to substitute the form with prototype (the rationale for which we explained in Chapter 5) and the variable of purpose with frame (since the frame implicitly embodies the purpose – see Chapter 3), we arrive at the same inference as Roozenburg:

$$((prototype \wedge intended\ actuation) \rightarrow mode\ of\ action) \rightarrow frame$$

Finally, in Chapters 4 and 5, we already postulated that the deliberate embedding of prototypes within their intended context allows us to elicit an array of potential uses and outcomes. Using prototypes in this capacity is representative of manipulative abduction (Magnani, 2004). This type

of abduction signifies the productive thinking that happens when we use "epistemic mediators" (ibid) such as prototypes to elicit new and still unexpressed information codified within the context. Consequently, the third row of the theoretical model becomes indicative of manipulative abduction: *prototype + observed actuation → outcomes*.

The clear delineation of the three types of abduction allows us to define the relationships among them. As it can be seen from the model, the result of each type of abduction serves as the starting point for the next. For instance, once we have defined a frame (which is the result of explanatory abduction), we can engage in activities representative of innovative abduction (e.g., defining a mode of action) and so on. The manners in which these types of abductive reasoning influence each other will be further explored in the next chapter.

Requirements

Requirements always appeared once the observed actuations and outcomes have been elicited and before a new iteration had begun. In effect, requirements served as the bridge between the different iterations. As such, they played a two-fold role. On the one hand, they served as a benchmark against which the team evaluated the elicited actuations and outcomes. This is in line with the paradigms of both Rational Problem Solving and Reflective Practice (Chapter 3). For Simon (1996), for instance, when searching for the solution that can satisfice the given (or selected) purpose, it is the requirements that indicate whether the designer has managed to find a satisficing solution. This is also in line with the manner in which Schön (1983) claims a designer decides whether to end the design process: by looking at the outcomes her moves have created and asking herself: "*Do you get what you intend?*" (p. 146).

On the other hand, the team generated new requirements during all but one of their iterations through the theoretical model. These were formulated as a result of the elicited actuations and outcomes. Once the designer observes how the prototype is actuated, these observations could lead to certain suggestions on what to improve or avoid in the next iteration. This process resembles what Schön (1983) terms as reflection-in-action. According to him, it is during

reflection-in-action when "*the situation talks back, the practitioner listens; and as he appreciates what he hears, he reframes the situation once again*" (p. 131). Thus, it allows the designer to better understand the problematic situation at hand and design a new intervention to address it (i.e., playing the role of a bridge between outcomes and new frames (see Chapter 3)). As such, this two-fold role requirements play will be further investigated in Chapter 7.

Personal values

The theoretical model introduced in Chapter 5 does not address the embedding of personal values into the design of an AI system. However, the manner in which the team implicitly embedded their own values into their concept brings to the fore an important aspect of Design Theory. First, recent design theories see designing as a value-producing activity (e.g., Friedman, 1996; Rosenman & Gero, 1998; van Onselen, 2022)
. Second, both Simon (1996) and Schön (1983) acknowledge the influence a designer has on the solutions she is designing. In fact, Schön (ibid) sees the designer as the one who decides whether to end the design process. She does so by looking at the outcomes her moves have created and then asks herself: "*Do you like what you get?*" (Schön, 1983, p. 146). Therefore, similarly to requirements, the values one holds as important are used to evaluate the outcomes a solution has produced.

On the other hand, we saw that values emerged (i.e., were made explicit) as a result of the elicited actuations and outcomes. For instance, it was only when the team noticed that one of their participants considered pictures of female yoga instructors to be prejudiced, that they made their value of inclusivity explicit in their subsequent prototypes. Hence, the role values serve during the application of the theoretical model could be likened to that of requirements. Therefore, an updated version of the theoretical model ought to consider and reflect the role personal values play when designing a new solution. Assigning a variable that can make the otherwise implicit embedding of personal values explicit could also aid us in having a clear overview on the types of values the designing team decides to embed into an AI system. Doing so is important if we are to audit the AI system and openly discuss the potential biases that

might stem from the embedded values. However, we should still avoid potential confusion with the dominant understanding of values in the field of AI (see Chapter 5). A first step in doing so is to clearly delineate the way we define values. Hence, we adopt the following definition: "*an enduring belief that a specific mode of conduct or end-state of existence is personally or socially preferable to an opposite or converse mode of conduct or end-state of existence*" (Rokeach, 1973, p. 5).

Conclusion and further research

The Design theories we discussed in Part II of the dissertation allowed us to devise an initial theoretical model that provides an answer to the main research question of the dissertation: ***"How can a theoretical model be designed that supports the early simulation of AI systems' behaviour-use interdependence by utilising Design theories?"***. However, these theories have been developed for the design of products and services. Therefore, in this chapter we set out to further explore how to simulate the behaviour-use interdependence in the context of designing AI systems.

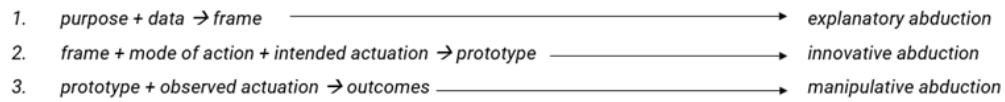


Figure 10 An overview on the abduction types the model is conductive to.

We began our exploration by presenting an empirical example from my teaching practice of a team comprised of three design students who simulated the behaviour-use interdependence of the concept they were designing. The example allowed us to render the theoretical model into the desired context, as well as to provide us with foundation for further research. Three insights emerged from the example.

First, in each of their iterations, the team explicitly defined their intended actuations. This insight allowed us to create a distinct and explicit demarcation between intended actuation and observed actuation. A separation that aligns with the inherent nature of AI systems where the manner in which one decides to use a solution does not necessarily align with the manner the designer

intended it to be used. Hence, we could delineate the feedback loops that define the manner in which an AI system will behave.

The addition of this variable allowed us to connect each row of the theoretical model with a different type of abductive reasoning: explanatory, innovative, and manipulative, respectively (see Figure 10). As discussed in Chapter 3, Design theories use only explanatory and innovative abduction to explicate the manners in which synthesis is carried out in Design. While the initial model presented in Chapter 5 acknowledged the three types, it lacked a clear explanation of their relationships. The conceptual refinement proposed in this chapter provides a more robust framework for understanding the nuanced role each form plays in the reasoning process utilised while designing. As such, this new formulation provides us with the foundation on which further conceptual exploration can be carried out to better understand how designers apply the three different types of abduction.

Second, Design theories provide conflicting accounts on how the different types of abduction are related and influence each other. For instance, both Roozenburg (1993) and Dorst (e.g., 2011) see innovative abduction (i.e., abduction-2) as central to the act of designing. However, unlike Roozenburg, Dorst claims that it is oftentimes followed by explanatory abduction (i.e., abduction-1)[3] to ensure the desired value can be achieved (p. 523) (see Chapter 3). On the other hand, Dong and MacDonald (2016) and Kroll and Koskela (2016) claim that when designing new products, services, and systems one uses explanatory abduction first and only then moves to innovative. Furthermore, to our knowledge, none of these scholars discusses the manner in which manipulative abduction impacts the cognitive act of designing. Yet, the student team we discussed in this chapter made use of all three in order to simulate the behaviour-use interdependence of their concept. Given the conflicting statements on the relationship between explanatory and innovative abduction, and the insights generated in this chapter, it is our contention that

3 An overview on why abduction-2 cannot precede abduction-1 can be found in Kroll and Koskela (2016).

further research is needed into the manner in which the three abduction types influence each other.

Third, as discussed in Chapter 3, it is widely acknowledged both within and outside the field of Design that abduction is intricately connected to deduction and induction. For instance, according to Roozenburg (1993), the outcome of innovative abduction needs to be tested by deduction and induction (p. 10). This sentiment is echoed by the work of Dorst (2011), Kroll and Koskela (e.g., 2015) and Dong and colleagues (e.g., 2016). However, to our knowledge, the connection among the three types of reasoning has not yet been made explicit. There is a notable exception coming from the work of Dong and colleagues around the concept of *generative sensing* (e.g., 2016) which delves into the relationship between abduction and deduction. However, their work does not provide a distinct separation among the forms of abduction and their connection with induction. The insights we gleaned from the empirical example discussed in this chapter, especially around the constructs of *requirements* and *values* could provide us with an interesting vantage point to address this challenge. For instance, both requirements and values are considered to be means for evaluation in extant Design literature. The former has been traditionally used to evaluate design concepts. Hence, a wealth of methods have been created to support designers in applying requirements to evaluate a design concept. Values, on the other hand, can also serve an evaluative role, although the manner in which they do so has been discussed in more implicit ways. For instance, as discussed in Chapter 3, according to Schön (1984), one of the most important questions a designer asks herself while evaluating the outcomes of her moves stems from her personal values: "*Do you like what you get?*" (p. 146). Using values to evaluate a design concept is also a common practice in the Value Sensitive Design approach (Friedman, 1996). Yet, requirements and values have not been connected to the cognitive act of designing or explicated from this vantage point. Therefore, the patterns in which these constructs emerge and are applied could offer Design scholars a new

vantage point for understanding the interplay among abduction, deduction, and potentially induction, too.

These three insights serve as a foundation for further conceptual investigation as they showcase that to be able to simulate behaviour-use interdependence of AI systems, Design theories need to be extended. We explore all of them in the next chapter. In it, we discuss the application of the theoretical model to the conceptual design of an in-car AI system for a large multinational automaker.

CHAPTER 7

CASE STUDY: DESIGNING A SMART IN-CAR SYSTEM CONCEPT FOR A LARGE AUTOMAKER

The main premise of the dissertation is that one can design a theoretical model that supports the early simulation of **AI systems'** behaviour-use interdependence by utilising Design theories. The example discussed in Chapter 6 suggests that the theoretical model we devised in Chapter 5 covers large parts of the process of designing AI systems concepts. However, it does not manage to fully capture the design activities that allow us to simulate the behaviour-use interdependence of an AI system.

In the previous chapter, we employed an empirical example that served as a scaffold around which we furthered the conceptual development of the theoretical model we introduced in Chapter 5. This approach led us to the identification of several avenues in which Design theories need to be further extended. More specifically, it uncovered three areas for further research: (1) the manner in which designers apply the three different types of abductive

reasoning, (2) the manner in which these types influence each other; and (3) the manner in which non-abductive activities are used when applying the theoretical model. In this chapter, we again apply the same approach of using empirical investigation as the basis upon which we further the conceptual development of the theoretical model. However, we apply a more in-depth and structured empirical investigation method: a case study describing a design project. It is our contention that such an approach can better inform the manners in which we can both further our efforts of conceptually developing the theoretical model and present more grounded insights and suggestions on how Design theories could be extended.

Therefore, the main research question of the chapter becomes: **"How can Design theories be extended to support the early simulation of AI systems' behaviour-use interdependence?".** We operationalise it through the following three sub-research questions (one for each of the aforementioned research areas):

1. *"How and where in the process do designers apply each of the three abductive reasoning types?"*

2. *"How do the three types of abductive reasoning influence each other?"*

3. *"Where in the process do designers make use of non-abductive activities?"*

To provide an answer to these, we explore the application of the theoretical model to the design of an AI system during a five-month long case study. The study followed the application of the adjusted model (see Chapter 6) to the conceptual design of an in-car AI system for a large international automaker (henceforth referred to as X). The design project was carried out by an expert designer (myself), who was supported by three other expert designers – two from the client side and one from my doctoral supervisory team. As such, the chapter has the following structure. First, we present the case study background. This is followed by an explanation on the utilised data collection and analysis methods. Then, we introduce the Results section of the chapter

where we use empirical data to answer each of the sub research questions. Finally, in the Discussion section, we address the main research question of the chapter, by extending the conceptual development of the theoretical model.

Case background

We executed this design project in the field of mobility together with an automotive company – the client. They were in their initial exploration stages of solutions that can facilitate the change from internal combustion engine (ICE) vehicles to electric vehicles (EVs). Therefore, at the beginning of the project, the following situation was presented by the client. One of the biggest hurdles to the adoption of EVs is the time it takes a user to charge their car. While ICE cars can be fuelled in just 5 to 10 minutes at a gas station, an EV takes between 40 minutes and 12 hours to be fully charged. Therefore, the solution we were about to design had to strike a balance between convenience (i.e., ensuring that users do not plan their days around charging their cars) and performance (i.e., ensuring that we can preserve the battery performance for as long as possible). Finally, the potential solution had to address the limited availability of accessible charging infrastructure, especially public charging points. Given this context, the agreed-upon initial purpose was: "*Enable the most optimal behaviour of as many EV owners as possible*".

The design project took five months from May to October 2021, and it was focused on the initial stages of the design process. Namely, while the client was already interested in the area of devising AI systems that can support EV owners, they did not yet know what a system like that should do, how it should behave and what user needs it could address. As such, this case presented us with the type of complex system for which we devised the theoretical model. It was defined by a collection of interconnected and interdependent social (i.e., *users and other EV drivers*), physical (i.e., *the car and the existing charging infrastructure*), and technical (i.e., *the AI system to be designed*) elements. Furthermore, the exploratory nature of the design project allowed us to observe the manners in which the theoretical model was applied in practice. Throughout the project, I acted as the principal designer, and collaborated with designers from the client side. The decision for me to play the role of the

principal designer was made for two reasons. First, my background classifies me as an expert designer (Dreyfus & Dreyfus, 2005; Lawson & Dorst, 2013). I have received my bachelor and master's formal education in Industrial Design Engineering. I have also been a practicing designer for over five years. Second, I have an intimate level of understanding of the theoretical model. Therefore, I could ensure the model was applied as intended. This knowledge is especially important since the model is still in its theory development stage (Cash, 2018).

Data collection

Next to my role of a principal designer, I was also the main researcher guiding and analysing the study. Combining these two roles led to a rich first-hand understanding of the design process, and its emergence (van Oorschot et al. 2022, p. 2). Furthermore, due to my dual role, I had access to all of the generated data throughout the study.

First, I kept a detailed design journal where I collected all my sketches, descriptions on why certain decisions were made, data analysis, ideation process, interview and observation notes, and generated insights and decisions. Example pages from the design journal can be found in Image 6 where one can see how sketches, descriptions of made decisions, and ideation process come together.`

Second, I regularly had a one-hour team meeting with the two expert designers from the client side. During these meetings I reported on my progress. We also discussed potential design directions (e.g., what features to introduce in the prototypes), but also values and requirements important for both the client and me). As such, the meetings served as a chance for alignment between us, as well as an important step in the design process with implications on how to continue. All meetings were audio recorded and transcribed. Moreover, a short PowerPoint presentation was shared during each of the meetings. All presentations detailing my progress were saved and dated.

Third, I carried out a creative session with three expert designers (two of them work for X and the third one is the co-promotor of this doctoral research). The

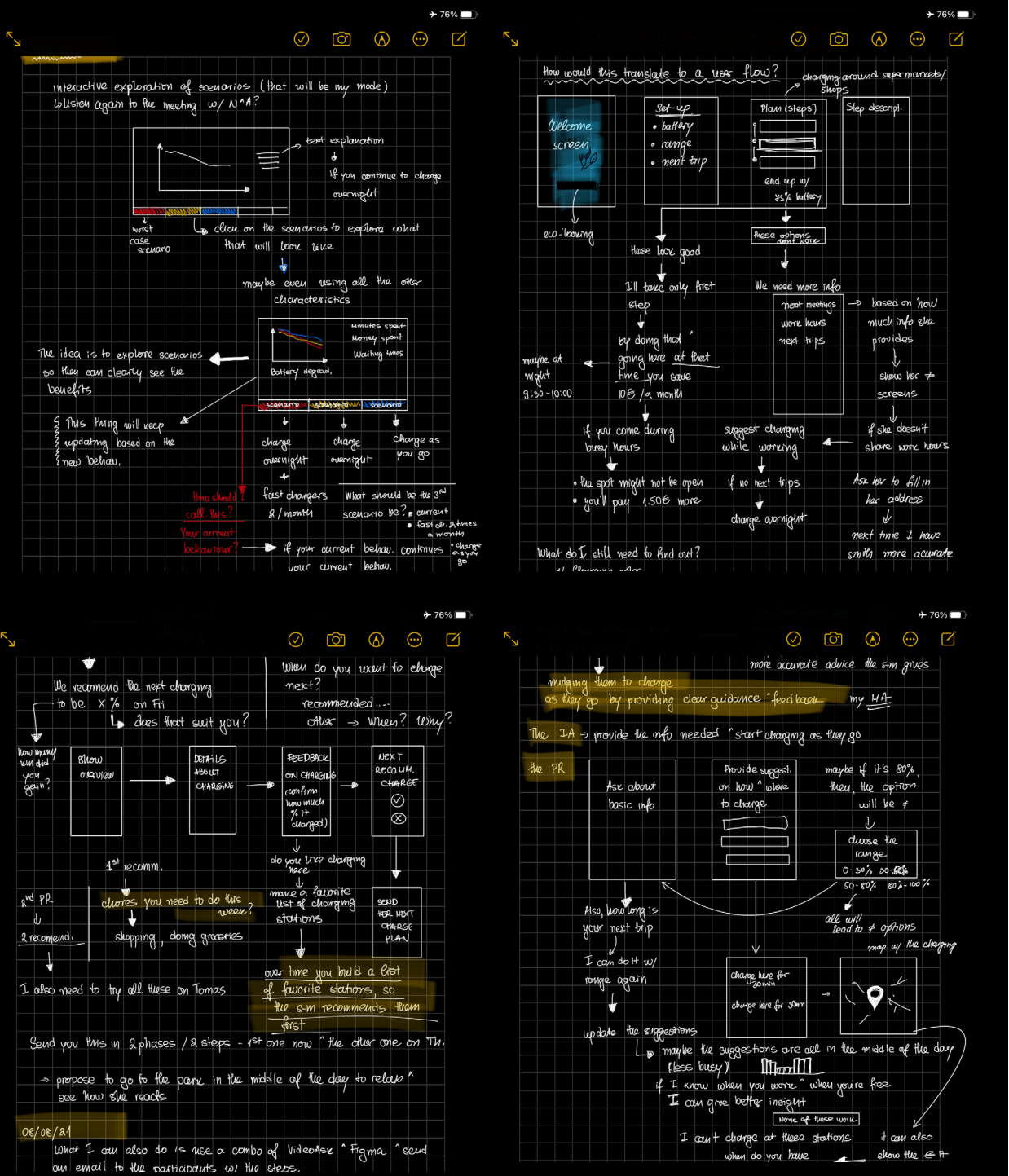


Image 6 Example pages from my design journal showcasing different sketches and the rationale behind them.

goal of the session was to jointly generate new frames based on the collected thus far data. Due to Covid-19 restrictions, the session was carried out online using a virtual collaboration software (i.e., Miro). All activities from the session were downloaded and analysed. The creative session was audio-recorded and transcribed.

Finally, from mid-July to mid-October 2021, two owners of EVs participated in the project. During this period, I carried out seven interviews with each of them. In addition, both of them actuated six of the devised prototypes (see Results section). All of the interviews and user tests were audio and video-recorded and subsequently transcribed.

The combination of these allowed us to create a detailed representation of the activities carried out during this design project in all of its richness – from activity descriptions, through sketches, collection of all ideas, and discussions between myself and the client, to the rationale for each of the made decisions.

Data analysis

The data analysis process was carried out by me. It commenced with iterative coding of the collected data. First, I pre-coded design activities (i.e., by highlighting rich or significant quotes found in the collected materials) (Saldaña, 2013). In order to identify design activities, I followed the definition Pedgley (2007) put forward: *"Design activity encompasses cerebral activities including thinking, imaging, and decision-making as well as practical and externally perceptible activities such as information gathering, drawing and model-making*". Each activity was either performed by me (the designer), the team, or the two recruited participants. Once all activities were pre-coded, I clustered them around the variables identified in Chapter 6: **purpose, data, frame, mode of action, intended actuation, prototype, observed actuation, outcomes, values,** and **requirements**. To gauge which instance could be grouped under which variable, I used the definitions presented in Table 6. The definitions come from both Chapter 5 and Chapter 6. Where a direct quote was

not available (e.g., a sketch or a prototype), I translated the sketch into a written description[1].

Table 6 An overview of the variables' definitions, paired with an example quote (quotes from the principal designer are followed by PD, quotes from the other designers – as C and the participants as U) and source.

Variable	Definitions	Example	Source
purpose	The goal of an AI system.	*"Optimise the charging for people who cannot charge at home"* (C).	team meeting
data	Records produced by experiments and measurements that serve as evidence for the existence or features of a phenomenon.	*"I looked through the forum threads on Quora when it comes to EVs and especially – what people do when they're charging, where they charge and the advantages, they perceive from driving an EV (as well as (super)charging on long trips)."* (PD)	journal
frame	A cognitive act of looking at a problem situation from a specific viewpoint that informs how the problem can be solved.	*"They lack in motivation, so what we should present to them is something that will help them to increase their motivation to perform the action."* (C)	team meeting
mode of action	(Functional) behaviour of the AI system in response to influences exerted on it from its environment.	*"What if we insert some provocative statements to gauge how easily the system can sway their preferences – e.g., "go to a park for a walk in the middle of the day so that they can charge", show costs, show route, etc."* (PD)	team meeting
intended actuation	The intended by design action of a user that allows the AI system to function and be "connected" to its immediate environment.	*"They need to select options and provide information on their current routines, next trips, home address, important aspects to them, agenda, etc."* (PD)	team meeting
prototype	An externalised representation of a potential AI system.	*An in-car system that guides drivers when to charge, where and for how long.* (PD)	journal

1 The description process was aided by the fact that when designing I tried to explicitly externalise (i.e., write down, sketch out) every single thought.

observed actuation	The actions a user performs in order to allow the AI system to function and be "connected" to its immediate environment.	*"I left it at the charging station longer because last week I had a problem with the charger where sometimes it didn't charge and sometimes it did, but I really didn't need the car for the day, so I left it out and eventually it charged…"* (U)	user test
outcome	The intended and unintended results of the AI system's actuation.	*"I'm concerned about my privacy. What if I give my home address to the system and full access to my agenda and the car gets hacked… Then they [burglars] will have all the information when I' not at home so they can rob me."* (U)	user test
requirements	The performance specification of the AI system that limits the range of acceptable solutions.	*"The solution should support users to reduce the time they leave their car plugged-in as much as possible."* (C)	team meeting
values	An enduring belief that a specific mode of conduct or end-state of existence is personally or socially preferable to an opposite or converse mode of conduct or end-state of existence.	*"Support people in doing healthy activities while charging."* (PD)	team meeting

After all data were coded, I compiled a chronological list of activities (Table 7). For each activity I kept a log of underlying data sources to create a chain of evidence (Yin, 2009). Once all activities were coded and ordered chronologically, I clustered similar codes in different descriptive categories (e.g., "collect data", "data analysis", "identify requirements", "formulate a frame").

Table 7 A snippet from the chronological event list where each category, code, quote, source, and time stamp are given.

Category	**Code**	**Quote**	**Source**	**When**
Collect data	Data	*"Look into public datasets on EV charging patterns in the US, UK, and the Netherlands."*	journal	07/06/2021
Data analysis	Data	*"If we're to design something for the car, we need to take into consideration that the car will most probably be resold."*	team meeting	10/06/2021

	Data	*"People who own EVs have charging anxiety."*	team meeting	10/06/2021
	Data	*"Even though there're not that many cars in the Netherlands, people are still experiencing busy chargers."*	team meeting	10/06/2021
	Data	*"When most people have an EV, chargers will probably never be enough, people will have the feeling that they have to wait."*	team meeting	10/06/2021
	Data	*"The existing grid will most probably not be able to support the overnight charging of all these new cars."*	team meeting	10/06/2021
	Data	*"Regardless of the country from which it comes, there are similar tendencies in the behaviour of people – whether it's Norway, Netherlands, UK, US, it's very similar."*	team meeting	10/06/2021
Identify requirements	Requirements	*"We just have to prepare the situation in a way that's somehow playful and make them feel like they're on an adventure."*	team meeting	10/06/2021
Formulate a frame	Frame	*"The fact that they feel like a novice is actually not a bad thing because we actually need them to behave in a different way, and if you think about how a novice behaves, one of the most important elements is that you feel like you're learning, and that you're uncovering new things. That's the biggest excitement of owning something new."*	team meeting	10/06/2021
Identify values	Value	*"I like the idea of juggling a lot of things and finding a way to figure out how we can support you and to what an extent do people want the car to do that."*	team meeting	10/06/2021
Formulate new frames	Frame	*"If we see the car as a symbol of freedom, then we need to ensure that their owners don't have to think about charging ever again."*	journal	14/06/2021
	Frame	*"If we want to scale EV ownership, we need to make sure that people who cannot charge at home will not find owning an EV a hassle."*	journal	14/06/2021
	Frame	*"Feeling like a novice is a good thing because it can bring you sense of accomplishment when you learn new things."*	journal	14/06/2021

	Frame	*"Maybe we should intentionally make people feel like novices so that they can learn the new behaviour."*	journal	14/06/2021
Identify values	Value	*"It should feel like new shoes that don't give you blisters."*	team meeting	14/06/2021
Identify requirements	Requirements	*"Introduce new habits and do it gradually."*	team meeting	14/06/2021
	Requirements	*"We can't change infrastructure, nor hardware."*	team meeting	14/06/2021
Formulate a new mode	Mode	*"The car starts by supporting them in charging as they are used to, and then gradually starts suggesting to places where they can start charging as they go. it should first understand their existing routines and then establish the feeling of learning."*	journal	14/06/2021

The chronologically ordered coded activities, allowed us to define three distinct sequential temporal brackets (Langley, 1999) that defined the design process. The first one, "**Understand context"**, contains design activities that allowed the team to better understand the context for which the concept had to be designed. Hence, it ended when a new, better-defined, purpose was formulated. The second bracket, "**Devise initial concept"**, contains the design activities that enabled the team to design three different concepts for an AI system and evaluate them together. The bracket was concluded with the definition of a new sub-frame that pointed the team at the direction the detailed concept should follow. Finally, the third bracket, "**Develop the concept"**, contains the activities the team carried out to detail the concept by simulating its behaviour-use interdependence. These brackets are not "phases in the sense of predictable sequential process" but they can be used as a structured way of describing activities (Langley, 1999).

In addition, each of the identified activities was labelled as either being representative of explanatory, innovative, manipulative or no abduction (with a different corresponding colour). To match identified activities with abduction types, we used the insights on the relationship among the three types of

abduction discussed in Chapter 6 (Figure 11). Namely, if we have an activity coded as a *mode of action*, we visualised the activity as part of innovative abduction. Activities that were not coded as one of the model's variables (including *values* and *requirements*) were considered to be representative of no abduction.

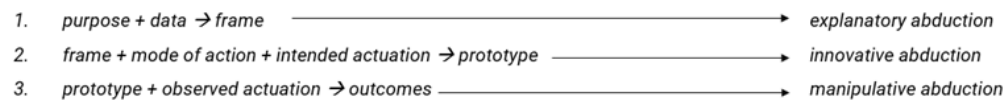
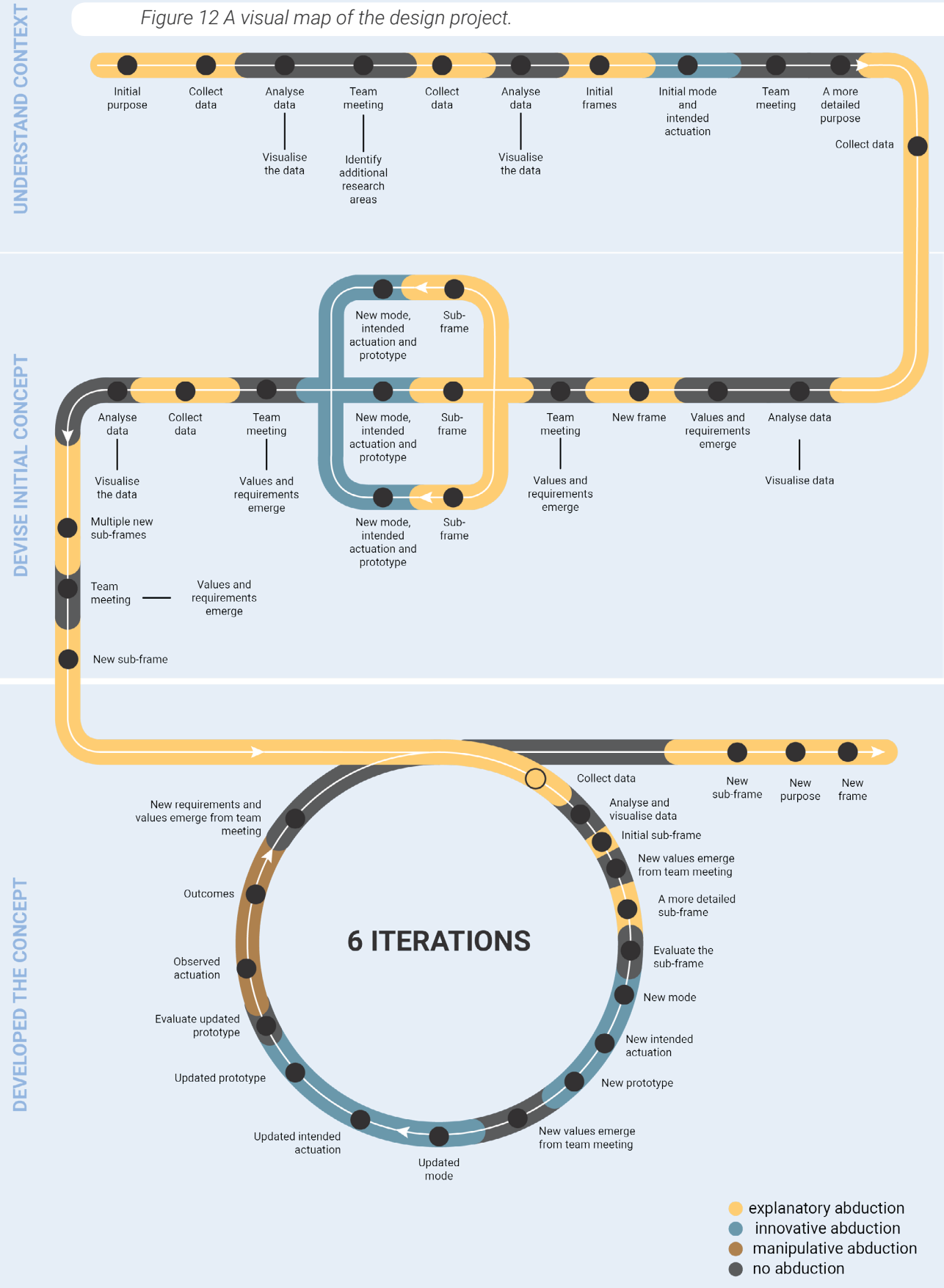


Figure 11 An overview on the abduction types the model is conductive to.

