



DESIGN FOR FAIRNESS IN AI

Cooking a fair AI dish

Design for fairness in AI

Cooking a fair AI dish

Author

Dasha Simons
dashasimons@gmail.com

Master thesis

MSc. Strategic Product Design
Faculty of Industrial Design Engineering
Delft University of Technology

Graduation committee

Chair | **Prof. Dr. E. Giaccardi**
Faculty of Industrial Design – Interactive Media Design

Mentor | **Dr. Ir. L.W. L. Simonse**
Faculty of