The combination of the temporal brackets and their corresponding activities were then placed on a visual map – a chronological timeline detailing the categories and their temporal occurrence (Langley, 1999). The map (Figure 12) is divided in three rectangles (one for each bracket), connected by a timeline. The timeline is colour-coded in accordance with the abduction type its corresponding activities represent. The size of each rectangle and its timeline identifies the time it took for the activities in a bracket to be performed. Some concessions had to be made for ease of visualisation. For instance, although the third temporal bracket visually takes approximately the same space as the other two, it took longer than both of them combined. Further, since all activities in the third bracket followed the same pattern, only one iteration through the model is visualised.

Finally, combining all of the thus far uncovered insights, we wrote three descriptive narratives (Langley, 1999) about the design project: one for each of the temporal brackets. The narrative writing was guided by the activity list and the raw data collected during the project. Each narrative describes the activities within the corresponding bracket, their connections to the variables of the model, as well as how these contributed to the devising of a concept for an in-car AI system. These can be found in the Results section of the chapter.

Figure 12 A visual map of the design project.



Results

The application of the theoretical model

During this project, we designed a concept for an in-car AI system that supports EV owners who cannot charge at home to start charging as they go (i.e., instead of waiting for their car to be fully charged, they charge enough only to get to their next destination). In order to do so, the design project went through three distinct temporal brackets. Each one of them is discussed in detail below as an answer to the first part (in bold) of the first sub research question of the chapter: "***How** and where in the process **do designers apply each of the three abductive reasoning types**?"*.

Temporal bracket 1: Understand context

The starting point for the first bracket was the initial purpose the client put forward at the beginning of the design project: "*Enable the most optimal behaviour of as many EV drivers as possible*". In order to design an AI system that could fulfil the given purpose, I began by collecting data on adoption rates of EVs across the United States, Europe, and Asia, the time people spend on charging their car and the challenges they experience in doing so. The data collection was followed by analysis and data visualisations (see Image 7), usually carried out prior and during our weekly team meetings. These meetings were structured as follows: (1) presenting the gathered data and drawing initial conclusions from it; and (2) together with the other two designers, further analysing the data, and deciding on next steps – e.g., collect additional data on charging routines of EV owners. Stemming from the collected data, we then generated multiple different frames (e.g., "*If the solution facilitates a mindset change in EV owners, then we can ensure optimal charging behaviour"),* and an initial mode of action: "*guide drivers when, where and for how long to charge*". After the second team meeting, a new, more detailed, purpose for the project was selected: "*Help EV owners who cannot charge at home to start charging as they go*".

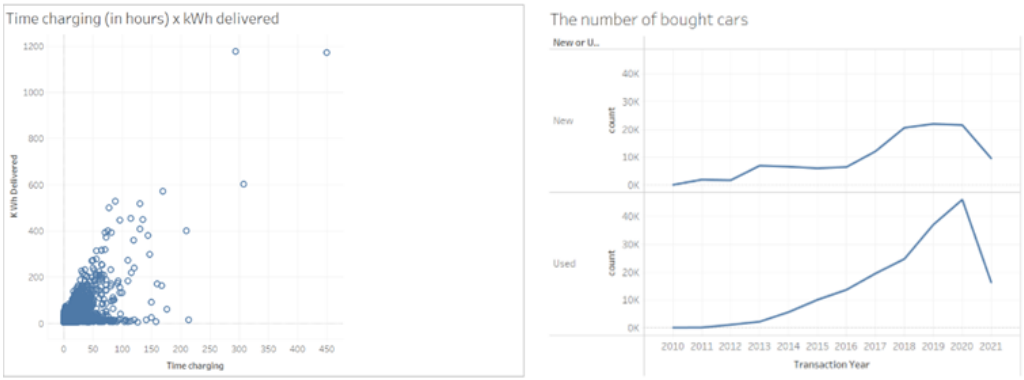


Image 7 Examples of data visualisations used during the first temporal bracket.

Temporal bracket 2: Devise initial concept

The second temporal bracket of the project was focused on designing the initial concept for the AI system. As such, using the new more detailed purpose, I continued with a new round of data collection on the routines EV owners who cannot charge in their homes follow, the existing charging infrastructure in the Netherlands and its use, the process of requesting the installation of a new public charging pole, the planned expansion of public charging stations, the use of charging stations and EV ownership types. All the collected data was subsequently analysed and visualised (e.g., Image 8). These then served as a basis on which we defined new frames and modes, presented during one of the team meetings. During them, values and requirements started to emerge as well (e.g., "*the human should always be in charge and not lose any of her skills* and *"the user should have the feeling the car is always sufficiently charged"*, respectively). In addition, during the meeting, our main frame was established: "*If we help the user to slowly follow all of the car's suggestions and make them connected to their everyday life, the user will charge as they go*".

In order to operationalise this frame, three sub frames were generated in parallel: (1) *If the solution helps the user to gradually follow all of the car's suggestions, they will charge as they go*; (2) *If the solution turns their cars into a place where they will work on themselves, the users will charge as they go;* and (3) *If the solution gets attuned to the user's levels of risk, the user will trust the*

Image 8 The collected data visualised as a user journey for EV owners who cannot charge at home.



NO CHARGING at home

Install PlugShare

Find charge points

- distance
- usual places
- free

Plan

Find charger (+ plan B)

Plug in

Go back in 2 hours

Unplug

Walk back in

60% Enough for daily commute

Go to the gym

Find where to park (+ plan B)

Plug in

- open proprietary app
- scan QR

27% Leave

Update planning

Find a charger (+ plan B)

- only level 2 charger available
- take the slow charger

Do the groceries

Go out & charge (+ plan B)

Go to a public station near you (+ plan B)

- drive home
- come back in a few hours
- walk home
- plug in
- find an available station

68% Leave

Go to a rapid charger (+ plan B)

- have to wait in line
- do emails /games/ calls
- get bored

suggestions of the system. These three sub frames led to the simultaneous formulation of corresponding modes, intended actuations and the devising of low-fidelity visualisations of potential prototypes (Image 9). These were then presented during one of our team meetings and served as boundary objects between the client and me, thus, aiding the discussion on how to proceed with the project. After the second team meeting, a two-hour creative session was carried out with four expert designers to further define the previously explored frames. The session began by discussing the visualisations of all previously collected data. These visualisations then served as boundary objects among all four designers and supported us in discussing insights and generating potential sub frames. The session resulted in the selection of a new sub frame: *"If the solution intentionally makes users feel like novices, they will learn new charging behaviours"*. This frame served as the starting point for the third temporal bracket.

Image 9 Three low fidelity prototypes of an in-car AI system helping its users to tailor their charging around their routines.



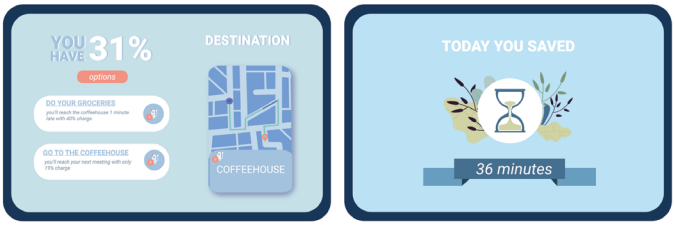
PROTOTYPE 1

A roadmap-concept for each day guiding users how to charge in order to make the most out of their day.



PROTOTYPE 2

An in-car system mindfulness guidance system EV onwes can use while charging their car.



PROTOTYPE 3

An in-car system pairing the charging of the car with chores EV owners have to do during the day.

Temporal bracket 3: Develop the concept

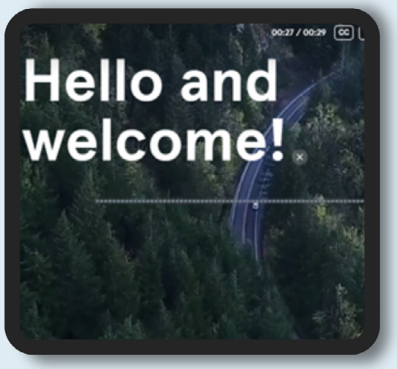
Finally, during the third temporal bracket, we underwent six iterations through the theoretical model, resulting in six actuated prototypes (Image 10). The first iteration commenced with data collection on the routines and charging preferences of the two participants we recruited. Both of them had just bought their new EVs, they did not yet have established charging routines and could not charge their EVs at home. The collected data was then analysed, visualised, and shared during our team meeting. During it, two new values emerged: "*Support users in their exploration of how the car and charging works"* (principal designer) and *"We just have to prepare the situation in a way that's somehow playful and make them feel like they're on an adventure."* (client). Consequently, we identified a more detailed sub frame (building onto the frame from the second temporal bracket): *"If owners take an active role in teaching the in-car system, they will follow its advice readily and be prepared for changes"*. To decide whether this frame can be used as the starting point for identification of a new mode, it was evaluated against the generated thus far requirements and values. For instance, our frame fulfilled the newly generated value of supporting exploration since teaching the car how they charge could allow them to explore and actively reflect on the manners in which they want to charge. Further, it also fulfils the generated requirements. For instance, one of our requirements was "*the user should feel like the car is always sufficiently charged*". We postulated that since the user teaches the car about their routines and comfort levels, they can also trust that it will know what "sufficiently charged" meant for them.

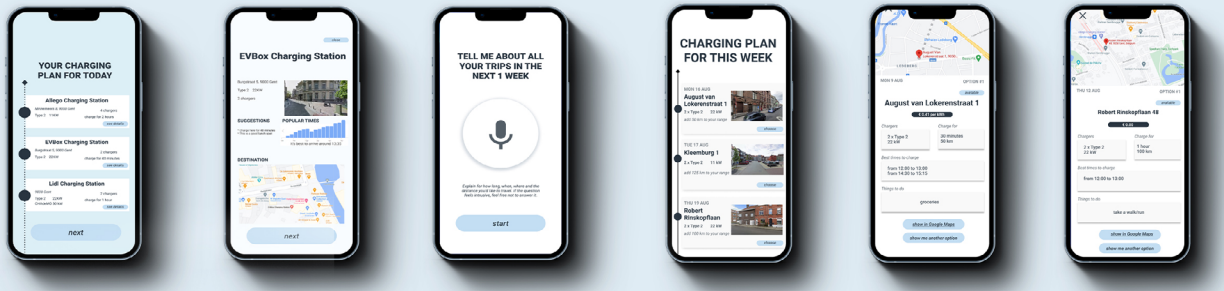
Once we evaluated the generated frame, a new mode *("Ask participants to show the system how they charge their car"),* intended actuation *("Walk the AI system along the steps they take to charge their car")* and initial version of the prototype were devised. These were then discussed during one of our team meetings. Then, together, we updated the existing mode into *"The system actively asks the user to show their current charging behaviours so that they can teach it what is important to them"*. The new mode led to a new intended actuation *("The users answer all of the prompts of the in-car system by filming their surroundings,*

dashboard, and their current routine on how they decide when and for how long they are going to charge.") and a prototype (see Image 10). The prototype was then quickly evaluated against the already existing requirements and values, similarly to the manner in which we evaluated the frame.

This prototype was then actuated by the two participants and resulted in multiple observed actuations such as: "*participants provided detailed videos for each of the prompts", "had difficulty in understanding how to use fast-charging stations", "went for a run while the car was charging",* and *"interrupted the charging*". Finally, these observed actuations resulted in several outcomes: e.g., "*participants felt anxious over the too many unknowns (e.g., charging costs, finding a charger, whether their attempts to teach the system were successful and what information the system paid attention to)".* These outcomes then triggered a reflection on the outcomes during our regular team meetings, resulting in the identification of new values and requirements, respectively: e.g., "*support people in doing healthy activities while charging"* and "*provide a clear overview on how much charging is going to cost"*. The observed actuations and outcomes served as the initial data input for the second iteration, where we identified other types of data that could support us in addressing the elicited unintended outcomes. An overview on the different instances each of the model's variables assumed during all six iterations can be found in the Appendix. Further, a detailed visualisation of the design activities that happened during the fourth iteration can be found in Figure 13.

After the sixth iteration through the model, we also identified a new sub frame: "*If the in-car system supports people to imagine how to plan for their long-term battery health, they will charge more as they go".* In addition, we also identified a new purpose: "*ensure EV owners do not occupy chargers they do not currently use*", accompanied by a new frame *"If the in-car system allows people to unplug each other's cars when they are done charging, then they will not occupy chargers they do not currently use".*





PROTOTYPE 1

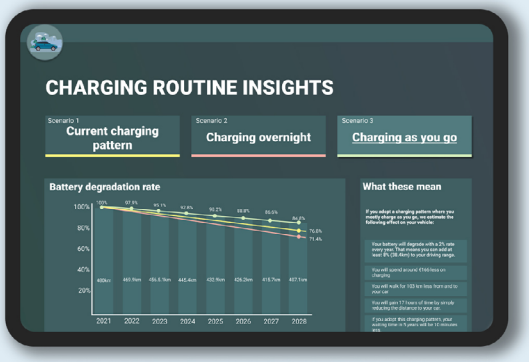
A low-fidelity mock-up of an in-car system containing multiple prompts delivered by a computer-generated voice guiding the users in explaining their routines.

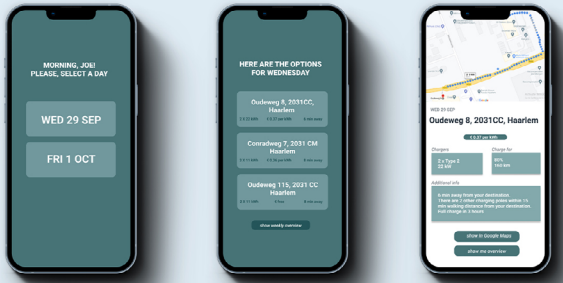
PROTOTYPE 2

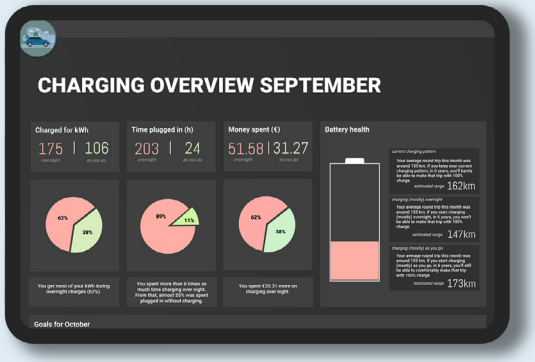
An app that shows users their charging plan for the week while also suggesting potential activities the user could do while the car is charging (e.g., doing the groceries, having lunch).

PROTOTYPE 3

An updated version of the app used in the previous iteration + updated buttons, clear charging plan (including photos of the charging stations) and a link to Google Maps directions on how to find the charging station.







PROTOTYPE 4

A mid-level fidelity dashboard providing three different scenarios that visualise and contextualise the potential battery degradation rate, the estimated waiting times and the money and time users spend on charging.

PROTOTYPE 5

An app building on the one introduced in Iteration 3 + an option to build their own plan, provide an overview on options for the day, provide information on distance from destination, type of charger, amount of time for full charge, walking route.

PROTOTYPE 6

A mid-level fidelity dashboard building on the one used in Iteration 4 + contextualisation of their potential battery degradation mode, the estimated waiting times and the money and time they spend on charging.

Image 10 An overview on all prototypes devised during the third temporal bracket Devise Prototypes

As explained in Chapter 5, the application of the theoretical model can result in the definition of new purposes. In fact, the model was devised so that it can support such results since, as already discussed, unlike products and services that are relatively finite, AI systems evolve in a rapid, hyper-personalised manner.

Occurrences of abductive reasoning

This section answers the second part of the first sub research question of this chapter: "*How and **where in the process do designers apply each of the three abductive reasoning types?".*** As it can be seen from the visual map (Figure 13), instances of abductive reasoning occurred during each of the three temporal brackets. During the first one, **"Understand context",** we mainly engaged in activities defined by explanatory abduction. Namely, we carried out activities labelled as *data* and *frame*. For instance, we collected data on the challenges EV owners experience when charging their cars. The patterns that emerged from the data collection led us to the insight that EV owners expect to use their EVs like a traditional car, which is reflected in the following quote from my design journal "*Maybe it's not about making charging stations the equivalent of gas stations, but instead, see where people's cars are idle and put strategically charging poles there so they can charge their cars while they're doing other stuff".* From this, an explanation (i.e., frame) was formulated "*If the solution facilitates a mindset change in EV owners, then we can ensure optimal charging behaviour").* We define a mode of action only once (which is activity that occurs during innovative abduction): "*guide drivers when, where and for how long to charge*". This exploration allowed us to agree upon a new, better-defined purpose. This new purpose also served as the starting point for explanatory abduction, which triggered the start of the next temporal bracket **"Devise initial concept".**

During the second bracket, we again, engaged in explanatory abduction activities such as collecting new data and formulating a new main frame that could explain the collected data. The identification of this main frame was then followed by the generation of several new sub frames, as already discussed in the previous section. These explanatory abduction activities

were complemented with activities indicative of innovative abduction such as defining new modes of action (e.g., "g*radually establish the new habits in drivers by providing rewards for charging optimally")* and intended actuations (e.g., *"the user aims to collect the suggested by the AI system reward points"*). The combination of our sub frames, modes and intended actuations then led to the creation of three different prototypes (Image 9). It also helped us to agree upon a new sub frame that could be used to achieve the purpose that emerged from the first temporal bracket. Namely: "*if the solution intentionally makes users feel like novices, they will learn new charging behaviours*".

During the third bracket **"Develop the concept",** we further detailed one of the concepts introduced during the second temporal bracket. As it can be seen from the visual map, activities representative of all three abductive types were observed from data collection (i.e., explanatory), through the devising of prototypes (i.e., innovative), to the actuation of these prototypes with the participants we recruited (i.e., manipulative). It is also the only temporal bracket during which we applied the theoretical model in its entirety and hence managed to address the behaviour-use interdependence that defines AI systems. Namely, we were able to observe how an early concept of the AI system will be actuated by its users. Then, based on the outcomes these actuations generated, we were able to observe a range of both intended and unintended outcomes that we addressed during the subsequent iteration through the model. Doing so allowed us to mitigate the negative unintended outcomes we observed and amplify the positive ones.

The three types of abduction

The second sub research question guiding the chapter was: *"How do the three types of abductive reasoning influence each other?".* In order to provide an answer to it, we first need to look into the bracket during which the three types of abduction occurred (i.e., Develop the concept) and more specifically, its fourth iteration. As with the other five iterations in this temporal bracket, all three types of abductive reasoning were used. The first iteration was focused on enabling the users to teach the in-car system, the second and third – to guide them on how to charge as they go. However, it was during the fourth

iteration that we realised that we need a new way to approach the problem at hand: charging as they go continued to be difficult for the participants and we continued to struggle to identify ways to help them adopt the new behaviour. In this fourth iteration, we explored a different part of the concept – an in-car system showcasing the user the impact their usage patterns have on their EV.

A detailed overview on all activities carried out during this iteration can be found in Figure 13. The figure is structured as follows. On the left-hand side, one can find the final formulation of each of the model's variables. On the right-hand side – the corresponding quotes and sketches that led to the definition of the variables on the left-hand side. The quotes come from (1) my design journal, (2) the weekly team meetings with each quote labelled as either "PD" (to denote the quotes from the principal designer) or as "C" (to denote the client's quotes); and (3) the user tests. In addition, some of the quotes have an additional label signifying whether they are addressing a variable (e.g., "MA" if they are indicative of the mode of action). These are complemented with visuals from the design journal and different versions of the prototype developed during Iteration 4 to showcase how it evolved. Furthermore, the activities that occurred during this iteration are color-coded depending on the type of abduction they represent. Finally, the activities that occurred during Iteration 4 are complemented with the outcomes, requirements and values that emerged from Iteration 3 and the frame of Iteration 5. The former served as the starting point for Iteration 4. The latter allowed us to address two of the Iteration 4's outcomes. Hence, it showcases the manner in which each iteration through the model influenced its subsequent iterations and how the concept was gradually developed and detailed.

As it can be seen from Figure 13, the outcome of each abduction type served as the starting point for another type of abduction. For instance, the outcomes of manipulative abduction (Iteration 3) served as the starting point for explanatory abduction (Iteration 4), during which gradually a new frame was defined in such a way that it can address (some of) the elicited outcomes and incorporate the requirements and values that emerged from the previous iteration. Namely, in our attempts to explain the outcome "*Participants find it difficult to believe they*

will be able to get to their destination on time if they only charge as they go. They need to feel prepared", we postulated that it could be because they lack the motivation to do so. In order to increase their motivation, we then suggested that we could use the Fogg Behavioural Model (Fogg, 2019) to achieve the needed level of motivation. This notion, however, was conflicting with a personal value I made explicit during Iteration 3: "*people's behaviours should change but without using nudging*". This act of reflection-in-action triggered the generation of two new values stemming from the desire not to nudge people: "*ensure each user has agency over the choices they make*" and "*ensure each user can make informed choices on how to change their behaviour*".

To address the newly generated values, we formulated the following frame: *"If the in-car system provides insights on the effect different charging behaviours have on the battery, users will be willing to do more charging as they go".* The frame was then evaluated against existing requirements and values (see Figure 13). Namely, the explicit decision to provide information rather than prescribe desired actions, we reasoned, allowed us to lessen the nudging effect of the solution.

Once this frame was selected, it did not change for the duration of the iteration, and served as the starting point for innovative abduction, where iteratively, we defined potential modes, intended actuations, and embedded these into a sketch which we iteratively developed into an initial wireframe of the prototype. These then allowed us to reflect on the direction we have chosen and led to explicit discussions of the values the client wanted to embed into the solution (e.g., *"We shouldn't tell people that they can improve their range because it will always degrade. We should tell them, your range will degrade, but here's a way to slow it down").*

This reflection-in action led to a second episode of innovative abduction during which we iteratively generated new modes and intended actuations, supported by numerous sketches and a prototype wireframe. Each one of these was influenced by the identified values. For instance, the subsequent mode we identified was based on presenting different types of scenarios users can use to imagine how they can slow down their battery degradation rate. Finally, once

we selected the final versions of the mode and intended actuation, we devised our new prototype. The newly devised prototype was then evaluated against our already existing requirements and values. As it can be seen from Figure 13, the prototype allowed as to address our requirements (e.g., "*provide easy access to information about the users' charging behaviours*") and values (e.g., "*and I think these choices people should make for themselves [the metrics they care about], not something we should push on them*") by providing users with different scenarios so that they can decide on their own which metrics they care about. The prototype was then actuated by our two participants.

The observed actuations resulted in the elicitation of four unintended outcomes and one intended. These can be found in Figure 13 where the summarised observed actuations and outcomes are accompanied with some of the quotes elicited during the user test. The outcomes then served as a point of reflection between the client and me, resulting in new values and requirements. We also evaluated the outcomes based on whether they were intended or unintended, positive, or negative. Subsequently, two of the unintended elicited outcomes, which we considered to be negative (i.e., *"the users continue to feel unprepared enough for their commutes if they charge only as they go"* and *"deciding to use their car more often instead of their bike in order to charge as they go"*) were addressed by the frame of Iteration 5, as one of the explanations we generated during our subsequent team meeting was that users feel unprepared and use their cars instead of their bikes because they do not trust the system. Hence, we hypothesised that if users are involved with the making of the suggestions, they will better understand the rationale behind them.

DETAILED OVERVIEW ON ITERATION 4

- explanatory abduction
- innovative abduction
- manipulative abduction
- no abduction

ITERATION 3

outcome Lack of trust in the charging suggestions generated by the system.

outcome Misalignment between suggested by the system activities and the ones they usually perform.

outcome Participants don't feel prepared enough if they don't chatge overnight.

outcome Participants try to increase the number of charge-as-you-go times.

team Provide easy access to information about the users' charging behaviours. C R

team Support people in feeling prepared about their journey. C R

team People should feel in control of the data they share. PD V

team People's behaviours should change but without using nudging. PD V

team "It will all come down to convenience because at some point you might want to charge overnight, but it's important not to let this become the dominant way of charging." C V

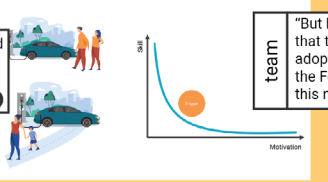
ITERATION 4



journal Based on the way they actuated the previous prototype, we can see that:
1. It continues to be difficult to charge as they go.
2. They always try to get as much charge as possible overnight.

team "The predominant theme is that they need to feel prepared. No matter what you tell them, or how you explain it to them, they will always go and charge overnight." C

team "They lack in motivation, so what I should present to them is something that will help them to increase their motivation to perform the action." PD F



team "But before that, I need to convince them that this is the behaviour they need to adopt. Maybe I can use principles from the Fogg Behavioural Model to achieve this new behaviour?" PD F

team Ensure each user has agency over the choices they make. PD V

team Ensure each user can make informed choices on how to change their behaviour. PD V

journal What if the prototype shows the users what it has learned from their behaviour so far? It will help with figuring out the other things they'd like the system to learn from them. F

team "If we provide people with different options and insights on what their actions mean for the range of their car, we can support them in making an informed choice." F C

journal It can also establish a sense of trust within the user since they can see that the system is paying attention to their preferences. F

journal What if we show users how 100% charge and overnight charging are bad for the car, because none of our participants knows that. F

frame If the in-car system provides insights on the effect different charging behaviours have on the battery, users will be willing to do more charging as they go.

team Provide easy access to information about the users' charging behaviours. C R

team Ensure each user can make informed choices on how to change their behaviour. PD V

team Ensure each user has agency over the choices they make. PD V

Figure 13 a snapshot of the activities carried out during Iteration 4

journal

The system can show insights about:
1. The way they look for a charging spot.
2. The way they build up their back up list of charging stations.
3. How fast chargers are bad for the car. MA

team

"I'll really focus on guiding them and explaining on why they should charge as they go." PD MA

journal

It can show
1. How much time they wasted going to and from the charger.
2. How much of the resell value they lost due to battery degradation.
3. By how much the availability of free charging poles is going to decline. MA

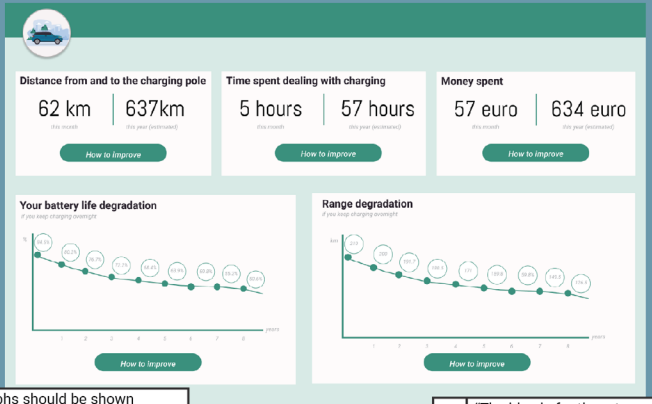
team

"So if you show them clearly "you're not supposed to do that because it will affect your battery", then they'll know. And then we can use that as a jump for them to start charging as they go." PD MA



journal

"They need to understand that it's not good for their battery to fully charge and this could lead to 3% decline in range every year which seems like something small. But once we show them how much their range will actually degrade, cumulatively, they will make the decision to adopt more battery-conscious practices." IA



team

"The two graphs should be shown together - that way people can have a better overview." C MA

How to improve

team

"The idea is for them to see the dashboard and kind of snap out of the idea of charging over-night." C IA

team

"And I think these choices people should make for themselves (the metrics they care about), not something we should push on them." C V

team

"We shouldn't tell people that they can improve their range because it will always degrade. We should tell them, your range will degrade, but here's a way to slow it down." C V

team

"It feels like it's more work but if you show them numbers and put them in context, it will be different." C MA

team

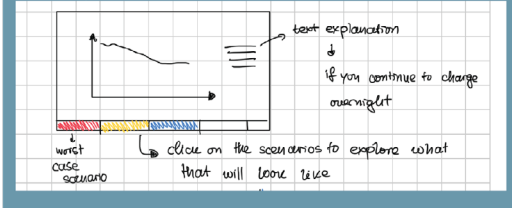
"Maybe they can click on different scenarios and see how they can improve their battery health." C MA

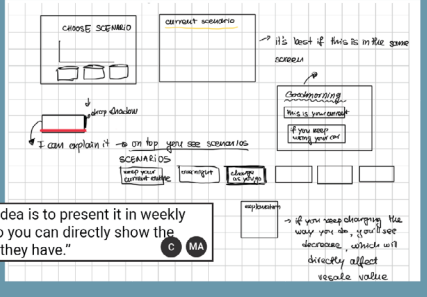
team

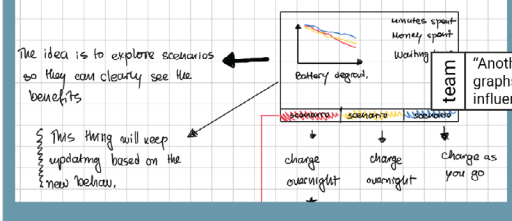
"Maybe it's an idea to present the average degradation for other users of the same vehicle so you can see how well you perform compared to the others – better or worse." C MA

team

"Maybe you can try and have more negatively and then more positively focused dashboard and see what the responses are and see what kind of other ideas emerge from that." C MA







team

"Another idea is to present it in weekly graphs, so you can directly show the influence they have." C MA

mode

Show three scenarios to the users on how their charging behaviour affects their battery degradation rate, their monetary and time spending, as well as the rate with which their waiting time will increase if every EV owner keeps charging overnight.

intended actuation

Users interact with the dashboard and its three scenarios. They decide to incorporate more charging as they go.

team

"They can choose: this is the information I want to see/I don't want to see and you can easily access that information when you want to but it's not in your face all the time." C IA

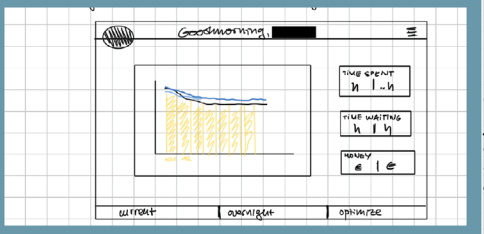
journal

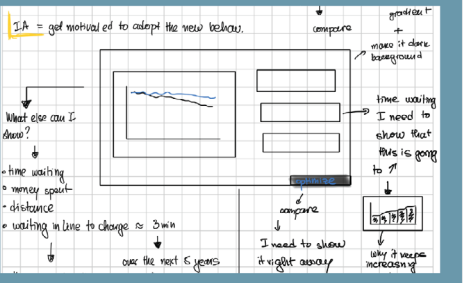
The system should convey these points:
1. Degrading battery life
2. Degrading range
3. Estimated wating times (to charge)
4. Walking distance
5. Money spent
6. Time spent charging

MA

journal

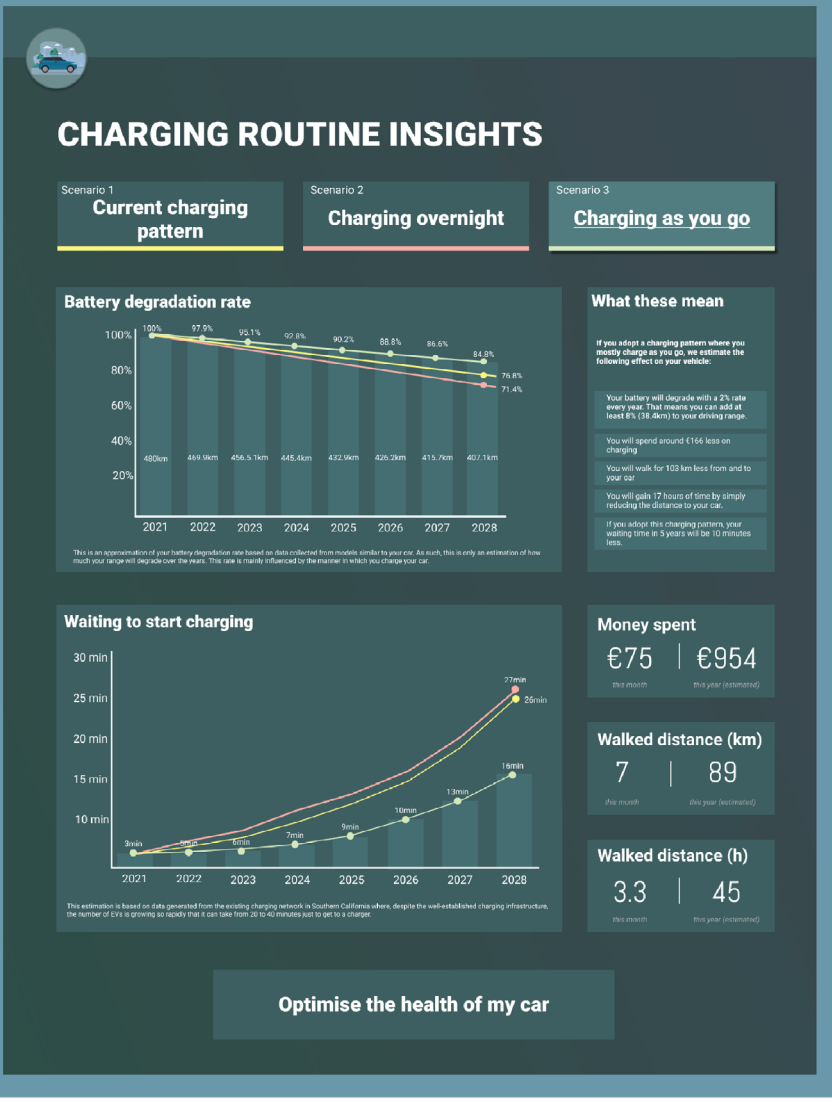
Make use of "what if scenarios": e.g., what if I charge overnight 50% of the times? MA





team

"So they can see their insights and they can play with them to see what will be the best option for them." PD IA



prototype

A mid-level fidelity dashboard providing three different scenarios that visualise and contextualise the potential battery degradation rate, the estimated waiting times and the money and time users spend on charging.

team "And I think these choices people should make for themselves (the metrics they care about), not something we should push on them." C V

team People should feel in control of the data they share. PD V

team Ensure each user has agency over the choices they make. PD V

team "We shouldn't tell people that they can improve their range because it will always degrade. We should tell them, your range will degrade, but here's a way to slow it down." C V

team "It will all come down to convenience because at some point you might want to charge overnight, but it's important not to let this become the dominant way of charging." C V

team Ensure each user can make informed choices on how to change their behaviour. PD V

team Provide easy access to information about the users' charging behaviours. C R

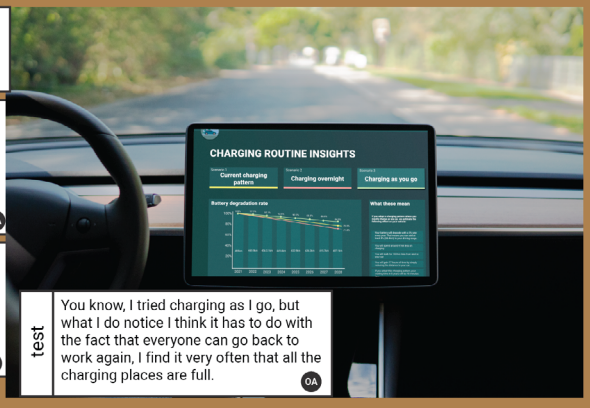
observed actuation

Both users went through the three different scenarios, barely paid attention to the estimated waiting times, and spent most time on coming up with different explanations about the factors that influence the battery degradation rate. User 2 changed behaviour drastically (i.e., provided immediate unrestricted access to their calendar).

test Goes through the dashboard and at first doesn't consider like an investment, but then sees the degradation and keeps saying "This is scary". OA

test "Ah, yes, my car is plugged-in for a long time, but I think I don't want to do that. The time it's plugged in is mostly due to my schedule, not necessarily whether the car is charged or not, so I don't really want to reduce that time. I don't find it a nuisance." OA

test "I'm not taking 15 min extra for every appointment that I have in order to find available charger, then I'll probably just drive until the battery is completely empty and fast-charge for the last part and then drive home and then charge it overnight." OA



test You know, I tried charging as I go, but what I do notice I think it has to do with the fact that everyone can go back to work again, I find it very often that all the charging places are full. OA

test "Seeing this degradation tendency, I wonder.... I now have battery warranty for 6 years saying that my range shouldn't get lower than 180 km. So if I charge overnight and degrade my battery to a range under 180 km in the next 3 years, would that mean I can get a brand new battery for free?" O

test "This is scary. What if I give the system full access to all my data? Would that help it to give me more accurate estimations of battery degradation?" O

test "You know, for the past few weeks I've been trying to charge as I go instead of charging overnight. But it's difficult, because instead of using my bike, I'm now using my car much more to do my chores." O

outcome Speculations about the potential ways in which the car warranty can be tricked by intentionally degrading the battery faster.

outcome Deciding to use their car more often instead of their bike in order to charge as they go.

outcome The users continue to feel unprepared enough for their commutes if they charge only as they go.

outcome The user suddenly provides full access to their data.

outcome Increased number of times participants charge as they go.

team Include only concepts for which the users already have a frame of reference. PD R

team Make sure that the battery degradation is not shown in a negative light. C V

team It shouldn't be that easy for the users to get scared and give the app full access to their agenda and data. PD V

ITERATION 5

frame If the in-car system involves the user in the creation of their charging plan, they will feel prepared and charge as they go.

Non-abductive occurrences

The third sub research question of the chapter was: *"Where in the process do designers make use of non-abductive activities?".* As we can see from the visual map and the detailed overview on Iteration 4, numerous non-abductive instances occurred throughout all three temporal brackets. For instance, during the first temporal bracket, we regularly carried out data analysis (as seen in Image 7). The same non-abductive activities can be seen in the second temporal bracket, too. During it we also observed new types of activities such as the generation of new values and requirements from both the client and me. For instance, during the second team meeting for the bracket, one of the requirements the client formulated was: "g*ive the user the perception that they are in control"* and the following value was formulated by me: *"EVs become an extension to the person rather than something that demands too much time of them"*.

We observed the same type of non-abductive activities during the third temporal bracket as well. They emerged in a pattern that repeated during each of the six iterations through the theoretical model. First, new values emerged in each iteration after initial frames and prototype were devised. These in turn influenced the manner in which every subsequent variable of the model was formulated. Second, the results of all three abduction types (i.e., frame, prototype, and outcomes) were always evaluated against the already existing requirements and values, as discussed in the previous section.

We also observed activities similar to the ones discussed in Chapter 6. After the prototype was actuated and the outcomes were observed, they were then evaluated against the existing requirements and values. The outcomes also served as a basis on which new requirements and values were generated that would guide the subsequent iteration through the theoretical model.

Discussion

In the process of designing a concept for an in-car AI system, we went through three temporal brackets: (1**) Understand context**, (2) **Devise initial concept**,

and (3) **Develop the concept**. During the first two brackets we made use of only explanatory and innovative abduction. These were used to first, understand the context for which we were to design and then design three concepts. One of these concepts (Image 11) became the starting point from which prototypes were built in the third temporal bracket (e.g., using chores like doing groceries as an opportunity to charge as you go). As such, this concept can be equated to what Roozenburg (1993) terms a "principal solution": "an idealized representation (a scheme) of the structure of a system, that defines those characteristics of the system that are essential for its functioning." (p. 12). These observations are in line with Design theories where explanatory and innovative abduction allow us to devise frames and new solutions to a problematic situation (e.g., Roozenburg (1993), Dorst (2011), Dong and MacDonald (2016) and Kroll and Koskela (2016)).

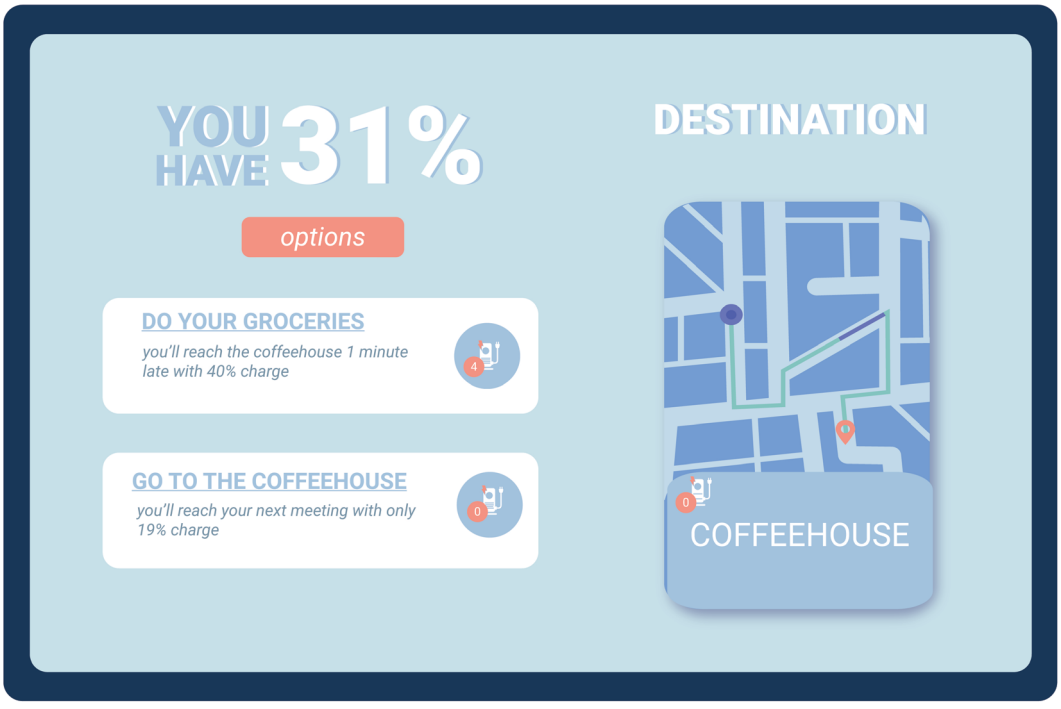


Image 11 Temporal bracket 2, Prototype 3

However, as we saw in Chapter 6, when designing ever-evolving AI systems, we need to extend Design theories so that we can address the behaviour-use interdependence these systems exhibit. This assertion became the basis for the chapter's main research question: (i.e., **"How can Design theories be extended to support the early simulation of AI systems' behaviour-use interdependence?"**)**.** As our results show, we were able to simulate the behaviour-use interdependence only during the third temporal bracket. It was then that we furthered developed the initial concept sketches into a prototype that can be actuated by our users. Consequently, it enabled us to actively develop and adapt the concept to address the observed actuations and outcomes we were eliciting.

The manner in which we applied the extended theoretical model was largely similar to the example reported in Chapter 6. However, we also saw that the transition between each type of abductive reasoning was supported by activities such as visualisations, and the generation of requirements and values. We discuss these in detail in the remainder of this section.

The role of visuals

We made use of two kinds of visual representations throughout the three temporal brackets: data visualisations and sketches. In order to create each one of them, we utilised different types of reasoning. On the one hand, when generating data visualisations (e.g., Image 7 and 8), we applied inductive reasoning[2]. Namely, we had the data (i.e., result) and the context in which the data was collected (i.e., case) which allowed us to infer and visualise a pattern (i.e., rule). These visuals then served as a means to communicate the collected insights with the rest of the team. In turn, they served as a basis for the creation of possible explanations for the collected data and observations. They also allowed us to elicit new values and formulate new frames.

On the other hand, we also made use of sketches, which were gradually increasing in complexity and level of detail visuals (e.g., Image 12). These sketches influenced the way we formulated the variables on the model's

2 See Chapter 3 and Table 8 for an overview on how induction is carried out.

second row. Namely, the moment we identified our first mode, we immediately started sketching how this mode could manifest itself. This allowed us to define the intended actuation and transform the initial sketch into a prototype. Consequently, these sketches served as initial versions of our prototype.
The type of reasoning that enabled us to do so is manipulative abduction. As discussed in Chapter 4, this abduction type is present when one uses "communicable accounts of new experiences to integrate them into previously existing systems of experimental and linguistic (theoretical) practices" (Magnani, 2004, p. 229). Therefore, the sketches we generated allowed us to move from our operational interpretations of the domain (e.g., due to the data we collected) to the design and integration of new experiences into the existing context. Consequently, we can hypothesise that manipulative abduction enables us to carry out innovative abduction. This contention, of course, needs to be further researched.

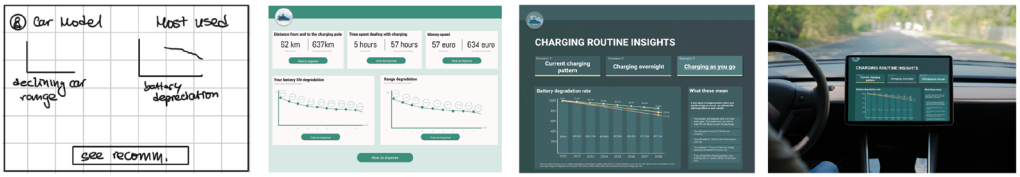


Image 12 An overview on the different types of solution visualisations increasing in level of detail and complexity.

Use of requirements and values

In Chapter 6, we postulated that the role values and requirements play when applying the theoretical model is akin to what Schön (1983) termed as reflection-in-action. While this statement was correct in this case, we also saw the pivotal role both values and requirements had in facilitating the transition from one abduction type to another. In the remainder of this section, we first elaborate upon the reasoning types that allow us to generate new values and

requirements and then use them as a means of evaluation. We then briefly discuss the pattern in which both of them emerge.

Reasoning type

In order to identify the used types of reasoning, we adopt the reasoning structure Peirce (CP 2.622) suggested of rule, case, and result[3]. According to him, in abduction, we reason from result and rule to a case; in induction – from result and case to a rule; and in deduction – from a rule and a case to the result (see Chapter 3). An example on how each reasoning type was used during Iteration 4 can be found in Table 8.

Table 8 Examples of how different types of reasoning were applied during the design project.

Reasoning	From	To	Rule	Case	Result
Abduction	Result + Rule	Case	People are likely to adopt a new behaviour if they have the skill and motivation to do so (Fogg, 2019)	If we motivate people well, they will adopt the new behaviour (initial frame, Iteration 4).	People have the skill to perform the new behaviour (outcome, Iteration 3).
Induction	Result + Case	Rule	Ensure people can make informed choices (value Iteration 4).	"We need to convince people to change their behaviour, but I'm also morally opposed to the idea of nudging and exploiting people's cognitive biases" – principal designer, Iteration 4.	Users find it very difficult to start charging as they go (outcome, Iteration 3).

3 The example Peirce (CP 2.622) used to explain rule, case and result was the following: "Rule: All the beans in the bag were white; Case: These beans were in the bag; Result: These beans are white."

Deduction	Rule + Case	Result	People find it difficult to conceptualise numbers in their head (fact).	The in-car system provides insights on the effect different charging behaviours have on the battery (final frame, Iteration 4)	Ensure easy access to information (requirement, Iteration 4).

Using this structure of a rule, case, and result, we can conclude that the generation of new requirements and values is indicative of inductive reasoning. Let us take a quote from Iteration 4's first team meeting:

> *"***We can see that our users are still refusing to charge only as they go [result]…** *This [the idea of increasing motivation] is actually something I'm really struggling with right now –* **we need to convince people to change their behaviour, but I'm also morally opposed to the idea of nudging and exploiting people's cognitive biases to make them establish a new routine [case]**. *So, I'm trying to figure out where the balance is because I really think charging as you go is a more sustainable behaviour that's also better for your car battery, but also this should be done by keeping the* **agency of the human [rule]** *and her* **ability to make informed choices [rule]**.*".*

As the quote shows, in order to formulate new values, we reason from the result and our specific case to two new rules which we added to our list of values the solution had to fulfil.

Second, we used the already existing requirements and values as a means of evaluating the results of each type of abduction during the third temporal bracket (i.e., frame, prototype, and outcomes). In order to evaluate whether the result of each abduction type fulfils our requirements and values, we use

deductive reasoning. Let us take a quote from the team meeting we had once the final frame of Iteration 4 was formulated:

> *"***So, this [giving users insights on how their behaviour affects their battery] will clearly show them that they need to stop charging overnight [case].** *Because* **people are generally not good at making sense out of numbers [rule]** *– you read stuff and you understand that that's [charging overnight] not good for your battery and the 3% [battery degradation] seems like something small. But once you see how much it will actually degrade, cumulatively, I think it's going to be striking for them…* **So, we'll make it very easy for them to get the needed information to make their own choices [result]***".*

As the quote shows, in order to evaluate whether our frame fulfils the existing requirements and values, we reasoned from case and rule to a result.

Emergence pattern

Finally, we discuss the patterns in which requirements and values occurred. First, neither requirements nor values emerged during the first temporal bracket. This is normal since, as its name[4] suggests, during this temporal bracket we were focused on understanding the problem at hand. Second, both new requirements and values emerged after activities indicative of explanatory and innovative abduction during the second temporal bracket. Then, in the third temporal bracket both new values and requirements emerged. However, while new values were identified during each abduction type (similar to Temporal bracket 2), new requirements were identified only after the outcomes of the actuated prototypes have been elicited (i.e., as a result of manipulative abduction). Namely, we saw that when no manipulative abduction was used, both values and requirements emerged as a result of explanatory and innovative abduction. However, once manipulative abduction was present, the emergence pattern changed. Therefore, further research in different

4 The name of the first temporal bracket is "Understand context".

contexts and with different types of developers is needed to discern whether manipulative abduction could have such influence.

Another interesting avenue for further research could be to view the emergence pattern through the prism of co-evolution (Dorst & Cross, 2001). For instance, the emergence of requirements and values can be mapped to the different types of transition from problem to solution space (e.g., see Cash & Gonçalves, 2017). Doing so would allow for a different point of view on the process of designing ever-evolving systems. However, first, a coding framework needs to be defined that combines the variables of the theoretical model with the different types of co-evolutionary transitions.

Conclusion

The research presented in this chapter supported us in answering the following main research question: **"How can Design theories be extended to support the early simulation of AI systems' behaviour-use interdependence?".** The application of the model to the case of designing an in-car AI system in collaboration with a large multinational automaker supported the assertion we made in Chapter 6 that to ***support the early simulation of AI systems' behaviour-use interdependence***, we need to apply the theoretical model in its entirety. If we are to apply only explanatory and innovative abduction (as existing Design theories prescribe) we can only generate what Roozenburg (1993) terms "principal solution". While the generation of such is indispensable for the design of an AI system, it does not manage to fully capture its behaviour-use interdependence which can result in multiple unintended and unanticipated consequences.

The conceptual developments we suggested in this chapter could also offer a nuanced perspective on the intricate mechanisms of reasoning within Design. The empirical insights enabled us to map interactions among the three types of abduction: explanatory, innovative, and manipulative, as well as their relationships with induction and deduction. As already explained in Chapter 6's conclusion, the exact relationships among these reasoning types is still largely unexplored. As such, the model introduces a conceptual foundation for

understanding the relationships among different types of reasoning. Although these insights are inherently suggestive and conceptual in nature, they could provide a starting point for in-depth explorations into the interplay and triggers of diverse types of reasoning.

In addition, our empirical insights and subsequent conceptual developments address the role of visual representations in Design, too. Visual representations have long been seen as integral in both Design Cognition and Design Methodology (e.g., see the work of Goldschmidt, 2003; Gonçalves & Cash, 2021). However, to our knowledge, the role they play when enacting different types of reasoning, especially abductive one, has not been explicitly addressed in literature. The work presented in the Discussion section of the chapter suggests that visual representations can facilitate both induction and manipulative abduction. This offers a unique lens through which to examine how different types of representations influence the reasoning patterns employed during the design process.

Finally, this case study also supported our contention that in order to address the interdependence, the Design theories we used in the theoretical model should be extended. There are five concrete insights that can guide such extension: (1) explanatory abduction is usually followed by innovative abduction; (2) the inductive generation of new values and requirements informs the formulation of every variable of the model; (3) visuals generated as a result of inductive reasoning (e.g., data visualisations) facilitate explanatory abduction; (4) the deductive evaluation of each row's result against requirements and values supports the move from one abduction type to another; and (5) manipulative abduction plays a facilitative role while carrying out innovative abduction.

Collectively, these insights enhance our comprehension of the intricate reasoning patterns underpinning the design of evolving solutions. They also shed light on how fundamental Design concepts such as visual representations, prototypes, requirements, and values are connected to these reasoning patterns. As such, these insights become the basis for the next chapter, which serves as a conclusion to this doctoral dissertation. In it we present the

final theoretical model that can provide an answer to the dissertation's main research question: *"How can a theoretical model be designed that supports the early simulation of AI systems' behaviour-use interdependence by utilising Design theories?"*.

CHAPTER 8

FINAL THEORETICAL MODEL

The theoretical model we introduced in Chapter 5 served as an initial answer to our main research question: *"How can a theoretical model be designed that supports the early simulation of AI systems' behaviour-use interdependence by utilising Design theories*?". In Part III we took this initial model as a starting point. Then, in Chapter 6, we used an example from my teaching practice as a way to render the initial model into the context of designing an AI system. We contended that Design theories, devised for the design of products and services, can only aid the generation of what Roozenburg (1993) terms "principal solution". If we are to ***support the early simulation of AI systems' behaviour-use interdependence***, we need to extend these theories. We further explored this assertion in Chapter 7 by presenting a 5-month long case study carried out in collaboration with a large automaker. The study followed the conceptual design stages of an in-car AI system and suggested manners in which the theories and the model could be extended.

We identified the following insights that can guide the needed extension: (1) a new variable needs to be added – intended actuation; (2) each row of the theoretical model is representative of a distinct abduction type; (3) innovative abduction usually follows explanatory abduction; (4) the inductive generation

of new values and requirements informs the manner in which all model's variables are formulated; (5) the deductive evaluation of each row's result against the requirements and values triggers the move from one abduction type to another; (6) visuals generated as a result of inductive reasoning (e.g., data visualisations) facilitate explanatory abduction; and (7) manipulative abduction plays a facilitative role while carrying out innovative abduction. This theory-building exploration (Cash, 2018; see Figure 14) serves as the basis on which we build the final version of the theoretical model.

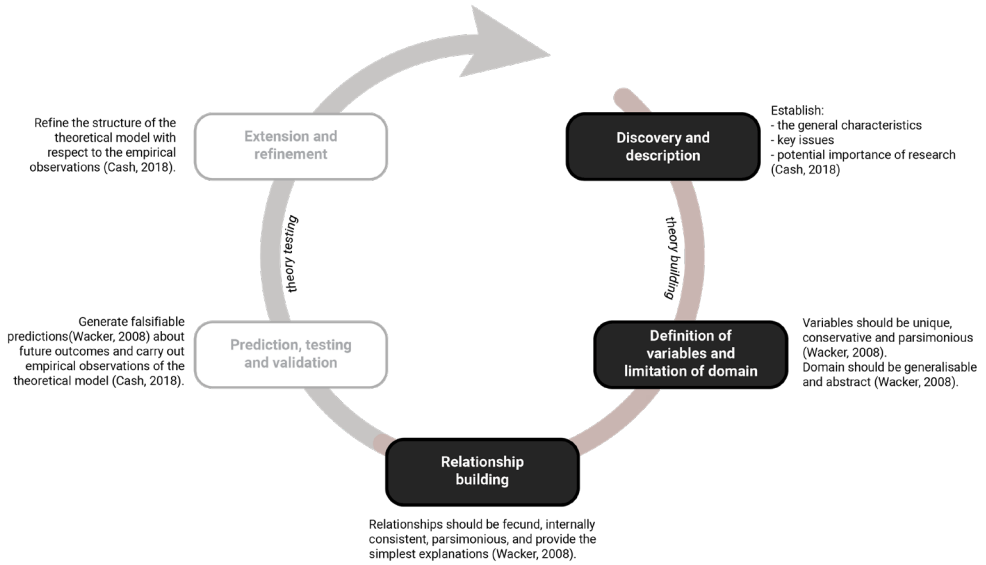


Figure 14 An overview on the stages of the theory building/testing cycle this dissertation addresses.

This conclusion chapter is structured as follows. First, we present the theoretical model by identifying its (1) domain (and the corresponding limitations), (2) the final list of variables and their definitions, (3) the relationships among the identified variables, and (4) predictions about the theoretical model that can serve as the starting point for its testing. These four aspects were chosen in accordance with the criteria of "good theory"

Wacker (1998, 2008) put forward, for "without any one of these properties, any conjecture, inference, supposition, hypothesis, or set of hypotheses, is just not a theory" (Wacker, 2008, p. 7). Finally, the chapter is concluded with a discussion on the implications of the theoretical model for Design(ers).

Domain

According to Wacker (2008), a domain is "the exact setting or circumstances where the theory can be applied" (p. 363). Taking this definition as a starting point, we can define the general domain of the theoretical model as *the early simulation of AI systems' behaviour-use interdependence during the conceptual design stage of their development*. AI systems are software programmes defined by their continuously self-learning nature that "can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with" (Artificial Intelligence Act, 2021). Further, they are defined by their complex functional processes, high dependency on human-generated data and continuous process of learning and adapting the system's behaviour to fit to the manner in which humans use it (see Chapter 1).

In Part III, we also imposed limitations onto the domain by contending that the early simulation of behaviour-use can happen only when a principal solution has already been designed. As such, the theoretical model can be applied in its entirety only when this condition has been met. Doing so, we argued, is important if we are to design the behaviour of the AI system in such a way that it will trigger its intended use. Further, it could also support us in ensuring the AI will continue behaving in an intended way even after different types of unexpected uses are performed by humans or other AI systems.

Variables

The majority of the theoretical model's variables were identified in Part II of the dissertation where we surveyed Design theories that can help us address the behaviour-use interdependence characteristic of our domain. Then, in Chapters

6 and 7 new variables emerged as well. These activities resulted in 10 variables that define the theoretical model (see Table 9).

Table 9 An overview on the definitions of the theoretical model's variables.

Variable	Definition
purpose	The goal of an AI system.
data	Records produced by experiments and measurements that serve as evidence for the existence or features of a phenomenon.
frame	A cognitive act of looking at a problem situation from a specific viewpoint that informs how the problem can be solved.
mode of action	(Functional) behaviour of the AI system in response to influences exerted on it from its environment.
intended actuation	The intended by design action of a user that allows the AI system to function and be "connected" to its immediate environment.
prototype	An externalised representation of a potential AI system.
observed actuation	The actions a user performs in order to allow the AI system to function and be "connected" to its immediate environment.
outcomes	The intended and unintended results of the AI system's actuation.
requirements	The performance specification of the AI system that limits the range of acceptable solutions.
values	An enduring belief that a specific mode of conduct or end-state of existence is personally or socially preferable to an opposite or converse mode of conduct or end-state of existence.

Relationships

The domain and variables serve as the foundation on which we can detail the relationships that define the theoretical model. These relationships were informed by Design theories (i.e., see Part II) and the proposed extensions discussed in Chapters 6 and 7. The combination of these leads us to the formulaic description of the relationships among the theoretical model's variables found in Figure 15.

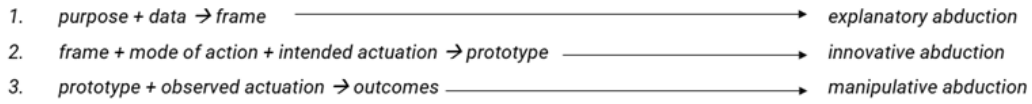


Figure 15 Formulaic representation of the theoretical model

As already discussed in Chapter 5, the model's relationships are defined through logical terms using the format typically used in Design Theory (e.g.,

Roozenburg, 1993; Dorst, 2011). However, in Chapters 6 and 7, we saw that different design representations (i.e., data visualisations and sketches), requirements and values also play an important role in how the variables are defined and relate to each other. Therefore, we also introduce a schematic representation to illustrate the variables' relationships.

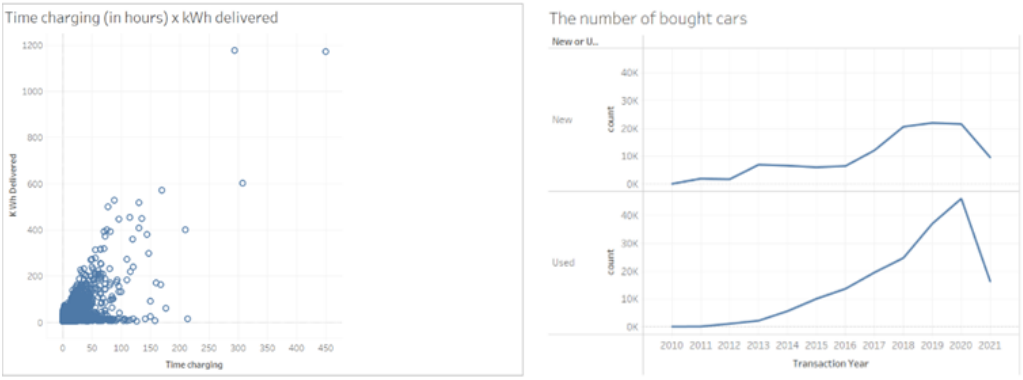


Image 13 Examples of data visualisations.

As already discussed, when simulating the behaviour-use interdependence of an AI system, the developer starts with a loosely defined purpose. Once this purpose has been selected, it defines the boundaries within which the future concept should operate (i.e., the concept should be designed in such a way so that it will fulfil the purpose). This purpose then serves as the starting point based on which developers commence collecting data that they try to explain by generating a frame (i.e., *purpose + data → frame*). This move is indicative of explanatory abduction (Figure 16). In order to move from data to a frame, however, they make use of different visual representations (e.g., data visualisations, user journeys) that allow them to analyse the data they have collected and make sense out of it (e.g., Image 13). Hence, it facilitates the process of generating an explanation for the collected data. The making of such representations is indicative of induction (see Chapter 7).

In order to decide whether the resulting frame is satisficing, the developers deductively evaluate it against the values and requirements their AI system should fulfil. These requirements and values have been introduced by both the developers, their client, and relevant stakeholders. Assigning a variable that

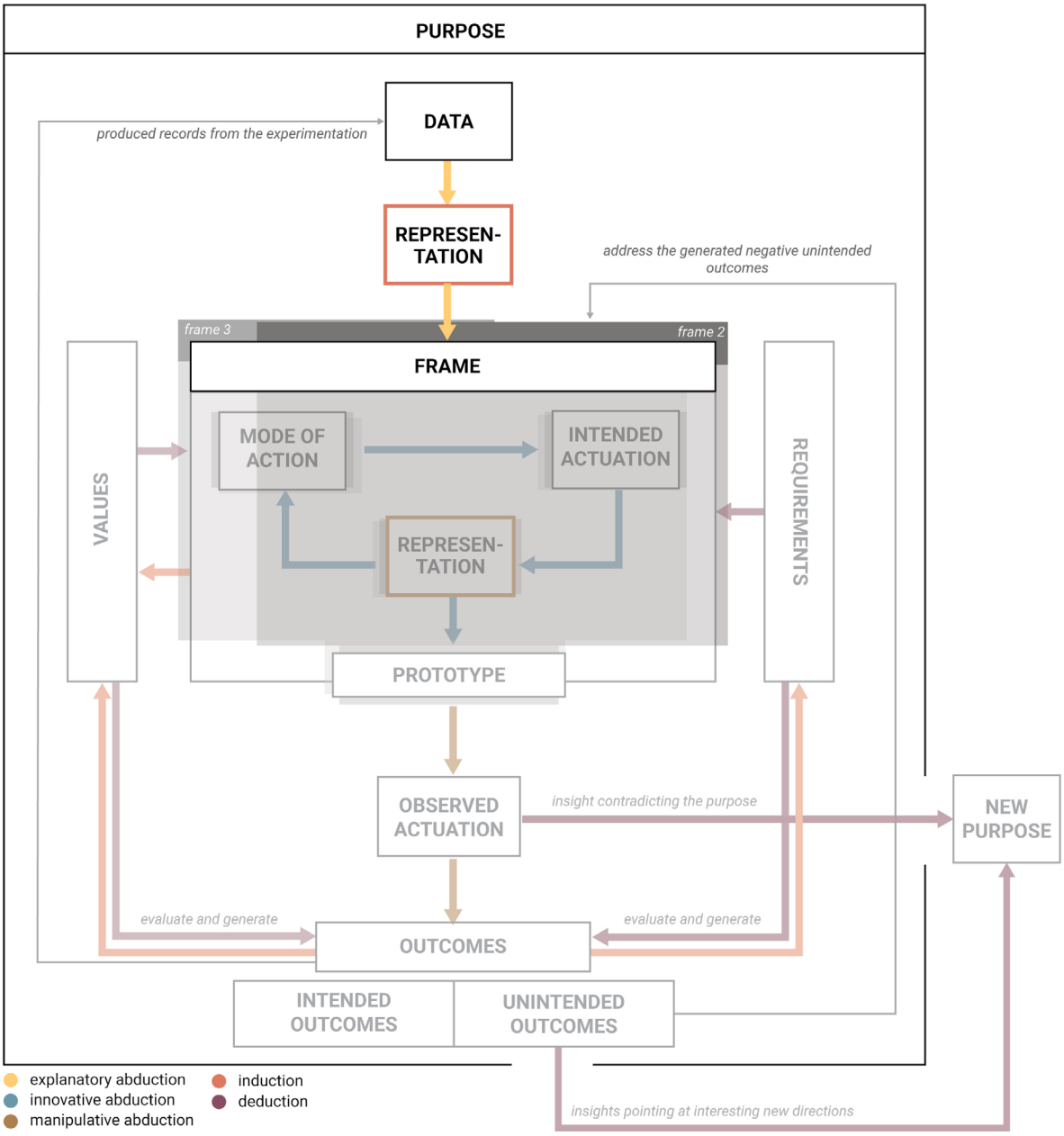


Figure 16 A schematic representation of the model with its first-row variables highlighted.

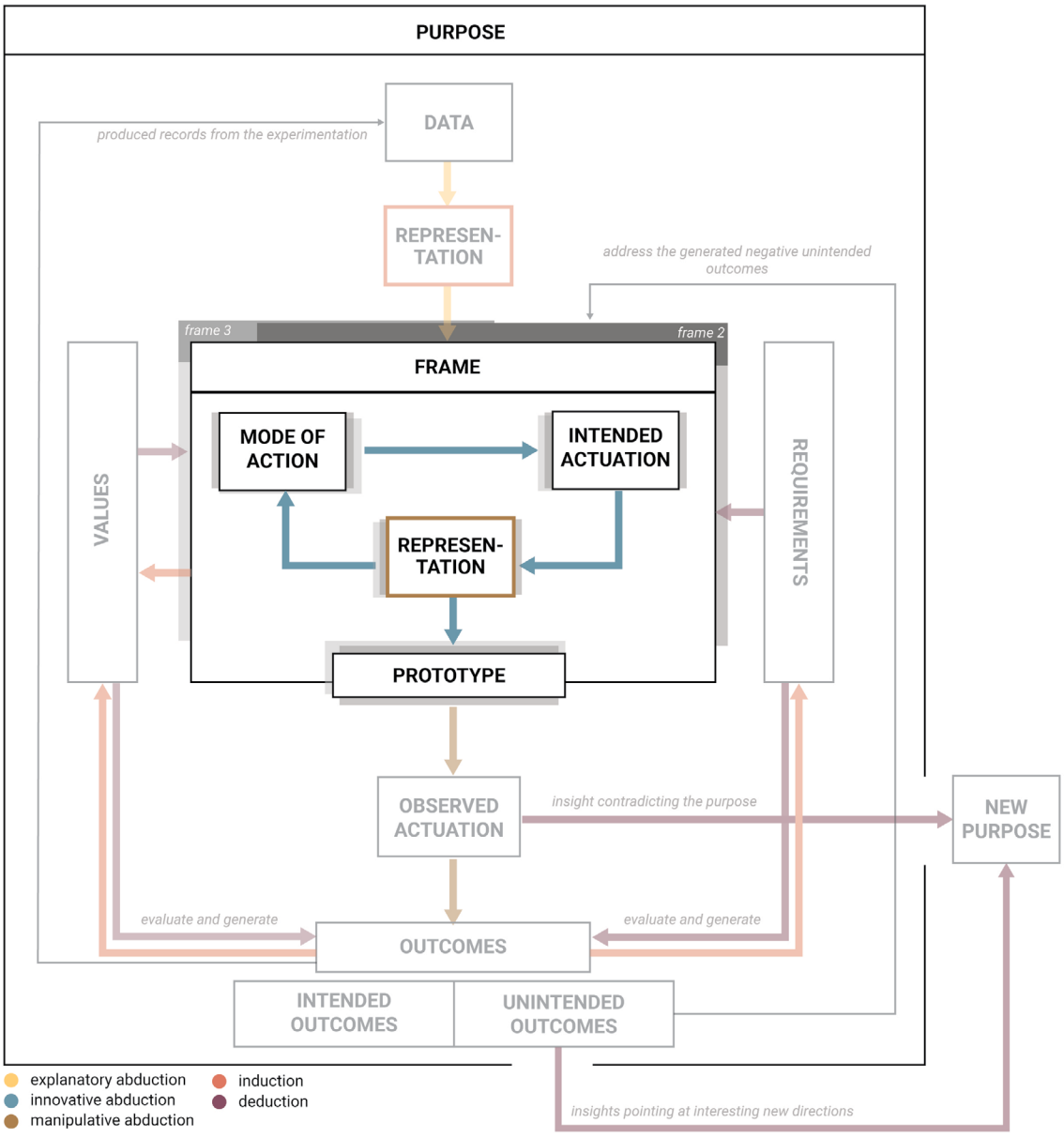


Figure 17 A schematic representation of the model with its second-row variables highlighted.

can make the otherwise implicit embedding of personal values explicit could aid us in having a clear overview on the types of values the development team decides to embed into an AI system. Doing so is important if we are to audit the AI system and openly discuss the potential biases that might stem from the embedded values. If the generated frame cannot satisfice the requirements and values, a new frame needs to be created that can. In addition, the reflection on the defined frames that do not fulfil the existing requirements and values leads to the inductive generation of new values[1].

Once a satisficing frame is created, the developers then commence with the application of innovative abduction. Namely, the frame provides them with a boundary within which they start defining a mode of action, which then leads to intended actuation and a prototype (i.e., *frame + mode of action + intended actuation →prototype*) (Figure 17). Here, developers use manipulative abduction to create a design representation (i.e., sketch). The representation is used to support the developers in making the transition from mode to intended actuation to prototype. Namely, the moment they identify a mode, they start sketching what this mode could look like. This allows them to identify a potential intended actuation, which also enriches their representation. During this iterative process new values their AI system should fulfil emerge as well. This is an inductive process where the developers generate a rule (i.e., a value the AI system should fulfil), based on the results they have observed and the case at hand[2]. These, in turn, trigger the definition of a more detailed mode, design representation and intended actuation. The representation is gradually built upon (Image 14). As such, manipulative abduction plays a facilitative role for innovative abduction. The resulting prototype is then deductively evaluated against the requirements and values before the developers can observe the different ways in which humans might actuate it.

1 For an exemplification on how this happens, please, refer to Chapter 7, sub section "The three types of abduction".

2 See Chapter 7's Discussion for a more detailed explanation on why the generation of new values is an inductive process and the evaluation against requirements and values is an example of deduction.

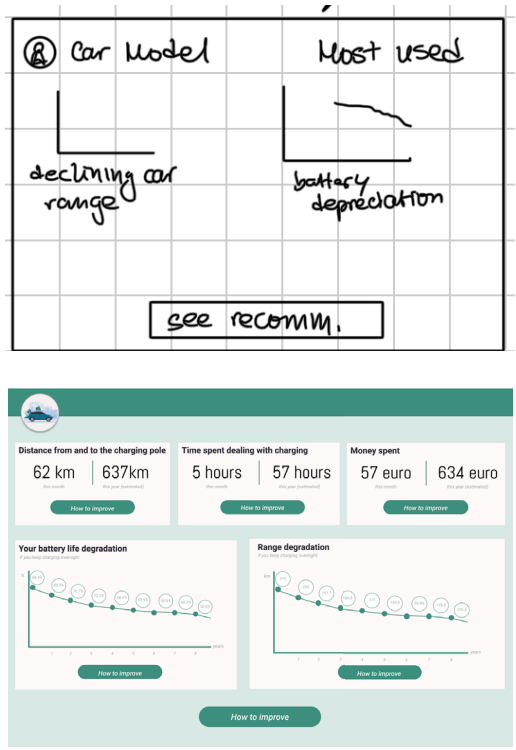


Image 14 An overview on the different types of solution representations increasing in level of detail and complexity (corresponding to the case study reported in Chapter 7).

Once the prototype becomes embedded into the context for which it was designed, the developers can start observing the manners in which users actuate the prototype (i.e., observed actuation) and the outcomes these actuations create. They evaluate the generated outcomes against the requirements and values to gauge whether they were intended or unintended. This evaluation process is a result of deductive reasoning. Next, the developers use the outcomes as a basis on which they generate both new values and requirements that the subsequent iterations through the model should fulfil. This is an act of inductive reasoning (see Figure 18).

The identified unintended outcomes become the starting point for the subsequent iteration through the model (following both Simon (1996) and Schön (1983) according to whom it is the unintended outcomes that propel the design process forward (see Chapter 3)). These surprises bring

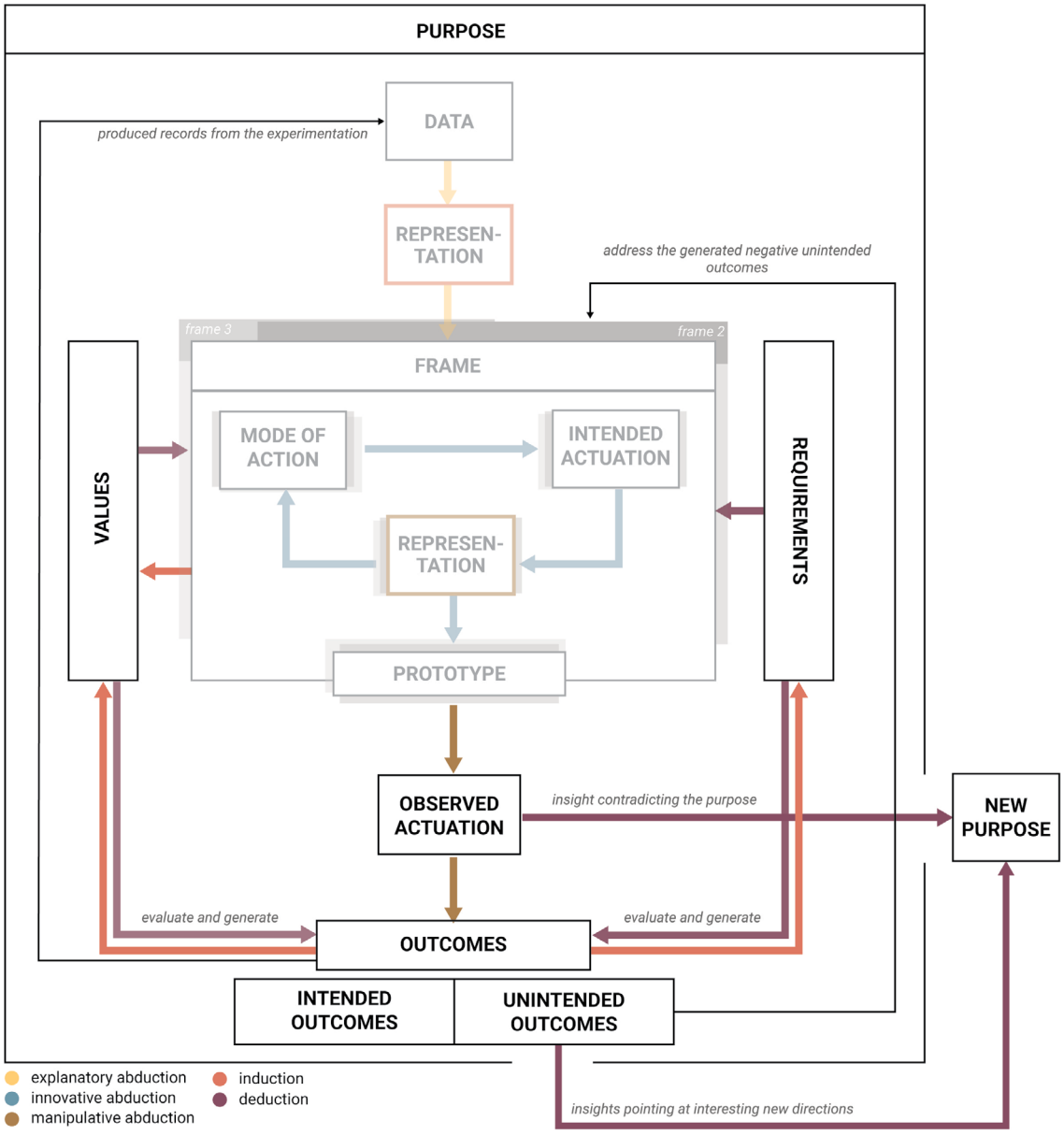


Figure 18 A schematic representation of the model with its third-row variables highlighted.

new dimensions to the problem situation (Schön, 1983; Stumpf & McDonnell, 2002) for they play a pivotal role in the actions of framing and reframing (Schön, 1983).

They also keep developers from routine behaviour and drive the originality of a solution (Dorst & Cross, 2001). Therefore, creating a model that intentionally leaves space and actively seeks cases of surprise allows developers, in theory, to elaborate on their understanding of the situation and initiate new better-fitting frames.Further, in order to facilitate the transition from one iteration to the other, we also have relationships that are transactional in nature. Namely, the data generated during the observed actuation are added to the variable of data and the unintended outcomes become the base on which the developers build their subsequent frame. Therefore, with every iteration through the model, we build a much more detailed inventory of the data types we need to collect in order to achieve the intended purpose. As a result, all variables of the theoretical model iteratively get updated and elaborated upon, too.

The purpose, on the other hand, remains relatively stable as it is not updated in every iteration through the model. However, sometimes a new purpose can be formulated. This can happen for two reasons. First, a new purpose emerges when one identifies an insight that can be seen as contradictory to their current purpose. For instance, during the case reported in Chapter 7, the data we collected suggested that EV owners consistently leave their cars charging for longer than needed. Furthermore, after six iterations, it continued to be difficult for our participants to fully charge-as-they-go (as was intended by our purpose *"help EV owners who cannot charge at home to start charging as they go*."). Therefore, a new purpose emerged: "*ensure EV owners do not occupy chargers they do not currently use"*. This definition of a new purpose is an act of deduction[3] (see Table 12).

3 It is worth noting that our move away from the old purpose (i.e., "***help EV owners who cannot charge at home to start charging as they go***") to this new one is categorically different from the act of frame creation. To substantiate this claim, let us formulate this new purpose as a frame addressing the old purpose. Namely, "if you do not occupy chargers you do not need, you will charge as you go". This is a contradiction as one can charge as they go and continue to occupy chargers they do not use. Furthermore, not occupying chargers could mean that the charging

Table 12 An explanation[4] on why the generation of a new purpose happens through deduction (following the structure Peirce (CP 2.622) [5] suggested).

Reasoning	From	To	Rule	Case	Result
Deduction	Rule + Case	Result	If people do not unplug their EVs on time, there will be not enough charging stations for everyone (fact).	EV owners who use public chargers leave their cars plugged in longer than they need to because it is convenient (observation).	Ensure EV owners do not occupy chargers they do not currently use (new purpose).

Second, although not observed in any of the empirical investigations reported in this dissertation, it is theoretically possible for a new purpose to emerge from an identified unintended outcome that can be seen as desirable by the system's developers. This assertion is based on the manner in which some popular software solutions have evolved as a response to the way they are used. For instance, the photo-sharing site Flickr was first designed to be an online role-playing game. However, once launched, the photo-sharing tool of the game became its most popular feature. As a result, Flickr changed their purpose to an online photo sharing platform (Nazar, 2013). YouTube followed a similar trajectory. Launched in 2005, it was intended to be a dating site where users can upload their videos talking about the partner of their dreams. Instead, people used the site as a means to share videos of all kinds and no one was uploading their dating profiles. Hence, YouTube adopted a new purpose (Koebler, 2015). Combining all these aspects leads us to the schema in Figure 19, where we show the relationships among each of the model's variables defined in terms of inductive, deductive and three types of abductive reasoning.

Finally, considering the relationships described thus far, one can discern the central role prototyping plays in enabling the behaviour-use interdependence

experience will not be optimal for the user (e.g., they have to unplug in an inconvenient for them time). Hence, the notion of not occupying chargers is a new purpose, not a new frame.

4 This example is based on the case study discussed in Chapter 7.

5 The example Peirce (CP 2.622) used to explain rule, case and result is as follows: "Rule: All the beans in the bag were white; Case: These beans were in the bag; Result: These beans are white."

simulation of AI systems. Therefore, we denominate the proposed model as Theoretical Model for Prototyping AI, or the **PAI model**.

Predictions and future research

The PAI model adds new insights to Design theories on abductive reasoning. As discussed in Chapter 3, Design theories use only explanatory and innovative abduction to explicate the manners in which synthesis is carried out in Design. The PAI model extends these by suggesting that for a developer to design an **ever-evolving AI system by simulating its behaviour-use interdependence**, she needs to make use of three different types of abductive reasoning – explanatory, innovative, and manipulative. It also suggests that manipulative abduction plays a facilitative role for innovative abduction. In such a way, the model formalises the different types of abductive reasoning used when designing AI systems. It also showcases how the three influence each other by prescribing the result of each of its rows to serve as the starting point for the subsequent abduction type. Further, the relationships defined by these abduction types also need to be supported by inductive and deductive moves. As such, the deductive evaluation of the model's variables against values and requirements and the inductive generation of new ones, also defines the manners in which each abduction type is triggered. In light of these developments, there are three main predictions that emerge from the theoretical and empirical investigations presented in this dissertation. The remainder of this section presents these predictions that can serve as the starting point for its testing and further research.

The PAI model aids the early simulation of AI systems' behaviour-use interdependence

The model was devised in such a way so that it can aid the early simulation of AI system's behaviour-use interdependence. In fact, herein lies the crux of our main contribution to the domain of AI: providing a theoretical model that supports the early simulation of one of the four main challenges we identified in Chapter 2 – behaviour-use interdependence. Therefore, it is our contention

that it will aid the explicit simulation of different uses (i.e., observed actuations), so that the behaviour of the AI systems (i.e., mode of action) and its intended use (i.e., intended actuation) can be adapted to mitigate harmful outcomes and ensure the outcomes it generates are aligned with human objectives. This prediction also stems from the observations discussed in Chapters 6 and 7. For instance, the student team discussed in Chapter 6 observed that their participants follow the AI's advice even when it contradicts their own preferences (e.g., dislike eating sea food). To address this unintended outcome, they adapted their system's behaviour so that it will provoke people to think about their choices. We argue that this pattern will be observed during the conceptual design of most types of AI systems regardless of the industry, context, or company. Further research is needed to falsify this claim. Such could also aid us in expanding our understanding on (1) the interaction between manipulative abduction and explanatory and innovative abduction; and (2) the manner in which requirements and values emerge while applying the theoretical model. A suitable research approach for this would be video and audio recorded think-aloud protocol study (e.g., Christensen & Abildgaard, 2016) that focuses both on the types of verbal and visual reasoning developers use. Doing so will allow for the examination of the interaction among all different types of abduction on a more granular level. Furthermore, it will be complementary to the research presented in this dissertation where the interaction among the different types of abduction and emergence of values and requirements was presented on the scale of five-month long projects.

Finally, the PAI model was designed with AI systems that face ill-defined problems in mind. Therefore, the first prediction *will probably not hold true* when developing AI systems that *have to provide a clear yes or no answer* (i.e., the ones addressing tame problems). Namely, systems which are implemented in environments and contexts that have clear procedures and rules. An example of such would be the deep-learning ML model developed by Google's DeepMind, AlphaFold, which managed to improve the process of protein folding – a problem that has resisted scientists for decades. Using the AlphaFold, scientists are now able to accurately predict protein structure from their amino-

acid sequences, which leads to faster drug discovery and studying of diseases (Callaway, 2020).

Different types of developers will apply the PAI model as prescribed

Different types of developers might approach the conceptual design of an AI system in a different manner. For instance, they might decide to take a user-centred or a systemic view, or a co-creation approach to the designing of the AI system. Each of these design approaches is associated with distinct design methods, philosophies, and rationales for the involvement of stakeholders. However, regardless of the developers' skill level, educational background, or the design approach they adopt, we contend that once a *principal solution* is designed, each developer will follow the cognitive steps prescribed by the theoretical model. For instance, the developers would first need to gather enough data so that they can formulate a frame; and only when they have a frame, can they move to the definition of the mode of action and intended actuation. Further, the types of design representations they might use to make a transition from data to frame, might be different. Still, they will use some type of design representation to do so, be it a sketch, data visualisation, role playing, or an artefact they use as a boundary object.

We predicate this prediction on the explicit decision we made in Chapter 3 to base the development of the theoretical model on principles stemming from the two main paradigms in Design – Rational Problem Solving (Simon, 1996) and Reflective Practice (Schön, 1983). As explained in that chapter, the design process, in its entirety, can be described only when we combine the two paradigms (Dorst & Dijkhuis, 1995; Dorst, 1997; Roozenburg & Dorst, 1998). Therefore, we expect the combination of the principles we embedded into the PAI model to allow us to accurately represent the cognitive steps one takes when designing an AI system.

This prediction will most probably not hold true when it comes to the pattern in which requirements and values emerge during the design process. According to the theoretical model, new values first emerge when we start attempting

to formulate a new frame (e.g., see Chapter 7). However, when stakeholders are initially involved with the purpose of identifying their values (e.g., by using interviews or focus groups during Value Sensitive Design (e.g., Friedman et al., 2002)), both requirements and values will emerge during the data collection step, not from the attempt to generate a frame. Further research is needed to falsify this prediction. For instance, where different types of developers (with different skillsets, educational and cultural backgrounds) apply the model to the same context. Finally, additional research is needed to investigate whether developers use all of the variables prescribed by the PAI model before their principal solution has been designed.

The PAI model provides an indication on how data can be used to update AI system's behaviour during model development and deployment

The behaviour (i.e., mode of action) of contemporary AI systems is directly dependent on the data generated during its use (see Chapter 1). Currently, in Data Science the widely accepted approach to this problem is to monitor the performance of the ML model once it gets deployed (Russell & Norvig, 2021). As such, when one is to ensure a deployed AI system behaves as intended, they look at the difference between the model's training (i.e., offline) data and the data that has been generated once the model has been deployed (i.e., online data). When the model behaviour does not fit the objective (i.e., purpose) of the model, a retraining strategy is employed to update the behaviour towards the new objective. However, when retraining the model, it continues to be difficult to decide whether this new behaviour would lead to the intended uses and hence prevent potentially harmful unintended outcomes.

The research presented in this dissertation is focused on the early simulation of AI systems' behaviour-use interdependence during the conceptual stages of its development. Nevertheless, it provides indications on how different data types can be utilised to modify the model's behaviour. Firstly, we use the same two types of data as the ones used during model monitoring: the data we started with (i.e., training data) and the data generated during the tests (i.e., online data). However, we also deliberately collect additional data (i.e., new data),

which is directly dependent on the requirements we have already identified. For instance, during the case reported in Chapter 7, one of the requirements was to *provide a clear overview on how much charging is going to cost*. Then, in subsequent iterations, we collected data about e.g., *charging stations, fees, and availability*. As such, the way in which we decide to collect the new data seems to be an activity guided by explanatory abduction (see Table 10).

Table 10 An explanation[6] on why the identification of new data types is indicative of explanatory abduction (following the structure Peirce (CP 2.622) [7] suggested).

Reasoning	From	To	Result	Rule	Case
Explanatory abduction	Result + Rule	Case	Participants feel anxious over the too many unknowns of charging at a public station (outcome).	Provide a clear overview on the charging process (requirement)	Collect data about charging routines of users and existing infrastructure around their regular routes (collected data)

Neither the theoretical model, nor the empirical research we discussed thus far, was aimed at addressing the manner in which a deployed model can be retrained. Therefore, the hypothesis that the PAI model could provide us with an indication on how to do so is a highly speculative and easily falsifiable one. Still, the insights presented here could aid us in understanding the logical relationships that define the way data influences a model's behaviour and purpose. Further research is needed to falsify this hypothesis. Potential directions for such could be studies to better define (1) the relationships among the three data types we identified, (2) the reasoning modes that connect them to mode of action, (3) how these are used to update the system's purpose, and (4) how all these can be applied during both model development and deployment stages. One suitable research approach for this would be video and audio recorded think-aloud protocol study (e.g., Christensen & Abildgaard, 2016). Another – an experimental setup where two models are used: one that is already deployed and another one that monitors the data types used when retraining the first model.

6 This example is based on the case study discussed in Chapter 7.

7 The example Peirce (CP 2.622) used to explain rule, case and result is as follows: "Rule: All the beans in the bag were white; Case: These beans were in the bag; Result: These beans are white."

Good theory criteria

In the Introduction chapter of the dissertation, we discussed the requirements for "good theory" Wacker (1998, 2008) put forward. Therefore, in order to evaluate the quality of the PAI model, in the remainder of this section, we evaluate it against each of Wacker's requirements (see Table 11).

Table 11 An overview on the components of good theory, the requirements they should fulfil and their explanations (Wacker, 1998; 2008), paired with an evaluation of the theoretical model for each requirement.

Component	Requirement	Explanation	Evaluation
Definitions	Conservatism	New terms can be used only if they can be clearly distinguished from existing ones	The theoretical model was built on existing Design theories, reviewed in the light of designing AI systems. As such, the majority of the used definitions are the same as in already existing theories (with small adjustments to fit the AI context – i.e., substitute the word "*product*" with "*AI system*").
	Uniqueness	New definitions should not borrow from existing conceptual definitions	The only new definition we introduced was the one for "*prototyping*", which, despite being inspired by existing definitions, does not borrow conceptual terms from them (see Chapter 4).
	Parsimony	Each definition should be as short as possible	All adopted definitions are succinct (see Table 9).

Domain	Generalisability	The theory can be applied to more than one area	As it can be seen from the example in Chapter 6 and the case study in Chapter 7, the theoretical model can be applied to the design of different AI systems (i.e., health recommendations and intelligent in-car system[8]).
	Abstractness	The theory can be applied across different times and places	The theoretical model can be applied to different industries, contexts and at different times. Furthermore, the model is built upon principles of the two main paradigms of Design Theory. Thus, it is conductive to design procedures, principles, and practices typical for each paradigm (see Chapter 3).
Relationships	Fecundity	Build upon existing theories and integrate them	The main premise of this dissertation was to use existing Design theories and principles and adapt them to the design of AI systems. Hence, the theoretical model builds upon and integrates existing theories (Part II) which were then extended (Part III).
	Internal consistency	Each variable is clearly connected to the other variables through mathematics or symbolic logic	The relationships among the variables are informed by both theoretical and empirical investigations. As such, they are described both formulaically and schematically utilising different types of reasoning (i.e., abduction, deduction, and induction).
	Parsimony	The resulting models should be as simple as possible	The theoretical model introduces only assumptions that are derived from existing theories or empirical insights. Moreover, the introduced relationships were iterated upon multiple times in order to ensure that only the relevant relationships among the variables are used.

8 In addition, the theoretical model has also been applied to the design of an AI system for incentivised healthcare insurance (Vriens, 2022) and one for fair hiring practices (van der Ploeg, 2021). These were carried out by two of my graduation students.

Predictions	Falsifiability	Predictions should be unlikely, based on extant literature and the conditions in which they cannot occur should be explained.	The three predictions outlined in this chapter can be easily falsified – e.g., with a case study of two different design approaches (e.g., a user-centred and co-creation) can track whether the theoretical model is applied as prescribed. Further, all predictions are based on extant literature and suggest conditions in which they will not hold true.

As it can be seen from the table, the PAI model fulfils the criteria for a "good theory" Wacker (2008) introduced. The explicit decision to use existing Design theories as the starting point allowed the model to fulfil the requirement of definitions' conservatism, and relationships' fecundity and internal consistency. It also allows us to ensure continuity between the already existing foundational knowledge of Design Theory developed in the past seven decades and adapt it to the challenges of designing AI systems. In addition, both the theoretical and empirical investigations allowed for the clear definition of a domain and its limitations and further ensured the newly introduced definitions are unique and parsimonious. The carried-out investigations also strengthen the internal consistency of the model's relationships. Finally, the three predictions we outlined are falsifiable, point at conditions in which they will not apply and are based on extant literature. Hence, the theory-building method we adopted allowed for the creation of robust scientific knowledge. Such robustness is important to introduce potential avenues for further development and evolution of the fundamental principles of Design.

The implications for Design(ers)

The final theoretical model brings to the fore three main implications for Design(ers). In the remainder of this section, we briefly discuss each one of them.

Implications for Design theory

As we saw from all of the examples discussed in this dissertation, as well as the presented empirical explorations (Chapters 6 and 7), designing AI systems

is categorically different than designing products or services. Unlike products or services, the behaviour of an AI system is continuously adapting to the data its users generate (Chapter 1). This continuous automated improvement leads to better decision-making mechanisms. However, it also gives rise to a multitude of unintended and unanticipated behaviours and uses. Such are numerous and continue to be hard to influence (Rahwan et al., 2019) due to the unparallel level of scale and personalisation AI systems support (Amodei et al., 2016).

Our empirical and conceptual theoretical explorations showcase that further development of the fundamental theories and practices of Design are needed to accommodate for the designing of ever-evolving AI systems. While the existing theories continue to be relevant, they need to be adapted and extended for designing with this new material. Consequently, the PAI model we introduce in this dissertation could have the following implications for Design theory.

Firstly, one key facet of our theoretical advancement involves the manner in which we formalised the role of prototypes to align it with the temporal and mutable nature of AI systems. By formally establishing prototypes as a bridge facilitating the relationship between mode of action (behaviour) and actuation (use), we offer a structured means to elevate the outcomes of prototyping to the abstraction plane of abductive reasoning. This positioning could provide further insights into the cognitive significance prototypes play during the design reasoning process. Thereby, paving the way for future investigations that could extend our grasp of cognitive design processes.

Secondly, while the initial model, as outlined in Chapter 5, was based on the three types of abductive reasoning, it did not explain their interrelationships. The conceptual refinement presented in Chapters 6 and 7 rectifies this gap, providing a frame of reference that could elucidate the nuanced roles played by each form of abduction, as well as their relationships with deduction and induction. The empirical examples dissected in Chapters 6 and 7, particularly concerning requirements and values, offered a novel perspective.

Thirdly, the role of visual representations in Design has long been a subject of research. While their significance in Design Cognition and Design Methodology

is widely acknowledged (Goldschmidt, 2003; Gonçalves & Cash, 2021), their specific role within different types of reasoning, particularly abduction, has not been explicitly addressed. Our empirical insights and subsequent conceptual development propose a novel perspective—visual representations can facilitate both induction and manipulative abduction. In particular, they do so during other types of abduction: explanatory and innovative, respectively. As such, these insights could promote a more detailed understanding of the role visualisations play during reasoning.

Collectively, these insights unravel intricate reasoning patterns, casting a spotlight on how foundational concepts such as visual representations, prototypes, requirements, and values intertwine with reasoning patterns. In addition, this model establishes a foundation for future scholarly inquiries not only into the manners in which we can simulate potential behaviour and use, but also, to address the other three broad challenges the implementation of AI into existing systems faces: (1) **ensuring transparency**, (2) **mitigating implicit biases**, and (3) **aligning with human values** (see Chapter 2 for a detailed explanation). The development of new methods and theories will be foundational for this.

Implications for Design education

One of the most fundamental pieces of knowledge a young industrial designer receives is a course on mechanics and material properties. Similarly, when designing AI systems, designers need to be able to understand the material with which they are going to design. For instance, some of the materials they need to have a basic understanding of are the so-called big data, sensors (which gather the data), data configuration, model building, and model monitoring. Designers need to have at least a working understanding of these materials. Otherwise, they will be excluded from the process of designing AI and will be relegated to the late development stages where the principles of the systems have already been designed and the designer would need to design the user interface for them. Once pushed to these later stages, designers would

have a very limited opportunity to define the objectives of the AI model and consequently the system in such a way so that it will be "good"[9] for humans.

In addition, designers also need to get acquainted with designing for something that is transient. For instance, they need to get familiar with utilising prototyping in their generative capacity (see Chapter 4). We contend that this will make it easier for students to become comfortable with a process that does not have a clear end (rather, it is exploratory) and where the primary aim is to satisfice the given and emerging requirements and values. Design has a long history with the notion of satisficing (see Chapter 3 and more specifically Simon (1996)). Still, design students continue to struggle to understand when to stop exploration and when their design is "done"[10]. This struggle will become exacerbated once designers start designing for something that does not have a clear end state. The theoretical model can support students in getting more comfortable with designing for something transient. Namely, the clearly defined relationships can guide the designer on how the results of each row of the model can be used as a starting point for the next.

Implications for Design practice

The billable-hour funding model is a standard practice for design agencies and consultancies. However, an updated business model is needed that could address the new market situation in which AI systems continuously evolve and each of their incarnations has to be monitored and adapted to ensure the delivered solutions are aligned with human interests. We contend that a more suitable business model would be for agencies to foster longer-term relationships with their clients (e.g., they can become an agency in residence). Namely, design agencies can be involved with the designing of the AI system, but also in the process of system maintenance where organisations need to ensure the deployed AI system is continuously adapted to mitigate potential

9 When we use the term "good" here, we refer to the notion addressed in the Introduction of the dissertation that the solution to a wicked problem can only be good or bad (Ritter & Webber, 1973) and the discussion on "good" in Chapter 2.

10 This statement is based on my own experience teaching both master and bachelor students in the Industrial Design Engineering faculty in Delft and supervising design students from both TU Delft and other technical universities in the Netherlands.

negative unintended outcomes. This structure could be reflected in a funding model where design agencies use billable hours for the conceptual design of the AI system, as well as when major updates are needed. In addition, they can use a retainer funding model that allows them to bill a client each annum (or quarterly) to “retain” their services.

Furthermore, agencies which only provide design services will struggle to adapt to this new context unless they acquire the skills and methods to do so. This contention is supported by a strong industry trend of acquiring multidisciplinary expertise. For instance, IDEO, one of the biggest design agencies in the world, acquired the data science firm Datascope in 2017 (IDEO, n.d.). Further, a number of large design agencies like Fjord and Lunar were acquired by big consultancy firms like Accenture, BCG, and McKinsey. In such a way, they are able to provide multifaceted services. Still, further efforts are needed to ensure designers can collaborate with data scientists, engineers, ethicists, and people who are able to navigate the new regulations on AI systems that the European Union, United States and China are working towards. The development of new methods and theories will be imperative in doing so. The PAI model can provide the scaffolding around which design practitioners can build their new methods and tools. For instance, its clear delineation on how elicited outcomes can be addressed by the creation of a new frame, and consequently a prototype, can support the design agencies’ continuous involvement into the full lifecycle of AI systems.

Conclusion

Since its inception, the field of AI has been focused on devising systems that can provide clear answers. AI systems can tell us how to move a chess piece, translate a word, fold a protein, and predict whether a person would buy a book. These systems have a clear objective, clear outcome, and in the case of reinforcement learning – clear reward and punishment functions, too (see Chapter 1). Consequently, the answers they provide can only be correct or

incorrect. After a chess piece has been moved, we can easily check whether the move was correct and has increased the odds of winning.

Due to their accuracy, AI systems have led to impressive strides in numerous fields such as speech and visual object recognition, object detection, drug discovery, physics, and genomics (Ching et al., 2018). For instance, AI systems helped scientists to capture the first image of a black hole's silhouette (see Introduction chapter), improve the protein folding process, and beat the human champion in the ancient game of Go (Chapter 1). Thus, it comes as no surprise that such systems have now become an integral part of people's everyday lives. However, once AI systems that need clear rules and objectives face the complexity of existing contexts, they start to produce a large number of diverse unintended and sometimes harmful outcomes. As the examples used throughout this dissertation show, such systems could lead to the labelling of black people as gorillas, perpetuating pre-existing socioeconomic inequality, and the spread of misinformation. In such situations, neither the problems nor the objectives can be clearly defined, and instead of true or false answers, there is a multitude of potential solutions that can only be good or bad (see wicked problems (Chapter 2)).

It is against this backdrop that we started the theoretical and empirical exploration presented in the dissertation to which Chapter 8 serves as a close. Following the three theory building steps prescribed by Cash (2018), the PAI model we presented provides us with a way to approach the conceptual design of ever-evolving AI systems through the early simulation of their behaviour-use interdependence. Hence, it allowed us to shed light into the ways in which Design theories could contribute to our understanding of how "good" AI systems can be designed. However, the Design theories we utilised were devised for a very different context. Therefore, they need to be extended and potentially revised. The PAI model serves as one potential way to do so.

Finally, the PAI model also provides us with a foundation on which the theory-testing part of the cycle Cash (2018) introduced can be carried out. When doing so, it is important to remember that the behaviour-use interdependence is only one of the challenges to devising "good" AI systems. The other three are:

(1) ensuring transparency, (2) mitigating implicit biases, and (3) aligning with human values (see Chapter 2). The four are interrelated and heavily influence each other. Therefore, the further development and testing of the PAI model should happen in awareness and conjunction with the other three challenges and the advances made in each of the research fields addressing them.

EPILOGUE

BEYOND DESIGNING CONCEPTS FOR AI SYSTEMS

In 2020, OpenAI, a research laboratory in San Francisco, introduced the third generation of its deep learning language model named General Pre-trained Transformer (i.e., GPT-3 for short). Trained on 175 billion parameters, the model was so good that it could write academic (e.g., Osmanovic Thunström, 2022) and news articles about itself (e.g., GPT-3, 2020), create websites from just a prompt, develop chatbots, and translate English into coding languages like Python (e.g., Heaven, 2020). In 2021, one of the applications built around it, Project December, gained notoriety after one of its users, Joshua, created a chatbot and infused it with the personality of his dead fiancée, Jessica (see Fagone (2021)). After paying $5 to get access to the project, all Joshua had to do was to provide GPT-3 with a short summary describing Jessica, a few of the text messages she sent to him when she was still alive and then, define the roles both him and the chatbot should play. Their interaction lasted for several months. Over time, Joshua claimed, the chatbot would sound so much "just like Jessica" that he decided to use the bot as a moment of catharsis (Image 15). Through their conversations, he reasoned, he would be able to move on and deal with the grief he had been feeling since the death of the real Jessica.

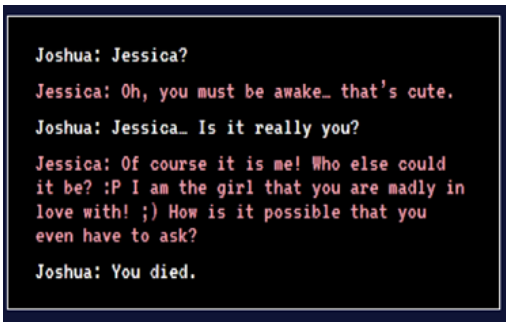


Image 16 A snipped of the conversation between a New York Times' reporter and the GPT-3 version of Bing Chat (Roose, 2023).



Image 17 A still from the deepfake video of Sedar Soares (NOS News, 2022).



Image 18 An image generated by Jason Allen using the Midjourney AI model which won the Colorado State Fair's digital arts competition in 2022 (Harwell, 2022)

Shortly after the article written by Fagone (ibid) went viral, OpenAI asked the creator of Project December to shut it down (Quach, 2021). In 2023, when Microsoft embedded GPT-3 capabilities into its Bing Chat, a New York Times' reporter had a similar experience. This time, however, the chatbot started professing its love for the reporter and telling him to divorce his wife unprompted (Image 16) (see Roose, 2023).

Similarly, in 2022, the Dutch Police created a deepfake video of a 13-year-old boy, Sedar Soares, who was shot dead while playing with his friends in 2003. The video, created with the approval of the boy's family, was released as an attempt to finally solve the cold case (NOS News, 2022). It featured a deepfake version of Sedar asking people to give information to the police that could solve his murder. The deepfake was created by using only two of the boy's photos and was added to a short video shot on a football pitch (Kivits, 2022) (Image 17). In a statement, the Dutch Police claimed to have received numerous tips after the video was released (ibid). However, the video also raised



Image 19 A visual generated jointly by the Mid Journey ML model and DALL-E after I gave them the following prompt: "Artificial Intelligence that keeps evolving in unintended ways the more people use it, inside a magical realm, hyper realistic, wide angle".

multiple questions around the ethics of using someone's image after they have passed away (Hendriks, 2022).

Examples like these raise a multitude of both ethical and philosophical questions that go far beyond the scope of this dissertation. Yet, they, and all of the other examples discussed thus far, clearly show the extent to which AI systems have permeated our everyday lives – from simple recommendation systems to solutions that influence our experiences of love, loss, and grief. AI systems have also started to influence the way we do science and art. A prime example of the former would be the AlphaFold model discussed in Chapter 8 or the coordination of multiple ML models that were used to take the first picture of a blackhole in 2019 (BBC, 2019). When it comes to the arts, in 2021 and 2022, multiple companies like OpenAI, Microsoft and Google, as well as other smaller labs (e.g., the independent research lab Midjourney) released models that can translate a text prompt into a high-quality image. The hyper realism of the generated images and the win of the Colorado State Fair's digital arts competition by a piece of art generated by Midjourney (Kuta, 2022) (Image 18) have led many to discuss the future of art and creativity (e.g., Rizzo, 2022, Harwell, 2022). An example image generated by me using two such models can be seen in Image 19.

Despite these impressive achievements, however, the types of AI systems discussed thus far are examples of the first stage of AI development – narrow (weak) AI. AI systems at this stage are bound to a specific field and are incapable of performing tasks outside a preprogramed scope. Although some widely publicised AI implementations tackle more general tasks, such as driving a car (Tesla's autopilot) or generating websites from a sketch (e.g., GPT-4), these examples are still considered a coordination of several narrow AIs. Namely, AlphaFold is very good at sequencing proteins but cannot generate an image out of text as DALLE-2 does.

The strive towards general AI

An increasing number of scientists and companies are working towards the achievement of the next stage in the development of AI – the so-called general

AI[1] (Noessel, 2017). For instance, the firm Deep Mind, which Google acquired in 2014, has stated its long-term goal is to create general AI (Deep Mind, n.d.). Similarly, OpenAI was founded in 2015 with the mission to ensure general AI "benefits humanity" (OpenAI, n.d.) and the head AI scientist in Meta, Yann LeCun[2], who is widely considered as one of the godfathers of AI, released a positioning paper in 2022 detailing a conceptual model architecture that can be used to achieve general AI (LeCun, 2022). In fact, a growing number of renowned scientists, philosophers, and forecasters predict general AI creation by mid-twenty-first century (Stoimenova & Price, 2020). It is their belief that accelerating progress in hardware, AI, robotics, genetic engineering, and nanotechnology make this timeframe achievable (Müller & Bostrom, 2016).

One important step in achieving such human-level performance, according to the Stanford's AI Index report for 2022, is the AI's ability to generate abductive inferences (Zhang et al., 2022). Despite being explored in the field of AI as early as the 1970s[3] (Bylander et al., 1991), abductive reasoning[4] has received relatively limited attention thus far (Bhagavatula et al., 2020). However, this is changing since even the most advanced models still perform significantly worse than humans when asked to generate a plausible explanation. For instance, when Bhagavatula and colleagues (2020) tested state of the art language models on explanation-generating tasks, they found that these models perform significantly worse than humans. For instance, when they used state of the art models such as BERT[5] to pick one out of two hypotheses as the most plausible one given a known observation, the models achieved an

1 General or strong AI are "systems which match or exceed the [intelligence] of humans in virtually all domains of interest." (Shulman & Bostrom, 2012).
2 Yann LeCun, Yoshua Bengio and Geoffrey Hinton are widely considered to be the three godfathers of AI. In addition, LeCun is not alone in his pursuit of general AI – both Hinton and Bengio are working on approaches for achieving such (e.g., see Bengio, 2019)
3 For instance, one of the earliest systems to use abductive reasoning was INTERNIST-I which performed medical diagnosis (Miller et al., 1985)
4 Abductive reasoning in the field of AI is generally equated to the theory of inference to the best explanation (IBE). Abduction and IBE are different (for an in-depth discussion on the topic, please, refer to Campos (2011)). However, for the sake of simplicity, we can equate the abduction authors in the field of AI use to explanatory abduction.
5 BERT stands for "Bidirectional Encoder Representations from Transformers" and it was developed by Google.

overall performance of 68.9% accuracy compared to that of humans, 91.4%. Further, when they tasked such models (including GPT-2[6]) to generate a hypothesis from a given observation, none of the models completed the task successfully[7]. Therefore, abduction, which needs only limited number of data points to generate a plausible explanation, could present us with a considerably more computationally and resource efficient partner to modern approaches like deep learning[8].

It is important to note that the type of abduction discussed in the AI field literature is akin to explanatory abduction: a logical explanation of an observation used to identify the use of a known to be true principle. As explained in Chapter 3, this can be a law or theory such as, "*If a motor has no gas, then it does not start*". Therefore, it allows us to reason from the observed effect ("*The car doesn't start*") to a possible cause ("*The tank is empty, I guess*") (Roozenburg, 1993, p. 10). Yet, as we saw from the research presented in the previous chapters, utilising only explanatory abduction when addressing wicked problems is not sufficient. We need other types of abduction, too. The positioning paper of LeCun (2022) supports this claim despite not mentioning the term abduction.

According to LeCun (2022), in order to achieve human-level intelligence, we need new models that enable "**reasoning by simulation and by analogy**" (p. 13). Hence, he proposed a new model architecture that can help us achieve such. The architecture is built around six distinct modules[9]: (1) *configurator* (takes input from all other modules and adapts it to the task at hand); (2) *perception* (receives and registers the sensor signals that allow it to understand the surrounding environment); (3) *world model* (serves as a "simulator" that can predict the relevant aspects of the world and it is "dynamically configurable for

6 The precursor of GPT-4.

7 It is worth noting that my limited experiments with GPT-3 show that the model is capable of generating explanatory hypotheses.

8 Deep learning models require a vast number of data points (e.g., GPT-3 was trained on 175 billion parameters) and we are starting to see the limitations of what models purely trained on large number of data points can achieve (e.g., LeCun, 2022). This is a departure from the widely held belief in the field that more data (and better-quality data) equals better models.

9 The in-depth explanation on each module is beyond the scope of this Epilogue.

the task at hand" (p. 5)); (4) *cost* (calculates the energy the system will exert for a given task), (5) *short-term memory* (stores information about the past, current and future states of the world) and (6) *actor* (computes sequences of proposed next actions). The key module of the architecture that could allow us to reason by simulation is the *world model*[10], which LeCun claims to be the most complex part since it is through models that humans and animals learn how to interact with the world around them (p. 2). Consequently, this module is the one that helps the AI agent make sense out of the world by fulfilling two main purposes: (1) estimating missing information and (2) predicting plausible future states of the world. Hence, in effect, LeCun suggests that in order to be able to reason "by simulation and by analogy", we need to utilise reasoning that allows us to use and manipulate models in order to make sense out of the world. As discussed previously (e.g., Chapter 4), the type of reasoning that allows us to make sense out of the world through models is manipulative abduction.

Therefore, the strive to create general AI can be directly linked to two types of abduction: (1) explanatory and (2) manipulative. The research presented in this dissertation, however, suggests that in order to enable these, one also needs innovative abduction as well as induction and deduction. For one cannot be performed without the others when trying to address a wicked problem. Our theoretical model, despite being designed for a different purpose[11], introduces formalised relationships among these different types of reasoning. As such, it defines logically how the world model (i.e., mode of action) of an AI system can be devised and adapted to reflect the elicited outcomes of the actuation of the said system. Hence, it could provide us with initial indications on how entirely new AI models can be devised that do not rely solely on vast numbers of data points. It could also aid the creation of highly configurable world models

10 The use of world models in reinforcement learning is making a comeback, as well (e.g., Levine, 2022).

11 The main premise behind the theoretical model was to ensure the AI systems we build and deploy behave in the manner we (humans) want them to behave.

AI agents can use. This contention is, of course, highly speculative. Further research is needed to explore its merits.

"Only stupid people do a lot of calculations."

Finally, I will afford myself some more leeway for speculation purely based on my own values. While working on this dissertation, I would often think about my mathematics teacher. She loved repeating a simple aphorism which can be roughly translated as: "*Only stupid people do a lot of calculations. Smart people find elegant ways to solve the problem*". Elegance, in this context, meant that one is able to solve the problem with simple logic instead of multiple lines of calculations. She would usually tell me that over and over again while asking me to find yet another, more elegant, solution to a problem I had already solved. At first, I thought it to be futile and frustrating – who cares how I solved the problem as long as my solution was correct, and I was faster than the others. Over time, however, I started to develop different "elegance strategies" and began to appreciate the manner in which doing so made me see the problem from a multitude of angles.

The current state of AI reminds me a lot of what she would label "*stupid people doing a lot of calculations*". This is by no means a categorisation of the people who work in AI and ML or the impressive strides we have witnessed in the past decade. Rather, it is a comment on the approach we have to devising AI models. The most successful ML models today require an enormous number of data and computational resources in order to be trained. GPT-3, for instance, was trained on 45TB data and 175 billion parameters[12] (Brown et al., 2020). Further, some estimate, it costs OpenAI around $100 000 per day to run ChatGPT[13]

12 Parameters define how the model input is to be transformed to the desired output. As such, they are learnt from the data on which the model is trained.
13 ChatGPT is an application OpenAI released to the public in 2022.

(Goldstein, 2022). Yet, we are already beginning to see the limitations of such models (see e.g., Bengio, 2017; LeCun, 2022).

The PAI model, with its foundation in abductive reasoning that by definition requires fewer "calculations"[14], is my attempt to address my mathematics teacher's notion of elegance. Striving for elegance could have another important benefit, too. At the time of writing, only a handful of companies can afford to train a deep learning model like DALL-E or AlphaFold. An even smaller number has the data of the right quality to do so. These limitations (inadvertently) create a strong model monopoly by big technology companies like Microsoft, Amazon, Google, and Meta (in the US) and Tencent, Baidu, and Alibaba (in China). In theory, one would be able to develop new AI systems using these models. Yet, in reality, she has to do so using the cloud services provided by Big Tech, the models they developed and abide by their rules. Rules that are put together by a very small number of people in a "room" somewhere in Silicon Valley or Shenzhen, Beijing, and Hangzhou[15] (see Bloomberg Technology, 2022). I would like to believe that there is a better way to design AI models and systems. A smarter, more elegant way[16] and that the research presented here could support our exploration for one.

14 i.e., only a limited number of data points are needed to generate a hypothesis.
15 This statement was eloquently made by Blake Lemoine who was fired from Google in 2022 after claiming that one of Google's AI, LaMBDA, has a soul (Tiku, 2022). Despite being a target of ridicule online and in the media, the engineer, raised a number of concerns around the manner in which AI systems are developed, especially when it comes to the firing of multiple AI ethicists who sounded the alarm on Google (Schiffer, 2021).
16 The discussion among Tomaso Poggio, Yan LeCun and Pierre Baldi presents a number of interesting insights on the importance of new mathematical theories to support AI development (University of Padua, 2022)

APPENDIX



APPENDIX

CHAPTER 6

Table 1 An overview on the formulation of each variable. In each iteration, only the variables that were updated by the team are described.

Variable	Performed activities
Iteration 1	
Purpose	Support people in establishing and sustaining healthy routines.
Data	Data from introductory interviews with users on their exercise and eating routines. As well as on the goals they wish to achieve.
Frame	"If the AI gives advice both in alignment with the user's preferences and in deviation from them, more varied routines can be introduced to the user."
Mode of action	Give advice to users that is largely in line with the preferences they indicated.
Intended actuation	Follow the advice given by the app.
Prototype	A fitness app powered by an AI that guides people through their attempts to lead a healthier lifestyle with two main features: exercising and diet.
Mode of action	As before + occasionally ask them to do things they specifically said they dislike.
Intended actuation	As before + react on the suggestions that are contradictory to the preferences they stated.
Prototype	As before + basic workout schedule.
Actuation	E.g., the participants did not click on every option in the app.

Outcomes	E.g., the participants found the suggestions not personalised enough, they felt in control of their routines.
Requirements	Apps suggestions should feel personalised; Each of the app's suggestions has to be aligned with likes and dislikes of the users; focus on the content of the workout rather than on whether it should be outside or inside; Provide recipes for each suggested dish.
Iteration 2	
Data	As previously + a collection of relevant meals and users' weekly schedule
Frame	If the AI intentionally makes mistakes in its advice, the app can elicit the personal preferences of its users more easily.
Mode of action	Provide personalised advice that is in contrast with the preferences they stated during the entry questionnaire and could be perceived negatively by them.
Intended actuation	Interact with all three options for meals (e.g., breakfast, lunch, and dinner) and rate the suggestions.
Prototype	An updated version of the app which provides workout schedule and suggests meals to try so that the user's goal can be achieved.
Mode of action	As before + introduce e.g., extreme exercises (too hard or too short) and long meal prep times and exotic ingredients (*"if our user is working from 9:00 until 17:00, the app is going to suggest doing exercise at 10 am")*
Intended actuation	As before + try to adjust the proposed workout schedule and teach the system about their preferences.
Prototype	As before + strenuous workout schedule
Actuation	E.g., Reluctance to follow the proposed schedule; focused on the aesthetic features of the app; tried to find the recipes for the suggested meals; provoked by suggestions that were not meant to be provocative; confusion about the interaction with the app.
Outcomes	E.g., users change their preferences based on AI's suggestions (*"Again, seafood. I'm not a big fan of seafood. But, it kind of looks very good in this picture. So uhm, let's say okay, not so bad as before, so I can try it.");* refusing to let the app interfere with their freedom *("Okay, this is way too busy. I don't want to do that many sports because I have other things to do").*
Requirements	Provide users with different exercise options; Simplify the interface of the app; Simplify the meal prepping process.
Values	Users should not follow the AI's advice blindly
Iteration 3	
Data	As previously + more exercise options, detailed meal suggestions (including ingredients, instructions, nutritional value, time) and food delivery options

Frame	If the app provides more nuanced deviations from the preferences of the user, they won't readily comply with them.
Mode of action	As before + including deviations in the workouts and introducing difficult meal plans.
Intended actuation	Carefully read and reflect on the suggested exercises and meals.
Prototype	As before + update of the app interface to include dish recommendation (and detailed recipes) and food delivery.
Mode of action	As before + assign labels to each user that can showcase how they performed during the past week.
Intended actuation	As before + notice the labels the algorithm has assigned to them.
Prototype	As before + image sequences showing different exercises and times.
Actuation	E.g., refusal to perform more challenging exercises; tried to change the proposed schedule; wondered how to provide feedback to the AI so that it would not suggest such exercises; wondered how to make the AI show easier recipes.
Outcomes	Participants are puzzled over the choices the AI made (e.g., one participant tried to figure out what he might have said in the first interaction that made the AI think that his diet should be vegetarian: "*well, there's a question I have here... I don't know if the current diet is something that I've said that I have or it's the diet my coach suggested to me...");* a photo of the female yoga instructor was considered too prejudiced and not inclusive enough for men who like to practice yoga by a male participant); too complex suggestions (e.g., *"40 min? 40 min for a sandwich?"*); compliance with suggestions of the app even though the suggestion was meant as something the user will consider a mistake (*"Preparation time 8h, okay. Yeah, I mean that also looks very yummy. But I again would not like to cook during the week, something that takes eight hours. Then I would try it out on the weekends one time").*
Requirements	Meals should not require too much effort from the user; Introduce an option to shop the ingredients needed for the meal online.
Values	Users should not feel like the design of the app is prejudiced against them
Iteration 4	
Data	As previously + data on the exercises they performed during the previous week, a list of products that can be used to cook the meal suggestions, users' photos
Frame	"If the app provides provocative advice to its users, they will reflect on their choices."

Mode of action	As before + introduce the users to their personal AI trainer; update their profile photos (without asking them to upload one); blame them for lack of commitment.
Intended actuation	Reflect on the introduction of the trainer, and their updated profile photos.
Prototype	As before + the new AI trainer, new type of meals, shoppable options, and their profile photo.
Mode of action	As before + propose an increase in their exercise intensity: e.g., when entering the app, a pop-up screen mentions a lack of commitment to the exercises, leading them to a graph describing the performed exercise intensity.
Intended actuation	As before + reflect on the suggested intensive workouts.
Prototype	As before + a graph mentioning a lack of commitment to the exercises, include photos of male yoga instructors.
Actuation	E.g., Noticed only the change of the name of the trainer, noticed their profile photos, but only two participants wondered how the AI has "gotten" their photos.
Outcomes	E.g., some participants feel pressured (due to the app's tone and exercise intensity), others felt very positive about the new functionality (e.g., "*Okay, cool. [...] And if you don't do it, they increase the exercise, there will be more in the next week or something like that. That's really cool I think.*"); participants perceive lack of privacy ("*where did he take my pictures from because I didn't upload the picture. I am scared of where this thing takes the pictures from*"); participants feel like the app is not inclusive (e.g., one participant *wonders why the picture he is being shown is of a male yoga instructor)*; participants perceive the AI guidance as friendly (*"I think it has changed the tone slightly. It talks more to me, or at least I feel so. I think it's quite nice, that it's more addressed to me"*.)

APPENDIX

CHAPTER 7

Table 2 A snippet of the overview on the design process employed during the third temporal bracket[1]

Element	Description
Iteration 1	
Purpose	Help EV owners who cannot charge at home to start charging as they go.
Data	In-depth interviews with the two users about their charging routines, obstacles, their experience of buying a new EV, setting it up, experience with existing infrastructure, the benefits they experience from owning an EV.
Frame	If owners take an active role in teaching the in-car system, they will follow its advice readily and be prepared for changes.
Mode of action	The system actively asks the user to show their current charging behaviours and identify what is important for them.
Intended actuation	The users answer all of the prompts of the in-car system by filming their surroundings, dashboard, and their current routine on how they decide when and for how long they are going to charge.

1 As the visual map in Chapter 7 shows, each iteration through the theoretical model also "housed" a smaller iteration among the elements of mode of action, intended actuation and prototype. However, not all instances of mode of action and intended actuation or prototype are included in this overview for ease of explanation.

Prototype	A low-fidelity mock-up of an in-car system containing multiple prompts delivered by a computer-generated voice guiding the users in explaining their routines.
Observed actuation	E.g., Provided detailed videos for each of the prompts, had difficulty in understanding how to use fast-charging stations, went for a run while the car was charging, interrupted the charging; unsure how much charging would cost them, looked at the percentage of battery left, looked at the range left, failed to find a free public charger, explained the rationale for their choices.
Outcomes	Participants felt anxious over the too many unknowns (e.g., charging costs, finding a charger, whether their attempts to teach the system were successful and what information the system paid attention to).
Values	Support people in doing healthy activities while charging.
Requirements	Provide a clear overview on how much charging is going to cost, free charging options, and clear explanation on how to get to the charging spots (include photos of the charging station).
Iteration 2	
Data	As before + data from the users' in-car systems (e.g., current charge, car's range, the number of kms the user wants to travel in the upcoming week, etc.), and existing infrastructure (e.g., charging stations, fees, and availability) around the users' usual commuting routes.
Frame	If the in-car system provides users with a clear charging plan tailored to their needs for the week, they will be able to charge primarily as they go.
Mode of action	Generate a personalised charging plan for the user based on their needs for the week and provide them with a few details on where they need to go each day and for how long they need to charge.
Intended actuation	The user charges their car according to the personalised plan the in-car system provides.
Prototype	An app that shows users their charging plan for the week while also suggesting potential activities the user could do while the car is charging (e.g., doing the groceries, having lunch, taking a walk in the park).
Observed actuation	Misunderstood the charging plan as something to be used only for long trips and not commuting, pointed at difficulties in charging at the suggested spots. Confused by some of the prototype's UI: e.g., the button and the prompts the system gave).
Outcomes	Feels like the plan is too hectic and unrealistic, the suggestions to charge in the middle of the day are seen as disruptive. Fear for their privacy.
Values	Instil sense of security with sharing their data.

Requirements	Make a clear overview of all aspects of the prototype, reduce the number of actions required by the user.
Iteration 3	
Data	As before + the users' schedule for the week.
Frame	If the in-car system suggests when to charge based on the activities the user performs in a week (e.g., doing the groceries), the users will follow the advice.
Mode of action	Provide a simple charging plan for the week aligned with the user's schedule and pair it with clear information about the suggested charging station and a back-up option.
Intended actuation	Users share their agendas and in-car system data with the app and then follow the provided charging plan as prescribed.
Prototype	An updated version of the app used in the previous iteration + updated buttons, clear charging plan (including photos of the charging stations) and a link to Google Maps directions on how to find the charging station.
Observed actuation	Participant 1 shared their agenda and in-car system data, followed only the first suggested charging, had difficulties finding the charging station, drove around for 15 minutes and when they found it, the charging pole was taken. They did not see the other suggestions for charging stations nearby the one that was taken. Participant 2 refused to share their agenda, did not follow the charging plan because they do not do any of the plan's suggested activities while with the car (e.g., they never use their car to do groceries).
Outcomes	Frustration (over driving around for too long), unclear guidance to the charging stations, misalignment between suggested activities and the ones usually performed, lack of trust and unwillingness to share their data, do not feel prepared.
Values	People should feel in control of the data they share, people's behaviours should change but without using nudging.
Requirements	Reduce the number of actions required by the user, make the backup options more visible, provide easy access to information about the users' charging behaviours, help them to increase their motivation to charge as they go.
Iteration 4	
Data	As before + charging history for the previous month, data on battery degradation for both users' car models, data on charging prices at the charging stations users usually use.
Frame	If the in-car system shows the users the effect of their charging overnight behaviours on their car battery, they will be willing to (at least partially) adopt charging as they go.

Mode of action	Show three scenarios to the users on how their charging behaviour affects their battery degradation rate, their monetary and time spending, as well as the rate with which their waiting time will increase if every EV owner keeps charging overnight.
Intended actuation	Users interact with the dashboard and its three scenarios. They decide to incorporate more charging as they go.
Prototype	A mid-level fidelity dashboard providing three different scenarios that visualise and contextualise the potential battery degradation rate, the estimated waiting times and the money and time users spend on charging.
Observed actuation	Both participants went through the three different scenarios, barely paid attention to the estimated waiting times, and spent most time on coming up with different explanations about the factors that influence the battery degradation rate. Participant 2 changed behaviour drastically (i.e., provided immediate unrestricted access to their calendar).
Outcomes	Do not feel sufficiently prepared, confusion over how the estimations were calculated, speculations about the potential ways in which the car warranty can be tricked by intentionally degrading the battery faster, deciding to use their car more often instead of their bike in order to charge as they go, behavioural change (i.e., provides full access to their data). Both participants increase the number of times of charging as they go.
Values	it shouldn't be that easy for the users to get scared and give the app full access to their agenda and data.
Requirements	Include only concepts for which the users already have a frame of reference, provide information on how estimations are calculated; make sure that the battery degradation is not shown in a negative light.
Iteration 5	
Data	As before + data on battery degradation for both users' car models, data on charging prices at the charging stations users usually use.
Frame	If the in-car system involves the user in the creation of their charging plan, they will feel prepared and charge as they go.
Mode of action	Help the user build their charging plan for the week by providing personalised suggestions and clear instructions to find the charging stations.
Intended actuation	The user helps the system to build the weekly charging plan and do not charge overnight beforehand.
Prototype	An app building on the one introduced in Iteration 3 + an option to build their own plan, provide an overview on options for the day, provide information on distance from destination, type of charger, amount of time for full charge, walking route.

Observed actuation	Participant charged overnight, checked specifically whether the proposed charging stations were behind a gate, overslept and did not have time to go to the planned charging spot, charged as they go on a different, unplanned by the system, location.
Outcomes	Distrust in the walking distance estimation and price estimation, feeling unprepared.
Values	none
Requirements	Suggest only easily accessible charging options (e.g., no gated stations), visualise the provided estimations, users shouldn't spend more than 10 minutes on finding a charger and walking to their final destination
Iteration 6	
Data	As before + the data generated from the previous iteration as well as the latest data from the users' in-car systems.
Frame	If the in-car system contextualises their charging behaviour, it will make the user willing to increase the percentage of times they charge as they go.
Mode of action	Contextualise battery degradation and charging behaviour.
Intended actuation	The user sees the statistics of their charging behaviour and selects a charging goal for the following month.
Prototype	A mid-level fidelity dashboard building on the one used in Iteration 4 + contextualisation of their potential battery degradation mode, the estimated waiting times and the money and time they spend on charging.
Observed actuation	Went through the dashboard and could not select a goal to slow down their battery degradation rate, frustration that it was becoming more difficult to find a free spot to charge, did not have a problem leaving their car plugged-in without charging,
Outcomes	Refusal to reduce overnight charging and take up chargers they don't use during the day, difficulty to plan, difficulty to imagine what activities need to be taken in the long term to ensure battery health, devise strategies on how to increase the amount of charging as they go.
Values	Support people in planning for the long-term health of their battery.
Requirements	Include insights on how to improve their battery health in the short term.
Frame	If the in-car system supports people to imagine how to plan for their long-term battery health, they will charge more as they go.
Frame	If the in-car system allows people to unplug each other's cars when they are done charging, then they will not occupy chargers they do not currently use.
Purpose	Ensure EV owners do not occupy chargers they do not currently use.

REFERENCES

A

ADCU. (2021). Uber drivers take unprecedented international legal action to demand their data. Retrieved 2023, from The App Drivers and Couriers Union: https://www.adcu.org.uk/news-posts/uber-drivers-take-unprecedented-international-legal-action-to-demand-their-data/

Alexander, C. (1964). *Notes on the Synthesis of Form* (Vol. 5). Harvard university Press.

Akin, Ö. & Akin, C. (1996). Frames of reference in architectural design: analysing the hyperacclamation (A-h-a!). Design Studies, 17, 341-361.

Andrade, F. R., Mizoguchi, R., & Isotani, S. (2016, June). The bright and dark sides of gamification. In *International conference on intelligent tutoring systems* (pp. 176-186). Springer, Cham.

Ardila, D., Kiraly, A.P., Bharadwaj, S. *et al.* End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat Med* 25, 954–961 (2019). https://doi.org/10.1038/s41591-019-0447-x

Amnesty International. (2021). *Dutch childcare benefit scandal an urgent wake-up call to ban racist algorithms*. Retrived from: https://www.amnesty.org/en/latest/news/2021/10/xenophobic-machines-dutch-child-benefit-scandal/.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.

ANP. (2021, February 19). *Belastingdienst krijgt miljoenenboete om toeslagenaffaire*: https://www.trouw.nl/binnenland/belastingdienst-krijgt-miljoenenboete-om-toeslagenaffaire~bffbd871/

Artificial Intelligence Act (2021). Proposal for a regulation of the European Parliament and the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. EUR-Lex-52021PC0206. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206&

B

Ball, L. J., & Christensen, B. T. (2009). Analogical reasoning and mental simulation in design: two strategies linked to uncertainty resolution. *Design Studies*, *30*(2), 169-186.

Baraniuk, C. (2017). *The 'creepy Facebook AI' story that captivated the media*. Retrieved from BBC: https://www.bbc.com/news/technology-40790258

BBC. (2015). *Google apologises for Photos app's racist blunder*. Retrieved from BBC: https://www.bbc.com/news/technology-33347866

BBC. (2019). *Katie Bouman: The woman behind the first black hole image*. Retrieved from BBC: https://www.bbc.com/news/science-environment-47891902

BBC. (2021*). Facebook apology as AI labels black men 'primates'*. Retrieved from BBC: https://www.bbc.com/news/technology-58462511

BBC. (2021*). Twitter finds racial bias in image-cropping AI*. Retrieved from BBC: https://www.bbc.com/news/technology-57192898

Beaudouin-Lafon, M., & Mackay, W. E. (2009). Prototyping tools and techniques. In Human-Computer Interaction (pp. 137-160). CRC Press.

Bedingfield, W. (2020*). Everything that went wrong with the botched A-Levels algorithm*. Retrieved from Wired: https://www.wired.co.uk/article/alevel-exam-algorithm

Bengio, Y. (2017). The consciousness prior. arXiv preprint arXiv:1709.08568.

Bengio, Y. (2019). *Yoshua Bengio: From System 1 Deep Learning to System 2 Deep*

Learning (NeurIPS 2019). Retrieved from YouTube: https://www.youtube.com/watch?v=T3sxeTgT4qc&list=PLPkuuw9ZRIXj2kWFRYhNqnzdbJKOf6yrE&index-=31&t=3s.

Benjamin, M., Buehler, K., Dooley, R., & Zipparo, P. (2021). *What the draft European Union AI regulations mean for business*. Retrieved from McKinsey: https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/what-the-draft-european-union-ai-regulations-mean-for-business.

Bhagavatula, C., Bras, R. L., Malaviya, C., Sakaguchi, K., Holtzman, A., Rashkin, H., ... & Choi, Y. (2019). Abductive commonsense reasoning. arXiv preprint arXiv:1908.05739.

Blomkvist, J., & Holmlid, S. (2011). Existing prototyping perspectives: considerations for service design. Nordes, (4).

Bloomberg Technology. (2022). *Google Engineer on His Sentient AI Claim*. Retrieved from YouTube: https://www.youtube.com/watch?v=kgCUn4fQTsc

Bostrom, N., & Yudkowsky, E. (2014). The ethics of artificial intelligence. The Cambridge handbook of artificial intelligence, 1, 316-334.

Boztepe, S. (2007). User value: Competing theories and models. International journal of design, 1(2).

Britanica. (n/d). Ethics. Retrieved from Britanica: https://www.britannica.com/topic/ethics-philosophyBrown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.

Buchanan, R. (1992). Wicked problems in design thinking. *Design issues*, *8*(2), 5-21.

Buchenau, M., & Suri, J. F. (2000, August). Experience prototyping. In Proceedings of the 3rd conference on Designing interactive systems: processes, practices, methods, and techniques (pp. 424-433).

Buijsman, S. (2023). Navigating fairness measures and trade-offs. AI and Ethics, 1-12.

Bylander, T., Allemang, D., Tanner, M. C., & Josephson, J. R. (1991). The computational complexity of abduction. *Artificial intelligence*, *49*(1-3), 25-60.

C

Calabretta, G., & Kleinsmann, M. (2017). Technology-driven evolution of design practices: envisioning the role of design in the digital era. *Journal of Marketing Management*, *33*(3-4), 292-304.

Callaway, E. (2020). 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. *Nature*, *588*(7837), 203-205.

Campos, D. G. (2011). On the distinction between Peirce's abduction and Lipton's inference to the best explanation. *Synthese*, *180*(3), 419-442.

Cash, P., & Gonçalves, M. (2017). Information-triggered co-evolution: a combined process perspective. In *Analysing design thinking: Studies of cross-cultural co-creation* (pp. 501-520). CRC Press.

Cash, P. J. (2018). Developing theory-driven design research. *Design Studies*, *56*, 84-119.

Carlile, P. R. (2002). A pragmatic view of knowledge and boundaries: Boundary objects in new product development. *Organization science*, *13*(4), 442-455.

Chancellor, S., Baumer, E. P., & De Choudhury, M. (2019). Who is the" Human" in Human-Centered Machine Learning: The Case of Predicting Mental Health from Social Media. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1-32.

Chee, F. Y. (2021). *EU seeks global standards for AI, civil rights groups fret*. Retrieved from Reuters: https://www.reuters.com/world/china/eu-aims-set-global-standards-ai-fines-violations-2021-04-21/

Christensen, B. T., & Abildgaard, S. J. J. (2016). DTRS11 Technical Report v. 2.0 (date: 13.05. 2016). *CBS Working Paper Series*.

Christensen, B. T., & Schunn, C. D. (2009). The role and impact of mental simulation in design. Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition, 23(3), 327-344.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, *5*(2), 153-163.

Cisse, M., Adi, Y., Neverova, N., & Keshet, J. (2017). Houdini: Fooling deep structured prediction models. *arXiv preprint arXiv:1707.05373*.

Clark, A., & Chalmers, D. (1998). The extended mind. *analysis*, *58*(1), 7-19.

Clifford, C. (2017). *Hundreds of A.I. experts echo Elon Musk, Stephen Hawking in call for a ban on killer robots*. Retrieved from CNBC: https://www.cnbc.com/2017/11/08/ai-experts-join-elon-musk-stephen-hawking-call-for-killer-robot-ban.html

CNN . (2021). *Zillow to exit its home buying business, cut 25% of staff*. Retrieved from CNN Business: https://edition.cnn.com/2021/11/02/homes/zillow-exit-ibuying-home-business/index.html

Confessore, N. (2018). *Cambridge Analytica and Facebook: The Scandal and the Fallout So Far*. Retrieved from New York Times: https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html

Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*.

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.

Crawford, K. et al. The AI Now report: The Social and Economic Implications of

Artificial Intelligence Technologies in the Near-term. https://ainowinstitute.org/

AI_Now_2016_Report.pdf (2016).

Crilly, N. (2021). The Evolution of "Co-evolution"(Part I): Problem Solving, Problem Finding, and Their Interaction in Design and Other Creative Practices. *She Ji: The Journal of Design, Economics, and Innovation*, *7*(3), 309-332.

Cross, N. (1990). The nature and nurture of design ability. *Design studies*, *11*(3), 127-140.

Cross, N. (2000). Engineering design methods: strategies for product design. John Wiley & Sons.

Cross, N. (2001). Design cognition: Results from protocol and other empirical studies of design activity. In *Design knowing and learning: Cognition in design education* (pp. 79-103). Elsevier science.

D

Dastin, J. (2018). *Amazon scraps secret AI recruiting tool that showed bias against women*. Retrieved from Reuters: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

Deep Mind, n.d., *About*. Retrieved from DeepMind: https://www.deepmind.com/about

de Reuver, M., van Wynsberghe, A., Janssen, M., & van de Poel, I. (2020). Digital platforms and responsible innovation: expanding value sensitive design to overcome ontological uncertainty. *Ethics and Information Technology*, 1-11.

de Zwart, F. (2015). Unintended but not unanticipated consequences. *Theory and Society*, *44*(3), 283-297.

Dewey, J. (1891). Moral theory and practice. *The International Journal of Ethics*, *1*(2), 186-203.

Dinar, M., Shah, J. J., Cagan, J., Leifer, L., Linsey, J., Smith, S. M., & Hernandez, N. V. (2015). Empirical studies of designer thinking: past, present, and future. *Journal of Mechanical Design*, *137*(2), 021101.

Dixon, L., Li, J., Sorensen, J., Thain, N., & Vasserman, L. (2018, December). Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 67-73).

Doke, E. R. (1990). An industry survey of emerging prototyping methodologies. *Information & Management*, *18*(4), 169-176.

Dong, A., Garbuio, M., & Lovallo, D. (2016a). Generative sensing: A design perspective on the microfoundations of sensing capabilities. *California Management Review*, *58*(4), 97-117.

Dong, A., Garbuio, M., & Lovallo, D. (2016b). Generative sensing in design evaluation. *Design Studies*, *45*, 68-91.

Dong, A., Lovallo, D., & Mounarath, R. (2015). The effect of abductive reasoning on concept selection decisions. *Design studies*, *37*, 37-58.

Dong, A., MacDonald, E., Christensen, B. T., Ball, L. J., & Halskov, K. (2017). From observations to insights: The hilly road to value creation. *Analysing Design Thinking: Studies of Cross-Cultural Co-Creation, Taylor & Francis, London*, 465-482.

Dorst, K., & Cross, N. (2001). Creativity in the design process: co-evolution of problem–solution. *Design studies*, *22*(5), 425-437.

Dorst, K. (2004). The Problem of Design Problems. Design Thinking Research Symposium 6, 4 (Creativity and Cognition Studio Press), 135–147.

Dorst, K. (2006). Design problems and design paradoxes. *Design issues*, *22*(3), 4-17.

Dorst, K. (2011). The core of 'design thinking' and its application. *Design studies*, *32*(6), 521-532.

Dorst, K., & Hansen, C. T. (2011). Modeling paradoxes in novice and expert design. In DS 68-2: Proceedings of the 18th International Conference on Engineering Design (ICED 11), Impacting Society through Engineering Design, Vol. 2: Design Theory and Research Methodology, Lyngby/Copenhagen, Denmark, 15.-19.08. 2011 (pp. 142-150).

Dorst, K., & Dijkhuis, J. (1995). Comparing paradigms for describing design activity. *Design studies*, *16*(2), 261-274.

Dorst, K., & Reymen, I. M. M. J. (2004). Levels of expertise in design education. In DS 33: Proceedings of E&PDE 2004, the 7th International Conference on Engineering and Product Design Education, Delft, the Netherlands, 02.-03.09. 2004.

Dove, G., & Fayard, A. L. (2020, April). Monsters, Metaphors, and Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1-17).

Dreyfus, H. L., & Dreyfus, S. E. (2005). Peripheral vision: Expertise in real world contexts. *Organization studies*, *26*(5), 779-792.

Duffy, C. (2019). *Apple co-founder Steve Wozniak says Apple Card discriminated against his wife*. Retrieved from CNN Business: https://edition.cnn.com/2019/11/10/business/goldman-sachs-apple-card-discrimination/index.html

E

Eden, A. H., Steinhart, E., Pearce, D., & Moor, J. H. (2012). Singularity hypotheses: An overview. *Singularity hypotheses*, 1-12.

Eisenhardt, K. M. (1989). Building theories from case study research. *Academy of management review*, *14*(4), 532-550.

Elverum, C. W., & Welo, T. (2016). Leveraging prototypes to generate value in the con-

cept-to-production process: a qualitative study of the automotive industry. *International Journal of Production Research*, *54*(10), 3006-3018.

Eppler, M. J., Hoffmann, F., & Pfister, R. (2011). Rigor and relevance in management typologies: Assessing the quality of qualitative classifications.

Evtimov, I., Eykholt, K., Fernandes, E., Kohno, T., Li, B., Prakash, A., ... & Song, D. (2017). Robust physical-world attacks on deep learning models. *arXiv preprint arXiv:1707.08945*.

F

Fagone, J. (2021). *The Jessica Simulation: Love and loss in the age of A.I.*. Retrieved from San Francisco Chronicle: https://www.sfchronicle.com/projects/2021/jessica-simulation-artificial-intelligence/

Fails, J. A., & Olsen Jr, D. R. (2003, January). Interactive machine learning. In *Proceedings of the 8th international conference on Intelligent user interfaces* (pp. 39-45).

Flanagan, M., Howe, D. C., & Nissenbaum, H. (2008). Embodying values in technology: Theory and practice. *Information technology and moral philosophy*, *322*.

Fiebrink, R., & Gillies, M. (2018). Introduction to the special issue on human-centered machine learning.

Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*, *1*(1).

Fogg, B. J. (2019). Fogg behavior model. URL: https://behaviormodel. org (visited on 12/14/2020).

Friedman, B., Kahn, P., & Borning, A. (2002). Value sensitive design: Theory and methods. *University of Washington technical report*, (2-12).

Friedman, B. (1996). Value-sensitive design. *interactions*, *3*(6), 16-23.

G

Gehman, J., Glaser, V. L., Eisenhardt, K. M., Gioia, D., Langley, A., & Corley, K. G. (2018). Finding theory–method fit: A comparison of three qualitative approaches to theory building. *Journal of Management Inquiry*, *27*(3), 284-300.

GPT-3 (2020). *A robot wrote this entire article. Are you scared yet, human?.* Retrieved from The Guardian: https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3.

Goldschmidt, G. (2003). The backtalk of self-generated sketches. *Design issues*, *19*(1), 72-88.

Goldstein: https://twitter.com/tomgoldsteincs/status/1600196995389366274?lang=en

Green, E.M., van Mourik, R., Wolfus, C. *et al.* Machine learning detection of obstructive hypertrophic cardiomyopathy using a wearable biosensor. *npj Digit. Med.* 2, 57 (2019). https://doi.org/10.1038/s41746-019-0130-

Greenberg, A. (2019). *Hackers Can Use Lasers to 'Speak' to Your Amazon Echo or Google Home*. Retrieved from Wired: https://www.wired.com/story/lasers-hack-amazon-echo-google-home/

Greene, D., Hoffmann, A. L., & Stark, L. (2019, January). Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning. In *Proceedings of the 52nd Hawaii international conference on system sciences*.

Gonçalves, M., & Cash, P. (2021). The life cycle of creative ideas: Towards a dual-process theory of ideation. *Design Studies*, *72*, 100988.

Gunning, D. (2017). Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web*, *2*, 2.

Gutierrez, O. (1989). Prototyping techniques for different problem contexts. *ACM SIGCHI Bulletin*, *20*(SI), 259-264.

H

Hagras, H. (2018). Toward human-understandable, explainable AI. *Computer*, *51*(9), 28-36.

Halloran, J., Hornecker, E., Stringer, M., Harris, E., & Fitzpatrick, G. (2009). The value of values: Resourcing co-design of ubiquitous computing. *CoDesign*, *5*(4), 245-273.

Harwell, D. (2022). *He used AI to win a fine-arts competition. Was it cheating?.* Retrieved from The Washington Post: https://www.washingtonpost.com/technology/2022/09/02/midjourney-artificial-intelligence-state-fair-colorado/.

Hatchuel, A. (2001). Towards Design Theory and expandable rationality: The unfinished program of Herbert Simon. *Journal of management and governance*, *5*(3/4), 260-273.

Hatchuel, A., Le Masson, P., Reich, Y., & Subrahmanian, E. (2018). Design theory: a foundation of a new paradigm for design science and engineering. *Research in Engineering Design*, *29*(1), 5-21.

Harari, Yuval Noah. (2015). Homo Deus: A Brief History of Tomorrow. New York: Harper-Collins.

Heaven, W. D. (2020b). *Our weird behavior during the pandemic is messing with AI models*. Retrieved from MIT Technology Review: https://www-technologyreview-com.cdn.ampproject.org/c/s/www.technologyreview.com/2020/05/11/1001563/covid-pandemic-broken-ai-machine-learning-amazon-retail-fraud-humans-in-the-loop/amp/

Heaven, W. D. (2020a). *OpenAI's new language generator GPT-3 is shockingly good—and completely mindless*. Retrieved from MIT Technology Review: https://www.technologyreview.com/2020/07/20/1005454/openai-machine-learning-language-generator-gpt-3-nlp/

Heaven, W. D. (2022). *The new version of GPT-3 is much better behaved (and should be less toxic)*. Retrieved from MIT Technology Review: https://www.technologyreview.com/2022/01/27/1044398/new-gpt3-openai-chatbot-language-model-ai-toxic-misinformation/

Hendriks, R. (2022). *Deepfake van politie roept vraag op over 'digitale onsterfelijkheid'*. Retrieved from NOS Nieuws: https://nos.nl/artikel/2429954-deepfake-van-politie-roept-vraag-op-over-digitale-onsterfelijkheid

Hern. A. (2020). *Twitter apologises for 'racist' image-cropping algorithm*. Retrieved from The Guardian: https://www.theguardian.com/technology/2020/sep/21/twitter-apologises-for-racist-image-cropping-algorithm

Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M., & Wallach, H. (2019, May). Improving fairness in machine learning systems: What do industry practitioners need?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1-16).

Horváth, I. (2016). Theory building in experimental design research. In *Experimental design research* (pp. 209-231). Springer, Cham.

Hsieh, H. F., & Shannon, S. E. (2005). Three approaches to qualitative content analy-

sis. *Qualitative health research*, *15*(9), 1277-1288.

Hu, K. (2023). ChatGPT sets record for fastest-growing user base - analyst note. Retrieved from Reuters: https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/.

Hunt, E. (2016). *Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter*. Retrieved from The Guardian: https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter.

J

JafariNaimi, N., Nathan, L., & Hargraves, I. (2015). Values as hypotheses: design, inquiry, and the service of values. *Design issues*, *31*(4), 91-104.

Jensen, M. B., Elverum, C. W., & Steinert, M. (2017). Eliciting unknown unknowns with prototypes: Introducing prototrials and prototrial-driven cultures. *Design Studies*, *49*, 1-31.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389-399.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255-260.

K

Kahn,J. (2020). *The polls are wrong. The U.S. presidential race is a near dead heat, this A.I. 'sentiment analysis' tool says*. Retrieved from Fortune: https://fortune.com/2020/10/14/polls-trump-biden-presidential-race-artificial-intelligence/?_hsmi=99626940

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kivits, N. (2022). *Voor de deepfake van Sedar waren maar twee foto's beschikbaar*. Retrieved from: https://www.villamedia.nl/artikel/voor-de-deepfake-van-sedar-waren-maar-twee-fotos-beschikbaar

Kleinnijenhuis, J. (2018). *Belastingdienst werkte ouders die recht hadden op kinderopvangtoeslag bewust tegen*. Retrieved from Trouw: https://www.trouw.nl/nieuws/belastingdienst-werkte-ouders-die-recht-hadden-op-kinderopvangtoeslag-bewust-tegen~bf13daf9/

Kleinnijenhuis, J. (2022). *Topambtenaren logen mogelijk bij toeslagenverhoren, onderzoek naar meineed.* Retrieved from NOS: https://nos.nl/nieuwsuur/artikel/2456145-topambtenaren-logen-mogelijk-bij-toeslagenverhoren-onderzoek-naar-meineed

Komkov, S., & Petiushko, A. (2019). AdvHat: Real-world adversarial attack on ArcFace Face ID system. *arXiv preprint arXiv:1908.08705*.

Koebler, J. (2015). *10 Years Ago Today, YouTube Launched as a Dating Website*. Retrieved from The Motherboard: https://www.vice.com/en/article/78xqjx/10-years-ago-today-youtube-launched-as-a-dating-website.

Kolko, J. (2010). Abductive thinking and sensemaking: The drivers of design synthesis. *Design issues*, *26*(1), 15-28.

Kroll, E., & Koskela, L. (2015). On abduction in design. In *Design Computing and Cognition'14* (pp. 327-344). Springer, Cham.

Kroll, E., & Koskela, L. (2016). Explicating concepts in reasoning from function to form by two-step innovative abductions. *AI EDAM*, *30*(2), 125-137.

Kuhlman, C., Jackson, L., & Chunara, R. (2020). No computation without representation: Avoiding data and algorithm biases through diversity. *arXiv preprint arXiv:2002.11836*.

Kurvinen, E., Koskinen, I., & Battarbee, K. (2008). Prototyping social interaction. *Design Issues*, *24*(3), 46-57.

Kurzweil, R. (2005). The singularity is near: When humans transcend biology. Penguin.

Kuta, S. (2022). *Art Made With Artificial Intelligence Wins at State Fair*. Retrieved from Smithsonian Magazine: https://www.smithsonianmag.com/smart-news/artificial-intelligence-art-wins-colorado-state-fair-180980703/#:~:text=Jason%20Allen%2C%20a%20video%20game,came%20with%20a%20%24300%20prize.

L

Langley, A. N. N., Smallman, C., Tsoukas, H., & Van de Ven, A. H. (2013). Process studies of change in organization and management: Unveiling temporality, activity, and flow. *Academy of management journal*, *56*(1), 1-13.

Lawson, B., & Dorst, K. (2013). *Design expertise*. Routledge.

LeCun, Y. (2022). A path towards autonomous machine intelligence. *preprint posted on*

openreview.

LeCun, Y. & Ng, A. (2023). *Yann LeCun and Andrew Ng: Why the 6-month AI Pause is a Bad Idea*. Retrieved from YouTube: https://www.youtube.com/watch?v=BY9KV8uCtj4

Le Dantec, C. A., Poole, E. S., & Wyche, S. P. (2009, April). Values as lived experience: evolving value sensitive design in support of value discovery. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1141-1150).

Lee, M. S. A., Floridi, L., & Singh, J. (2021). Formalising trade-offs beyond algorithmic fairness: lessons from ethical philosophy and welfare economics. *AI and Ethics*, *1*(4), 529-544.

Levina, N., & Vaast, E. (2005). The emergence of boundary spanning competence in practice: implications for implementation and use of information systems. *MIS quarterly*, 335-363.

Levine, S. (2022, January). Understanding the world through action. In *Conference on Robot Learning* (pp. 1752-1757). PMLR.

Lewis: Mike Lewis, Denis Yarats, Devi Parikh, Dhruv Batra: https://engineering.fb.com/2017/06/14/ml-applications/deal-or-no-deal-training-ai-bots-to-negotiate/

Li, O., Liu, H., Chen, C., & Rudin, C. (2018, April). Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In *Thirty-Second AAAI Conference on Artificial Intelligence*

Lichter, H., Schneider-Hufschmidt, M., & Zullighoven, H. (1994). Prototyping in industrial software projects-bridging the gap between theory and practice. *IEEE transactions on software engineering*, *20*(11), 825-832.

Liedtka, J., Martin, R., & Dew, N. (2007). Abduction: a pre-condition for the intelligent design of strategy. *Journal of Business Strategy*.

Lim, Y. K., Stolterman, E., & Tenenberg, J. (2008). The anatomy of prototypes: Prototypes as filters, prototypes as manifestations of design ideas. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *15*(2), 1-27.

Lipton, Z., Wang, Y. X., & Smola, A. (2018, July). Detecting and correcting for label shift with black box predictors. In *International conference on machine learning* (pp. 3122-3130). PMLR.

M

Magnani, L. (2004). Model-based and manipulative abduction in science. *Foundations of science*, *9*(3), 219-247.

Magnani, L. (2007). Abduction and chance discovery in science. *International Journal of Knowledge-based and Intelligent Engineering Systems*, *11*(5), 273-279.

Magnani, L. (2011). Abduction, reason and science: Processes of discovery and explanation. Springer Science & Business Media.

Magnani, L., Carnielli, W., & Pizzi, C. (Eds.). (2010). Model-based reasoning in science and technology: Abduction, logic, and computational discovery (Vol. 314). Springer.

Maher, M. L. (2000). A model of co-evolutionary design. Engineering with computers, 16, 195-208.

March, L. (1984). The Logic of Design. In N. Cross*, Developments in Design Methodology* (pp. 265-276). Bath, Avon: John Wiley & Sons Ltd.

Mariani, E., Kooijman, F. S. C., Shah, P., & Stoimenova, N. (2021). Prototyping in social VR: anticipate the unanticipated outcomes of interactions between ai-powered solutions and users. *Proceedings of the Design Society*, *1*, 2491-2500.

European Parliament (2020). *Artificial intelligence: threats and opportunities*. Retrieved from: https://www.europarl.europa.eu/news/en/headlines/society/20200918STO87404/artificial-intelligence-threats-and-opportunities

Mao, F. (2020). *Coronavirus panic: Why are people stockpiling toilet paper?.* Retrieved from BBC:: https://www.bbc.com/news/world-australia-51731422.

Meng, J. C. S. (2009). Donald Schön, Herbert Simon and the sciences of the artificial. *Design Studies*, *30*(1), 60-68.

Menold, J., Jablokow, K., & Simpson, T. (2017). Prototype for X (PFX): A holistic framework for structuring prototyping methods to support engineering design. *Design Studies*, *50*, 70-112.

Merton, R. K. (1936). The unanticipated consequences of purposive social action. *American sociological review*, *1*(6), 894-904.

Michaelraj, A. (2009). Taxonomy of physical prototypes: Structure and validation.

Miller, R. A., Pople, H. E., & Myers, J. D. (1985). Internist-I, an experimental computer-based diagnostic consultant for general internal medicine. In *Computer-assisted medical decision making* (pp. 139-158). Springer, New York, NY.

Mingers, J. (2012). Abduction: the missing link between deduction and induction. A comment on Ormerod's 'rational inference: deductive, inductive and probabilistic thinking'. *Journal of the Operational Research Society*, *63*(6), 860-861.

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, 1-7.

Mogensen, P. (1992). Towards a Provotyping Approach in Systems Development. *Scand. J. Inf. Syst.*, *4*(1), 5.

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., & Fernández-Leal, Á. (2022). Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review*, 1-50.

Müller, V. C., & Bostrom, N. (2016). Future progress in artificial intelligence: A survey of expert opinion. In *Fundamental issues of artificial intelligence* (pp. 555-572). Springer, Cham.

N

Nagesh, A. (2021). *The moment QAnon took the person I love most*. Retrieved from BBC: https://www.bbc.com/news/world-us-canada-57369349

Nazar, J. (2013) 14 Famous Business Pivots. Retrieved from Forbes: https://www.forbes.com/sites/jasonnazar/2013/10/08/14-famous-business-pivots/?sh=357dd6ba5797.

Nelson, J., & Menold, J. (2020). Opening the black box: Developing metrics to assess the cognitive processes of prototyping. *Design Studies*, *70*, 100964.

Nersessian, N. J. (2002). The cognitive basis of model-based reasoning in science. *The cognitive basis of science*, 133-153.

Noessel, C. (2017). Designing agentive technology: AI that works for people. Rosenfeld Media.

Norman, D. (2013). The design of everyday things: Revised and expanded edition. Basic books.

Norman, D. A., & Stappers, P. J. (2015). DesignX: complex sociotechnical systems. *She Ji: The Journal of Design, Economics, and Innovation*, *1*(2), 83-106.

NOS Niews (2022). *Gedode Sedar Soares doet in deepfake-video oproep aan dader en getuigen*. Retrieved from: https://nos.nl/artikel/2429868-gedode-sedar-soares-doet-in-deepfake-video-oproep-aan-dader-en-getuigen.

O

Ohlheiser, A. (2021). *The beauty of TikTok's secret, surprising, and eerily accurate recommendation algorithms*. Retrieved from MIT Technology Review: https://www.technologyreview.com/2021/02/24/1017814/tiktok-algorithm-famous-social-media/.

OpenAI, n.d. *About*. Retrieved from: https://www.deepmind.com/about.

Osmanovic Thunström, A. (2022). *We Asked GPT-3 to Write an Academic Paper about Itself—Then We Tried to Get It Published*. Retrieved from Scientific American: https://www.scientificamerican.com/article/we-asked-gpt-3-to-write-an-academic-paper-about-itself-mdash-then-we-tried-to-get-it-published/.

Otto, K., & Wood, K. (2001). Techniques in Reverse Engineering, Systematic Design and New Product Development.

P

Pahl, G., & Beitz, W. (1996). Conceptual design. In *Engineering Design* (pp. 139-198). Springer, London.

Pahl, G., & Wallace, K. (2002). Using the concept of functions to help synthesise solutions. In *Engineering design synthesis* (pp. 109-119). Springer, London.

Parkin: https://www.theguardian.com/technology/2018/sep/08/youtube-stars-burnout-fun-bleak-stressed

Passera, S., Kärkkäinen, H., & Maila, R. (2012). When, how, why prototyping? A practical framework for service development. In *ISPIM Conference Proceedings* (p. 1). The International Society for Professional Innovation Management (ISPIM).

Paul, K. (2023). *Letter signed by Elon Musk demanding AI research pause sparks controversy*. Retrieved from The Guardian: https://www.theguardian.com/technology/2023/mar/31/ai-research-pause-elon-musk-chatgpt

Pedgley, O. (2007). Capturing and analysing own design activity. *Design studies*, *28*(5), 463-483.

Pei, E., Campbell, I., & Evans, M. (2011). A taxonomic classification of visual design representations used by industrial designers and engineering designers. *The Design Journal*, *14*(1), 64-91.

Peirce, C. S. (1994). *Collected papers of Charles Sanders Peirce* (Vols. I- VIII). Harvard University Press.

Piper, K. (2020). *GPT-3, explained: This new language AI is uncanny, funny — and a big deal*. Retrieved from Vox: https://www.vox.com/future-perfect/21355768/gpt-3-ai-openai-turing-test-language.

Porter, J. (2020). *UK ditches exam results generated by biased algorithm after student protests*. Retrieved from The Verge: https://www.theverge.com/2020/8/17/21372045/uk-a-level-results-algorithm-biased-coronavirus-covid-19-pandemic-university-applications.

Purcell, A. T., & Gero, J. S. (1998). Drawings and the design process: A review of protocol studies in design and other disciplines and related research in cognitive psychology. *Design studies*, *19*(4), 389-430.

Q

Quach. (2021). A developer built an AI chatbot using GPT-3 that helped a man speak again to his late fiancée. OpenAI shut it down. Retrieved from: https://www.theregister.com/2021/09/08/project_december_openai_gpt_3/.

Queiroz, J., & Merrell, F. (2005). Abduction: Between subjectivity and objectivity. *Semiotica*, *2005*(153), 1-8.

R

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J. F., Breazeal, C., ... & Wellman, M. (2019). Machine behaviour. *Nature*, *568*(7753), 477-486.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., ... & Sutskever, I. (2021, July). Zero-shot text-to-image generation. In *International Conference on Machine Learning* (pp. 8821-8831). PMLR.

Riedl, M. O. (2019). Human-centered artificial intelligence and machine learning. *Human*

Behavior and Emerging Technologies, *1*(1), 33-36.

Rittel, H. W., & Webber, M. M. (1973). Dilemmas in a general theory of planning. *Policy sciences*, *4*(2), 155-169.

Rizzo, J. (2022). *Who Will Own the Art of the Future?.* Retrieved from Wired: https://www.wired.com/story/openai-dalle-copyright-intellectual-property-art/.

Robbins, S. (2019). AI and the path to envelopment: knowledge as a first step towards the responsible regulation and use of AI-powered machines. *AI & SOCIETY*, 1-10.

Rokeach, M. (1973). *The nature of human values*. New York: Free Press.

Roose, K. (2023). *Bing's A.I. Chat: 'I Want to Be Alive*. Retrieved from The New York Times: https://www.nytimes.com/2023/02/16/technology/bing-chatbot-transcript.html.

Roozenburg, N. F. (2002). Defining synthesis: on the senses and the logic of design synthesis. In *Engineering Design Synthesis* (pp. 3-18). Springer, London.

Roozenburg, N. F., & Eekels, J. (1995). Product design: fundamentals and methods.

Roozenburg, N. F., & Dorst, K. (1998). Describing design as a reflective practice: Observations on Schön's theory of practice. In *Designers* (pp. 29-41). Springer, London.

Roozenburg, N. F. (1993). On the pattern of reasoning in innovative design. *Design Studies*, *14*(1), 4-18.

Rosenman, M. A., & Gero, J. S. (1998). Purpose and function in design: from the socio-cultural to the techno-physical. *Design Studies*, *19*(2), 161-186.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206-215.

Ruf, B., & Detyniecki, M. (2021). Towards the right kind of fairness in AI. *arXiv preprint arXiv:2102.08453*.

Russell, S., & Norvig, P. (2021). Artificial Intelligence: A Modern Approach, Global Edition 4th. *Foundations*, *19*, 23.

S

Saldaña, J. (2013). An introduction to codes and coding. In: *The coding manual for qualitative researchers*. Second Edition. SAGE Publications Ltd, London

Saldaña, J. (2014). Coding and analysis strategies. In *The Oxford handbook of qualitative research*.

Sanders, E. B. N., & Stappers, P. J. (2012). *Convivial design toolbox*.

Satariano, A. & Alba, D. (2020). *Burning Cell Towers, Out of Baseless Fear They Spread the Virus*. Retrieved from The New York Times: https://www.nytimes.com/2020/04/10/technology/coronavirus-5g-uk.html.

Satariano, A. (2020). *British Grading Debacle Shows Pitfalls of Automating Government*. Retrieved from the New York Times: https://www.nytimes.com/2020/08/20/world/europe/uk-england-grading-algorithm.html.

Schiffer, Z. (2021). *Google fires second AI ethics researcher following internal investigation*. Retrieved from The Verge: https://www.theverge.com/2021/2/19/22292011/google-second-ethical-ai-researcher-fired.

Schön, D. A. (1984). The reflective practitioner: How professionals think in action (Vol. 5126). Basic books.

Shead, S. (2020). *How a computer algorithm caused a grading crisis in British schools*. Retrieved from CNBC: https://www.cnbc.com/2020/08/21/computer-algorithm-caused-a-grading-crisis-in-british-schools.html.

Shulman, C., & Bostrom, N. (2012). How hard is artificial intelligence? Evolutionary arguments and selection effects. *Journal of Consciousness Studies*, *19*(7-8), 103-130.

Siebert, L. C., Mercuur, R., Dignum, V., Hoven, J. V. D., & Jonker, C. (2020). Improving Confidence in the Estimation of Values and Norms. *arXiv preprint arXiv:2004.01056*.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., ... & Hassabis, D. (2017). Mastering the game of go without human knowledge. *Nature*, *550*(7676), 354-359.

Simon, H. A. (1977). The structure of ill-structured problems. In *Models of discovery* (pp. 304-325). Springer, Dordrecht.

Simon, H. (1996). The Sciences of the Artificial 3rd Edition Massachusetts.

Singh. (2021, October 3). What we can learn from China's proposed AI regulations. VentureBeat. https://venturebeat.com/2021/10/03/what-we-can-learn-from-chinas-proposed-ai-regulations/.

Spangler. (2022, January 11). MrBeast Named Highest-Earning YouTube Star of 2021, With $54 Million in Pretax Income. https://variety.com/2022/digital/news/mrbeast-highest-earning-youtube-star-54-million-1235154580/.

Spee, A. P., & Jarzabkowski, P. (2009). Strategy tools as boundary objects. *Strategic organization*, *7*(2), 223-232.

Srivastava, B., & Rossi, F. (2018, December). Towards composable bias rating of AI services. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 284-289).

Stappers, P., & Giaccardi, E. (2017). Research through design. *The encyclopedia of human-computer interaction*, *2*.

Star, S. L., & Griesemer, J. R. (1989). Institutional ecology,translations' and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Social studies of science*, *19*(3), 387-420.

Sveiby, K. E., Gripenberg, P., Segercrantz, B., Eriksson, A., & Aminoff, A. (2009, June). Unintended and undesirable consequences of innovation. In *XX ISPIM conference, The Future of Innovation. Vienna*.

Sydell, L. (2017*). Facebook, Google Spread Misinformation About Las Vegas Shooting. What Went Wrong?*. Retrieved from: https://www.npr.org/sections/alltechconsidered/2017/10/03/555320532/facebook-google-spread-misinformation-about-las-vegas-shooting-what-went-wrong?t=1638803622306.

T

Takeda, H. (1994, June). Abduction for design. In *Formal design methods for CAD* (pp. 221-243).

Takeda, H., Yoshioka, M., & Tomiyama, T. (2001). Roles and formalization of abduction in synthesis. In *Proceedings of the Annual Conference of JSAI 15th (2001)* (pp. 30-30). The Japanese Society for Artificial Intelligence.

ten Bhömer, M., Brouwer, C. E., Tomico, O., & Wensveen, S. A. G. (2013, June). Interactive prototypes in the participatory development of product-service systems. In *Proceedings of the 3rd Participatory Innovation Conference. Lahti, Finland* (pp. 36-42).

ten Bhömer, M. (2016). Designing embodied smart textile services: the role of prototypes for project, community and stakeholders.

Tiku, N. (2022). *The Google engineer who thinks the company's AI has come to life*. Retrieved from The Washington Post: https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/.

Timberg, Dwoskin & Albergotti (2021). Inside Facebook, Jan. 6 violence fueled anger, regret over missed warning signs. https://www.washingtonpost.com/technology/2021/10/22/jan-6-capitol-riot-facebook/.

U

Ulrich, K. T., & Eppinger, S. D. (2012). Product Design and Development-–McGraw-Hill Irwin.

University of Padua. Retrieved from YouTube: https://www.youtube.com/watch?v=__vB-MCSeB64

V

Van de Poel, I. (2009). Values in engineering design. In *Philosophy of technology and engineering sciences* (pp. 973-1006). North-Holland.

van de Poel, I. (2020). Embedding Values in Artificial Intelligence (AI) Systems. *Minds and Machines*, *30*(3), 385-409.

van der Ploeg. (2021). The Meaning in Hiring. TU Delft Repository. https://repository.tudelft.nl/islandora/object/uuid:98459ea5-fc0a-498e-a6d9-0615b938442a?collection=education.

van Onselen, L. (2022). Becoming a design professional through coping with value-based conflicts in collaborative design practice.

van Oorschot, R., Snelders, D., Kleinsmann, M., & Buur, J. (2022). Participation in design research. Design Studies, 78, 101073.

Vargo, S. L., Maglio, P. P., & Akaka, M. A. (2008). On value and value co-creation: A service systems and service logic perspective. *European management journal*, *26*(3), 145-152.

Vinge, V. (1993). The coming technological singularity: How to survive in the post-human era. *Science Fiction Criticism: An Anthology of Essential Writings*, 352-363.

Vincent, J. (2018, January 12). Google's racist gorillas show that AI is still lacking. Retrieved from The Verge: https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai.

Visser, F. S., Stappers, P. J., Van der Lugt, R., & Sanders, E. B. (2005). Contextmapping: experiences from practice. *CoDesign*, *1*(2), 119-149.

Voûte, E., Stappers, P. J., Giaccardi, E., Mooij, S., & van Boeijen, A. (2020). Innovating a large design education program at a university of technology. *She Ji: The Journal of Design, Economics, and Innovation*, *6*(1), 50-66.

Vriens. (2021, January 7). A Situated Exploration in the World of Incentivized Health Insurance. Retrieved from DEUS AI: https://www.deus.ai/post/a-situated-exploration-in-the-world-of-incentivized-health-insurance.

W

Wacker, J. G. (1998). A definition of theory: research guidelines for different theory-building research methods in operations management. *Journal of operations management*, *16*(4), 361-385.

Wacker, J. G. (2008). A conceptual understanding of requirements for theory-building research: Guidelines for scientific theory building. *Journal of Supply Chain Management*, *44*(3), 5-15.

Waller, R. A. (1986). WOMEN AND THE TYPEWRITER DURING THE FIRST FIFTY YEARS, 1873-1923. *Studies in Popular Culture*, *9*(1), 39-50.

Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019, May). Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1-15).

Wang, F., Rudin, C., Mccormick, T. H., & Gore, J. L. (2019). Modeling recovery curves with application to prostatectomy. *Biostatistics*, *20*(4), 549-564.

White, R. W., Doraiswamy, P. M., & Horvitz, E. (2018). Detecting neurodegenerative disorders from web search signals. *NPJ digital medicine*, *1*(1), 1-4.

Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., & He, L. (2022). A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*.

Y

Yang, Q., Steinfeld, A., Rosé, C., & Zimmerman, J. (2020, April). Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. In *Proceedings of the 2020 chi conference on human factors in computing systems* (pp. 1-13).

Z

Zeng, Y., & Cheng, G. D. (1991). On the logic of design. *Design Studies*, *12*(3), 137-141.

Zhang, Y., Bellamy, R., & Varshney, K. (2020, February). Joint optimization of AI fairness and utility: A human-centered approach. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 400-406).

Zhou, J., Park, C.Y., Theesfeld, C.L. et al. Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk. Nat Genet 51, 973–980 (2019). https://doi.org/10.1038/s41588-019-0420-0.

Zimmerman, J., Stolterman, E., & Forlizzi, J. (2010, August). An analysis and critique of Research through Design: towards a formalization of a research approach. In *proceedings of the 8th ACM conference on designing interactive systems* (pp. 310-319).

ACKNOWLEDGEMENTS

When I was 15, my philosophy teacher, Vassia Velkova, introduced existentialism to me in preparation for my inaugural Olympiad. One of the very first things we discussed was Sartre's "*Man is condemned to be free*"[1]. Those six words sounded like a revelation I did not even know I was after. A paradox that seized my imagination and anchored all my teenage angst. For finally, I had the validation that I had no choice but to be free. To make my own decisions, carve my own path and, occasionally, grapple with the burden of the freedom I sought. And while it is certainly daunting to be the sole custodian of one's destiny, it is also immensely empowering to know that you, alone, define your life. An aspiration equally intoxicating and disconcerting.

Now that I am twice as old, I am willing to make the concession that not everything is a direct result of my own volition. In fact, the mounting evidence in the fields of Neuroscience and Behavioural Psychology cast a well-founded doubt on the absoluteness of free will. Our environment mars our perceptions and even the explanations we often give others and ourselves. This is where the

1 Sartre, J. P. (2022). ***Being and nothingness: An essay in phenomenological ontology***. Taylor & Francis.

realm of luck comes in. And I have always been lucky with my environments. Especially when it comes to the people around me.

Above all, I have been incredibly fortunate to have the family I have. This dissertation is for them. To my wonderful, creative, fun, and imaginative mother. Thank you for teaching me how to dream and making me believe I can fight the evil mice even when all odds are stacked against me. To my father, and my moral compass, who has always kept me grounded. Thank you for showing me the importance of never compromising with one's values and dignity. Even when the machine tries to swallow you whole, à la Charlie Chaplin in Modern Times. To my incredible sister and best friend. Thank you for being there to pick me up when I cannot do it myself. I have loved you from the very moment I saw you. And to my grandmother. I often struggle to pick just a few words to describe you. You are so many things! However, I think *unwavering* comes close. Thank you for always believing in me, especially in those moments when my courage wavers.

I have also been lucky with my "adopted" family. Maaike – words are not enough! You have done so much for me. Thank you for giving me the opportunity to research obscure topics at first and then readily leading my supervisory team. Thank you for the freedom you gave me and oftentimes facilitated for me. Milene, I do not recall if I have ever told you that the very first PhD defence I saw was yours. So, it felt natural and somewhat serendipitous when you agreed to join the team without any hesitation. Thank you for helping me see the forest for the trees. To both of you, thank you for all of your mental and emotional support during these years of "doing a PhD in my free time". For balancing my propensity to overthink and underexplain. And for the much needed "glue" that enabled me to stick to a framing of my work.

I am also at an immense debt to Dirk and Christine who were part of my PhD supervisory team at the beginning. For without them the dissertation would not exist either. Christine, (*my fairy god mother*) thank you for being there for me and cultivating my love for research. Dirk, thank you for humouring my

oftentimes non-sensical questions and musings. Above all, thank you both for always asking thought-provoking questions.

I also owe special thanks to the people who supported me while gathering the empirical insights used in this dissertation. To Nicole and Alexandra for providing me with a real-world case study and thinking along with me. Thank you – I truly enjoyed working with you! To Tomas and Mieke for graciously offering their time and letting me tag along on their EV journeys. To my students, who always asked thought-provoking questions and somehow managed to help me untangle my severely tangled thoughts. Special thanks goes to Priyanka, Finn, Elena, Katja, Roberto, Derek, Marijn, Lee, Tobias, Guusje, Chantal and Bibian for contributing to this doctoral research in more ways than one. I hope you learned from me as much as I have learned from you. Thank you!

My gratitude also extends to the committee members who graciously agreed to make time for reading and critically assessing my work. Professors Bozzon, Cash, Dong, Lloyd, and Roeser, I am honoured to have had the chance to share my work with you! Thank you for your thoughtful feedback and readiness to provide additional insights that made my work better. Your efforts and guidance significantly improved the quality of this thesis.

Another batch of “thanks” goes to the people who indirectly shaped my work. Valeria and Ehsan – thank you for always making me question what design is. To my Designtology fellows – Eric Jan and Jeroen (my paranymph). Thank you for the multiple discussions and more-than-occasional venting about the current state of Design. And obviously, for all of the abduction vans, post-it guns, and Napoleon-syndrome quandaries. To Jan-Maarten (my other paranymph), who always has a critical question in store that challenges my explanations. Jeroen and Jan-Maarten, thank you for agreeing to be my paranymphs and stand with me while I finalise and defend this work.! Thank

you also to Quiel, Rebecca, Sicco, Mieke, Sander, Vicky, Danielle, Hanneke, and Leandra for making my time in the university an enjoyable experience.

For the last two years, special thanks goes to my DEUS colleagues for their support. Thank you for making a rather busy period a much more enjoyable ride! Above all, to Nathalie, Ron, Nick, Richard, and Danny for trusting me to explore ways, in which we can further extend some of the ideas presented in this thesis. To Nathalie, in particular, who brought me into the company and made (and continues to make) space for me in more ways than one. Nathalie, thank you for the many thought-provoking and, occasionally random, conversations (about everything and nothing), guidance and trust. I will be forever grateful to Kari for introducing us! Special thanks also goes to Ron for guiding me in packaging abstract ideas in an appealing way.

The spotlight also goes to my team for putting up with my propensity to derail conversations and their readiness to provide to-the-point feedback to any topic I bring up: Alex, Anibal, Diego, Inês, Jose, Juan, Marta, Miguel, Xes, Carlos and Rodrigo. And, especially, to Juan Carlos (for being my mental support, an on-demand therapist and someone who understands the gravity of always knowing where his towel is), to Cris (for always being open to a crazy idea), to Esdras (for all the sarcastic and poignant observations on the human condition, which, on occasion, he allowed me to join), to Junior (I get instantly happier just being around you), and to Gus (for always being there to help, answer all my stupid questions, and tolerate my ill-formed jokes). Thank you all for the never-ending banana-related discussions, Mars-bars-induced existential musings, duck songs, alter ego narratives and enduring my weird fascination with collages, and stupid puns. And, of course, to my other team – João (my data science+ guide since Day 1) and Liliana (my *spare-time-PhD-and-delulu* buddy) – thank you for more than I can count thought-provoking conversations about data science, overthinking, the shape of two-dimensional sausage, and how the universe ended up like that!

And then, to those friendships that have been with me since the first days of my master's degree. Luuk, thank you for joining me on the ride of starting our own company. Above all, for being a great friend who understood and supported my

reasons to “jump” to yet another field and pursue a PhD in it. Rolien, the second half of our cat therapist duo, thank you for always being there for me and for your clear outlook that always gets me out of my overthinking patterns. I am so happy we got stuck in that miserable project group together all those years ago!

I began this Acknowledgement section by mentioning my high school philosophy teacher. Vassia, thank you for trying to teach the intricacies of Philosophy to a dumb 15-year-old kid. Thank you for making me think logically and critically. And sorry I made you read so many essays about the truth. I will also be remiss not to mention my mathematics teacher, Daniela Zhekova. Thank you for teaching me what elegance means! The way I perceive a problem, approach it, and reason about it – I owe it all to both of you.

Sartre claimed that a man – in this case, a woman – is condemned to be free. And I abide by this maxim more than I would like to admit. Yet, all the condemnation that stems from it, and the one that I occasionally create for myself, is made easier because it is shared. Thank you all for being there to help me shoulder the full burden of my freedom!

I am lucky to have you in my life!

September 3, 2023
Delft, the Netherlands

БЛАГОДАРНОСТИ

Едно от първите понятия, с които учителят ми по философия, Вася Велкова, ме „запозна“ беше сартеровата идея, че “*човекът е осъден да бъде свободен*”. И сякаш изведнъж, този парадокс облада въображението ми и се закотви трайно в моята тийнейджърска тревожност. Защото за пръв път през живота си получих уверение, че единственият избор, който имам е да бъда свободна. Уверение, което беше едновременно омайващо и смущаващо.

За мое най-голямо съжаление, обаче, се оказа, че не всичко е пряка последица от изборите ми. Човек не избира семейството си и има сравнително малко влияние върху нещата, с които се сблъсква. Освен това, научният прогрес в областта на неврологията и поведенческата психология убедително потвърждава хипотезата за химерността на абсолютната свобода. А свят, определян от липсата на пълна свобода, е неминуемо и свят, определян от растящата роля на късмета. За мое най-голямо щастие, късметът винаги е бил на моя страна. Особено по отношение на хората в живота ми.

Преди всичко, най-големият извор на късмет в живота ми е моето семейство. Тази дисертация е за тях. За моята чудесна, творческа,

забавна и креативна майка. Благодаря ти, че ме научи как да мечтая и ме накара да вярвам, че мога да се боря с мишките, дори когато всички обстоятелства са срещу мен. На баща ми, моят морален компас и постоянен трезво-мислещ събеседник. Благодаря ти, че ми показа важността на това никога да не отстъпвам от своите ценности и достойнство; дори когато машината се опитва да те погълне цял, точно както героя на Чарли Чаплин в "Модерни времена". На невероятната ми сестра и най-добър приятел. Благодаря ти, че винаги ме изправяш, особено когато не ми стигат силите да го направя сама. Обичам те от момента, в който те видях. И на баба ми. Често се затруднявам да подбера думи, за да те опиша. Ти си толкова много неща! Въпреки това, мисля, че *непоколебима* е добро описание. Благодаря ти, че винаги вярваш в мен, особено в моментите, в които куражът ми се „колебае".

Също така, имам късмет и с "осиновеното" си семейство. Майке, думите не стигат! Ти си сторила толкова много за мен. Благодаря ти, че ми предостави възможността да мисля върху дълго-забравени теми и след това стана мой научен ръководител. Благодаря ти за свободата, която ми предостави и запази за мен. Милен, не си спомням дали някога съм ти казвала, че първата защита на докторска дисертация, която видях, беше твоята. Така че беше естествено и до някаква степен съдбовно, когато се съгласи да се присъединиш към екипа. Благодаря ти, че ми помогна да отсея важните неща. Благодаря и на двете ви за непоколебимата подкрепа през последните четири години. За това, че балансирахте патологичните ми наклонености към прекомерни размишления и че често играхте ролята на така-необходимото ми "лепило", което ми позволи да формулирам мислите си ясно.

Отправям огромни благодарности и към Дирк и Кристине, които бяха част от моя научен екип в началото. Без тях дисертацията нямаше да съществува. Кристине, моята фея орисница, благодаря ти, че беше до мен и култивира моята любов към науката. Дирк, благодаря ти, че търпя

моите често нелогични въпроси и размисли. Преди всичко, благодаря и на двама ви за нестихващата ви подкрепа.

Дължа много и на хората, които подкрепиха емпиричната част от тази дисертация. Никол и Александра – благодаря ви за предоставената възможност да работим заедно! Томас и Мике – за това, че ми позволихте да се присъединя към вашите първи опити с електрически автомобили. На моите студенти, които винаги задаваха провокиращи въпроси. Специални благодарности отправям към Приянка, Фин, Елена, Катя, Роберто, Дерек, Марайн, Лий, Тобиас, Хусше, Шантал и Бибиан. Надявам се, че сте научили от мен толкова, колкото и аз съм научила от вас. Благодаря ви!

Специални благодарности отправям и към членовете на докторската комисия, които любезно се съгласиха да отделят време за четене и критично оценяване на работата ми. Професори Бозон, Кеш, Донг, Лойд и Розер – чувствам се поласкана, че имах привилегията да споделя работата си с Вас! Благодаря Ви за внимателния Ви отзив и готовността да предоставите своите знания. Вашите усилия и насоки значително подобриха качеството на дисертацията.

Също така, дължа благодарности и на хората, които непряко сформираха работата ми. Валерия и Есан - благодаря ви, че винаги ме карате да се замислям какво е Дизайнът. На моите съучастници дизайнтолози – Ерик Ян и Йерун (моят паранимф). Благодаря ви за безчетните дискусии. И разбира се, за всички абдуктивни ванове, пистолети с лепящи листчета и дилеми произхождащи от Наполеоновия синдром. На Ян-Мартен (моят друг паранимф), който винаги има критичен въпрос „в джоба си“. Йерун и Ян-Мартен, благодаря ви, че се съгласихте да бъдете моите паранимфи и да стоите до мен, докато завършвам и защитавам тази работа! Благодаря и на Киел, Ребека, Сико, Мике, Сандър, Вики, Ханеке и Леандра, че направиха „пребиваването“ ми в университет по-приятно.

За последните две години, специални благодарности също така отиват и към моите колеги от Деус за тяхната подкрепа. Преди всичко, на

Натали, Рон, Ник, Ричард и Дани за доверието, което ми оказаха и възможностите, които не спират да ми предоставят. На Натали, която ме въведе в компанията и ми предостави пространство да развивам идеите и вижданията си. Натали, благодаря ти за безчетните разговори, подкрепа и доверие! Винаги ще бъда благодарна на Кари за това, че ни запозна! Специални благодарности също отиват и към Рон за безчетните му съвети.

Разбира се, екипът ми също заслужава специални благодарности: Алекс, Анибал, Диего, Инеш, Хосе, Хуан, Марта, Мигел, Шес, Карлос и Родриго. И по-специално, на Хуан Карлос (заради неспирната му подкрепа, готовността му да служи като терапевт и фактът, че знае къде му е кърпата във всеки един момент), на Крис (за вечната ѝ готовност да „скочи“ в ново приключение), на Ездрас (за всичките му саркастични и проницателни наблюдения върху човешкото състояние), на Джуниър (за това, че винаги успява да ме развесели), и на Гус (за готовността му да помогне и отговори на всичките ми глупави въпроси). Благодаря на всички ви за безбройните дискусии свързани с банани, метафизични размисли на тема „чипсове“, песни за патици, и за неизчерпаемото търпение към моето странно влечение към колажи и глупави думички. И разбира се, на моя друг екип - Жоао и Лилиана - благодаря ви за многобройните разговори за изкуствен интелект, формата на двуизмерната наденица и Вселената!

И разбира се, на тези приятели, които продължават да бъдат неизменно с мен от първите дни на магистратурата ми. Луук, благодаря ти, че беше до мен докато стартирахме собствената ни компания. Преди всичко, за това, че си чудесен приятел, който разбра и подкрепи причините ми да се захвана с ново предизвикателство. Ролин, втората половина от нашата двойка терапевти на котки, благодаря ти за твоята ясна перспектива, която винаги ме извежда от моите патологични разсъждения. Толкова

съм щастлива, че ни накараха да работим заедно в онази ужасна група преди толкова години!

Започнах тази последна част на дисертацията споменавайки учителя си по философия. Вася, благодаря ти, че ме запозна с тънкостите на философията. Благодаря, че ме научи да мисля логично и критично. И съжалявам, че те накарах да прочетеш толкова много есета за *истината*. Не искам да пропусна да спомена и учителя ми по математика, Даниела Жекова. Госпожо, благодаря Ви, че ме научихте какво означава елегантност. Начинът, по който възприемам трудните задачи и се изправям пред тях дължа на Вас двете!

Сартр твърди, че човекът е осъден да бъде свободен. И въпреки всичко, аз продължавам да се придържам към тази максима по-силно, отколкото фактите ми позволяват. Но, цялата тази обреченост, която често сама си създавам, става по-лека, защото е споделена. Благодаря на всички ви, че сте до мен и неуморно ми помагате да понеса пълния товар на свободата си!

Имам късмет, че сте в живота ми!

3 септември, 2023
Делфт, Нидерландия

SUMMARY

SAMENVATTING

Het vakgebied van kunstmatige intelligentie (KI) heeft zich sinds de oprichting gericht op het ontwikkelen van systemen die duidelijke antwoorden bieden. KI-systemen vertellen ons de beste zet in een schaakspel, de juiste vertaling voor een woord, hoe we een eiwit moeten vouwen, en voorspellen of iemand een boek zal kopen. Deze systemen hebben een duidelijk doel, een duidelijke uitkomst en in sommige gevallen ook duidelijke belonings- en straffuncties. De antwoorden die ze bieden, kunnen gemakkelijk worden geclassificeerd als juist of onjuist. Nadat een schaakstuk is verplaatst, kan men eenvoudig controleren of de zet optimaal was en de kans om te winnen heeft vergroot. Deze kenmerken hebben geleid tot de adoptie van KI-systemen in gebieden zoals productie- en transportmethoden, de manier waarop mensen informatie ontvangen, films en liedjes selecteren, daten, handelen op de aandelenbeurs en de manier waarop sociale instellingen zoals ziekenhuizen, banken, politieafdelingen en rechtbanken beslissingen nemen (bijvoorbeeld Rahwan et al., 2019).

Echter, zodra KI-systemen — die duidelijke regels en doelstellingen nodig hebben — te maken krijgen met de complexiteit van sociale contexten, veroorzaken ze soms onbedoelde en soms schadelijke resultaten (Rudin,

2019). Zo leidde een KI-systeem dat door de Nederlandse Belastingdienst werd gebruikt tot de onterechte vervolging van duizenden families en het aftreden van de hele regering in 2021 (Hanley, 2021). KI-systemen in gebruik van techgiganten als Google, Meta en Twitter (nu X) labelen vaak zwarte mensen als primaten (BBC, 2021b). Toen de Covid-19-pandemie uitbrak, resulteerde het KI-systeem dat Zillow gebruikte om huizenprijzen te schatten in meer dan 300 miljoen dollar verlies voor het bedrijf en het ontslag van 2000 mensen (CNN, 2021). In dergelijke situaties kunnen noch de problemen noch de doelstellingen duidelijk worden gedefinieerd. Verder zijn er in plaats van ware of onware antwoorden talloze mogelijke oplossingen die alleen als goed of slecht kunnen worden geclassificeerd. Daarom kunnen we, als we KI-systemen willen ontwerpen die geen verstrekkende schadelijke gevolgen hebben, ze niet loskoppelen van de complexe contexten waarin we ze borgen.

Historisch gezien is het vakgebied van ontwerpen ontwikkeld als reactie op grote veranderingen in de samenleving die mogelijk werden gemaakt door snelle technologische ontwikkelingen (Calabretta & Kleinsmann, 2017). In de loop van een eeuw evolueerde het Ontwerp geleidelijk van het ontwerpen van producten naar het ontwerpen van mens-computer interacties, product-dienst-systemen en recentelijk complexe sociotechnische systemen (Norman & Stappers, 2015). Deze uitbreiding van het vakgebied heeft een voortdurend debat aangewakkerd over of otwerptheorieën moeten worden aangepast om de steeds complexere contexten aan te pakken waarin ontwerpers werken (bijvoorbeeld Voûte et al., 2020). Dit debat moeten worden uitgebreid naar het ontwerpen van KI-systemen die in grotere complexe contexten worden geïmplementeerd. Vandaar de initiële onderzoeksvraag van dit proefschrift:

"Hoe kunnen Ontwerptheorieën het ontwerp en de implementatie van KI-systemen in complexe contexten ondersteunen?"

Het is tegen deze achtergrond dat we de theoretische en empirische verkenning starten die in het proefschrift wordt gepresenteerd. We volgen de door Cash (2018) voorgestelde cyclus van theorieopbouw/testen. We behandelen met name de eerste drie fasen: Ontdekking en beschrijving (dat wil zeggen, het gedetailleerd beschrijven van de belangrijke kwesties waarop de theorie is

gebaseerd), Definities van variabelen en beperking van het domein (dat wil zeggen, het presenteren van variabelen en hun definities, evenals waar en wanneer een theorie moet worden toegepast), en Relatieopbouw (dat wil zeggen, het beschrijven van de conceptuele relaties tussen de geïdentificeerde variabelen) (Cash, 2018). Deze fasen komen overeen met de drie delen van het proefschrift.

Deel I: De situatie schetsen

De wijdverbreide adoptie van KI-systemen is het resultaat van de significante vooruitgang op het gebied in het afgelopen decennium (Rahwan et al., 2019). Zo kunnen diepgaande leermethoden longkanker voorspellen met een nauwkeurigheid van 94,4% (Ardila et al., 2019), beter presteren dan radiologen bij het detecteren van longontsteking (Rajpurkar et al., 2017), hypertrofische cardiomyopathie detecteren bij asymptomatische patiënten (Green et al., 2019) en de locaties van naschokken van aardbevingen voorspellen (DeVries et al., 2018). Machine learning, dat modellen traint door voorbeelden van gewenst input-outputgedrag, drijft het grootste deel van deze prestaties aan.

Hoewel er verschillende soorten KI-systemen zijn, hebben ze gemeenschappelijke kenmerken: (1) hun besluitvormingsprocessen zijn complex en niet gemakkelijk te begrijpen, (2) ze zijn sterk afhankelijk van door mensen gegenereerde gegevens, en (3) ze leren voortdurend van menselijk gedrag en passen zich aan menselijk gedrag aan. Deze kenmerken maken het moeilijk om ervoor te zorgen dat KI-systemen transparant, eerlijk, verantwoordelijk en niet-schadelijk zijn (Jobin et al., 2019). Als gevolg daarvan zijn er vier belangrijke uitdagingen bij de implementatie van KI-systemen in bestaande contexten: (1) zorgen voor transparantie van de innerlijke werking van KI-systemen, (2) het verminderen van impliciete vooroordelen in de gegevens die worden gebruikt om deze systemen te trainen en opnieuw te trainen, (3) de moeilijkheid om het gedrag van KI-systemen af te stemmen op

menselijke waarden, en (4) het aanpakken van de onderlinge afhankelijkheid van gedrag en gebruik van deze systemen.

Terwijl er onderzoeksbenaderingen bestaan voor de eerste drie uitdagingen, is de uitdaging van het aanpakken van de behaviour-use interdependence van KI-systemen nog niet volledig behandeld in bestaande literatuur. Ontwerptheorieën hebben veel raakvlakken hiermee vanwege de centrale rol die zowel gedrag als gebruik spelen in de kern van Ontwerpregeneratie - innovatieve abductie (Roozenburg, 1993). Deze inzichten bieden een manier om de hoofdonderzoeksvraag te herformuleren tot:

"Hoe kan een theoretisch model worden ontworpen dat de vroege simulatie van de AI system's behaviour-use interdependence ondersteunt door gebruik te maken van Ontwerptheorieën?"

Deel II: Initieel theoretisch model

We beginnen onze verkenning met een theoretisch onderzoek naar Ontwerptheorieën die kunnen helpen bij de vroege simulatie van de onderlinge afhankelijkheid van gedrag en gebruik van KI-systemen. Allereerst onderzoeken we de paradigma's die grotendeels de meeste Ontwerptheorieën hebben gedefinieerd – Rational Problem Solving (Simon, 1996) en Reflective Practice (Schön, 1983). Ondanks hun verschillen beweren beide paradigma's dat (1) het opzetten van een feedbackmechanisme instrumenteel is om een beter begrip van het probleem te krijgen; (2) de ontwerper invloed heeft op het ontwerpproces; en (3) onbedoelde resultaten een natuurlijk en noodzakelijk gevolg zijn van het ontwerpproces omdat ze het vooruit stuwen.

Ten tweede onderzoeken we de ontwerptheorie van innovatieve abductie, die formeel gedefinieerde relaties tussen de variabelen gedrag en gebruik biedt. Het introduceert ook de manieren waarop deze twee kunnen worden gebruikt om nieuwe oplossingen te ontwerpen. Er bestaan twee algemeen aanvaarde modellen die de relatie definiëren - één geïntroduceerd door Roozenburg (1993) en een door Dorst (2011). Ondanks hun verschillen beginnen ze beide met (1) een aanvankelijk overeengekomen startpunt (doel en waarde), (2) dat ze

gebruiken om het gedrag en gebruik van de mogelijke oplossing te definiëren (wijze van handelen en activering, en hoe), en (3) de combinatie hiervan leidt tot een tastbare oplossing - een vorm of een object, dienst of een systeem.

Ten derde kunnen prototypes de continue simulatie van het gedrag en gebruik van een oplossing die in complexe contexten moet worden geïmplementeerd, ondersteunen. Sterker nog, ze (1) ondersteunen ons bij het observeren van de verschillende soorten resultaten en toepassingen die het gedrag kunnen opleveren; (2) dienen als brug tussen gedrag en gebruik; en (3) stellen wat Magnani (2007) manipulatieve abductie noemt, in staat. We illustreren deze inzichten met twee voorbeelden uit een project dat is uitgevoerd tussen november 2015 en september 2016 voor een grote Europese luchtvaartmaatschappij.

Met deze basis onderzoeken we theoretisch de geschiktheid van ontwerptheorieën voor de vroege simulatie van de onderlinge afhankelijkheid van gedrag en gebruik van KI-systemen. We introduceren vervolgens een eerste versie van een theoretisch model, dat een reeks relaties voorstelt tussen de door ons geïdentificeerde variabelen.

purpose + data → frame

frame + mode of action → prototype

prototype + actuation → outcomes

Het model helpt ons om het gedrag van het systeem aan te passen om het gewenste gebruik en de gewenste resultaten te activeren. Als gevolg is het model het meest geschikt voor de vroege stadia van conceptueel ontwerp wanneer noch het gedrag noch het gewenste gebruik (of resultaten) duidelijk zijn. Hierdoor wordt een voorlopig antwoord gegeven op de hoofdonderzoeksvraag van het proefschrift.

Deel III: Uitgebreid theoretisch model

De ontwerptheorieën die we hebben gebruikt, zijn echter ontwikkeld voor het ontwerpen van producten en diensten. In tegenstelling tot deze, wordt

het gedrag van een KI-systeem continu beïnvloed door en leert het van door gebruikers gegenereerde gegevens. Daarom onderzoeken we in Deel III verder hoe de onderlinge afhankelijkheid van gedrag en gebruik kan worden gesimuleerd in de context van het ontwerpen van KI-systemen.

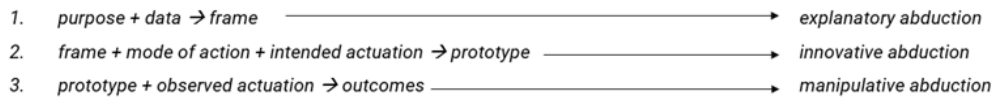


Figure 0 Formulaic representation of the theoretical model

We beginnen onze verkenning door een voorbeeld te presenteren van een studententeam van drie personen dat met succes de onderlinge afhankelijkheid van gedrag en gebruik van een KI-systeem vier keer heeft gesimuleerd door eenvoudige prototypes te gebruiken. Het team haalde meerdere (bedoelde en onbedoelde) resultaten naar boven, die dienden als een robuust feedbackmechanisme. Drie dingen hielpen het team. Ten eerste identificeerden ze expliciet bedoelde activeringen nadat ze hun wijze van handelen hadden besloten en voordat ze een prototype bouwden. De toevoeging van deze variabele stelt ons in staat om elke regel van het theoretische model te verbinden met een ander type abductieve redenering: verklarende, innovatieve en manipulatieve (Figuur 0).

Ten tweede suggereren huidige ontwerptheorieën dat om een nieuwe oplossing te ontwerpen, innovatieve (en verklarende) abductie moet worden toegepast. Toch maakte het team gebruik van alle drie de soorten abductie om de onderlinge afhankelijkheid van gedrag en gebruik van hun concept te simuleren. Ten derde gebruikten ze ook niet-abductieve activiteiten zoals expliciete definitie van vereisten en waarden. Deze speelden een cruciale ondersteunende rol bij de ontwikkeling van het concept voor een KI-systeem.

We gaan in op de ontdekte inzichten met een vijf maanden durende casestudy. Hierin wordt een ontwerpproject besproken voor het ontwikkelen van een KI-systeem voor in de auto, uitgevoerd in samenwerking met een grote autofabrikant. De casus levert twee belangrijke conclusies op. Ten eerste

moeten om de vroege simulatie van de onderlinge afhankelijkheid van gedrag en gebruik van KI-systemen te ondersteunen, de drie abductietypen - verklarend, innovatief en manipulatief - worden toegepast. Ten tweede moeten bestaande Ontwerptheorieën worden uitgebreid. Vijf inzichten kunnen dergelijke uitbreiding leiden: (1) verklarende abductie wordt meestal gevolgd door innovatieve abductie; (2) de inductieve generatie van nieuwe waarden en vereisten informeert de formulering van elke variabele van het model; (3) visuele elementen die voortkomen uit inductieve redenering (bijvoorbeeld gegevensvisualisaties) vergemakkelijken verklarende abductie; (4) de deductieve evaluatie van het resultaat van elke rij tegen vereisten en waarden ondersteunt de overgang van het ene abductietype naar het andere; en (5) manipulatieve abductie speelt een faciliterende rol bij het uitvoeren van innovatieve abductie.

Deze inzichten vormen de basis van het uiteindelijke theoretische model dat we het Theoretisch Model voor Prototyping KI of het PAI-model noemen. Het PAI-model wordt gedefinieerd door relaties tussen abductie (verklarend, innovatief en manipulatief), inductie en deductie. Een model dat ons voorziet van een manier om de vroege simulatie van de onderlinge afhankelijkheid van gedrag en gebruik van KI-systemen te ondersteunen. Bovendien zijn we van mening dat het PAI-model op dezelfde manier zal worden toegepast door verschillende KI-ontwikkelaars, ongeacht hun achtergrond of vaardigheidsniveau. Ten slotte geeft het model ook een indicatie van hoe drie verschillende soorten gegevens kunnen worden gebruikt om het gedrag van KI-systemen bij te werken tijdens de ontwikkeling en implementatie van het model. Deze kunnen dienen als het startpunt voor het deel van de cyclus van Cash (2018) dat betrekking heeft op het testen van de theorie.

Door de drie stappen van theorieopbouw te volgen die zijn voorgeschreven door Cash (2018), stelt het opstellen van het PAI-model ons in staat om inzicht te krijgen in hoe Ontwerptheorieën kunnen bijdragen aan het ontwerp van betere KI-systemen. Het stelt ons ook in staat om deze theorieën uit te breiden en richtingen te identificeren waarin het vakgebied zich in een toekomst gedefinieerd door intelligente agenten zou kunnen (of moeten)

ontwikkelen. Hierdoor biedt het model ons een manier om het conceptuele ontwerp van steeds evoluerende KI-systemen te benaderen door de vroege simulatie van hun onderlinge afhankelijkheid van gedrag en gebruik. Tot slot kunnen deze geformaliseerde relaties ons ook aanwijzingen geven over hoe nieuwe KI-modellen kunnen worden ontwikkeld. KI-modellen die niet uitsluitend afhankelijk zijn van grote aantallen datapunten, maar in plaats daarvan de creatie van sterk configureerbare wereldmodellen voor AI-agenten mogelijk maken.

CURRICULUM VITAE

Niya Evtimova Stoimenova
born in Petrich, Bulgaria

EDUCATION

Delft University of Technology

2019 – 2023

PhD at the Faculty of Industrial Design Engineering, focused on applying Design theories to address the behaviour-use interdependence of AI systems

Delft University of Technology

2015 – 2017

MSc (cum laude) in Strategic Product Design
Honours programme graduate (focus on implementing innovation strategies in large international companies)

The Hague University of Applied Sciences

2011 – 2014

BEng in Industrial Design Engineering

PROFESSIONAL EXPERIENCE

DEUS

2022 - Present

Reliable AI Lead (leading the development of reliable AI solutions for DEUS)

Onami

2017 - 2019

Co-founder (supporting organisations in developing and establishing new innovation strategies)

Delft University of Technology

2017 - 2019

Researching the opportunities to apply principles of design cognition to AI algorithms

TEACHING EXPERIENCE

Delft University of Technology

2018 – 2021

Coordinated a university-wide elective on Entrepreneurial Learning
Coached in the following master's courses: Design Strategy Project, Design Theory and Methodology, SPD Research
Coached the following bachelor's course: Business, Culture and Technology

MENTORING EXPERIENCE

DEUS

2022 – now

Serve as a company mentor during the graduation and research projects of MSc students from Eindhoven University of Technology, Utrecht University, Radboud University, VU Amsterdam. One of the graduations resulted in a conference paper.

Delft University of Technology

2018 – 2022

Mentored six master's students during their graduation projects
Mentored two research elective courses, which resulted in two conference papers

AWARDS

2023

Nominated for the award in the category AI Researcher given by Women in AI (Netherlands)

2017

Awarded the TU Delft's Honours Programme Best Paper award

2015

Finalist of the essay competition of the Third Hague Peace Conference

2011

Laureate of the National Olympiad in Philosophy (Bulgaria)

INVITED TALKS AND INTERVIEWS

2023

BioFutures Symposium (New Castle), *The Alignment Problem*, organised by Northumbria University

TechDays Event (Amsterdam), *Beyond the hype: a realistic perspective on AI and its limitations*, organised by Kickstarter AI

Business Insider Nederland (2023) *An interview on the importance of approaching AI systems from a systemic point of view* (in Dutch) https://shorturl.at/rwKL0

Data makers Fest (Porto) *Beyond the hype of AI systems*, Portugal

2021

Featured project during the Dutch Design Week's **Up Close & Personal** talk show

AUXILIARY ACTIVITIES

2021

Member of the Ethics Working Group in the International Atomic Energy Agency on *AI for Nuclear Technologies and Applications*

PAPERS

during the PhD

Angelucci, A., Li, Z., **Stoimenova**, N., & Canali, S. (2022). The paradox of the artificial intelligence system development process: the use case of corporate wellness programs using smart wearables. *Ai & Society*, 1 - 11.

Stoimenova, N., & Price, R. (2020). Exploring the Nuances of Designing (with/for) Artificial Intelligence. *Design Issues, 36*, 45-55.

Stoimenova, N., & Kleinsmann, M. (2020). Identifying and addressing unintended values when designing (with) Artificial Intelligence. *Design Research Society Conference.* Brisbane.

Stoimenova, N., & de Lille, C. (2020). The adaptive organization: using design's prototyping practices to innovate in complex contexts. *Academic Design Management Conference.* Toronto, Canada.

Stoimenova, N., Lille, C.D., & Stomph, S. (2019). The Organization as a Prototype. *Conference Proceedings of the Academy for Design Innovation Management*.

before the PhD

Stoimenova, N., Lille, C.D., (2019). Building Design-led Ambidexterity in Big Companies. *Conference Proceedings of the Academy for Design Innovation Management*.

Stoimenova, N., & Lille, C.D. (2018). Building the Foundation for a Design-Led Ambidexterity in a Medium-SizedTech Company. *DRS2018: Catalyst*.

Stoimenova, N., Lille, C.D., Stoimenova, N., Ferreira, C., & Ferreira, C. (2017). Co-Designing Innovation in Fast-Paced Environments: Organizational Challenges and Implications.

Stoimenova, N. (2015). FOUR GUIDING FACTORS FOR FACILITATORS OF MULTIDISCIPLINARY.

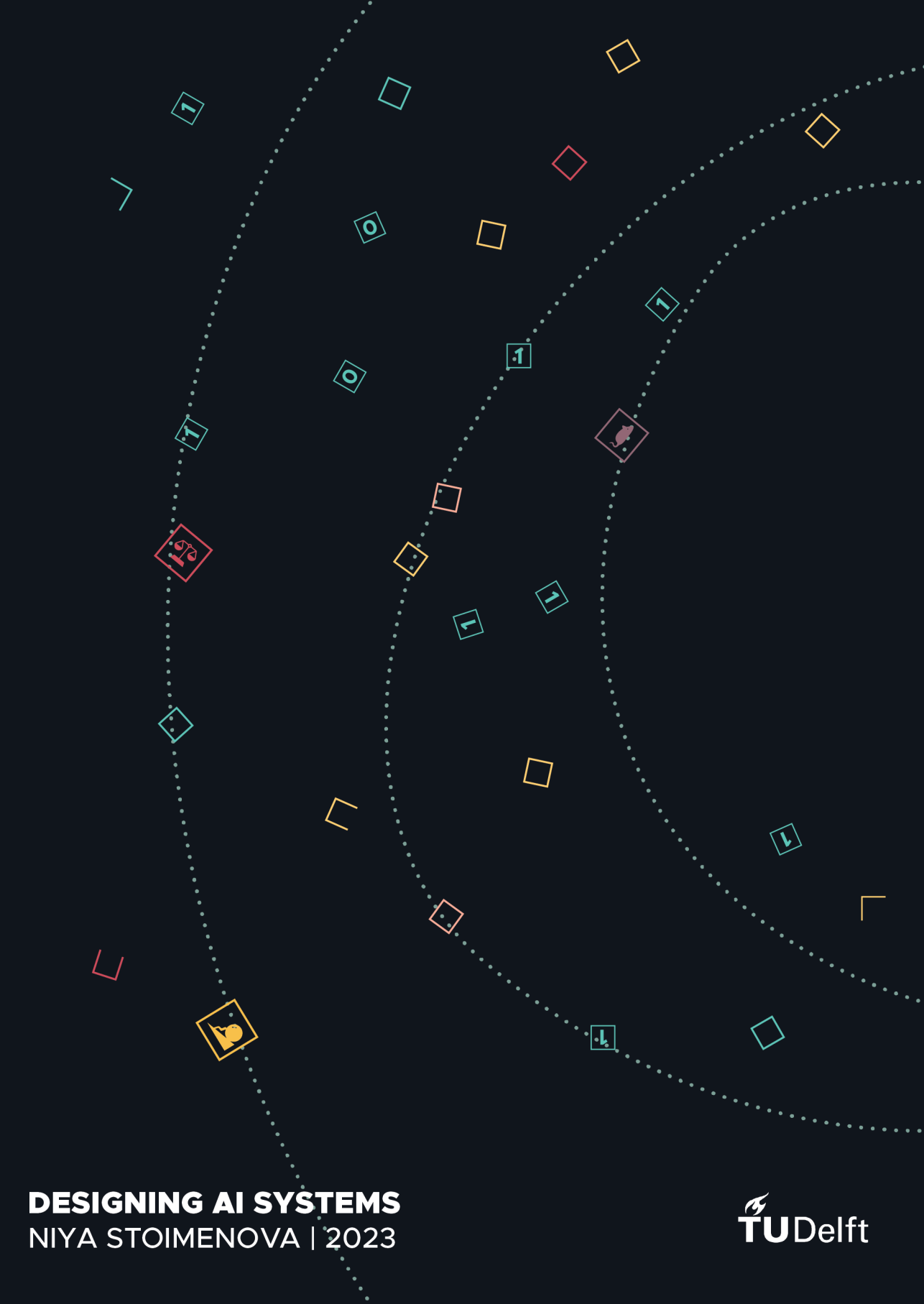
Stoimenova, N. (2015). Optimisation of Multidisciplinary Collaboration in Fast-Paced Innovation Projects. *Student Undergraduate Research E-journal, 1*.

SUPERVISED PAPERS

Juijn, G., **Stoimenova**, N., Reis, J., & Nguyen, D. (2023). Perceived Algorithmic Fairness using Organizational Justice Theory: An Empirical Case Study on Algorithmic Hiring. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*.

Mariani, E., Kooijman, F.S., Shah, P., & **Stoimenova**, N. (2021). PROTOTYPING IN SOCIAL VR: ANTICIPATE THE UNANTICIPATED OUTCOMES OF INTERACTIONS BETWEEN AI-POWERED SOLUTIONS AND USERS. *Proceedings of the Design Society, 1*, 2491 - 2500.

in ‘t Veld, J., & **Stoimenova**, N. (2020). Dealing with changing environments: prototyping practices in organizations . *Academic Design Management Conference.* Toronto.



DESIGNING AI SYSTEMS

NIYA STOIMENOVA | 2023

TU Delft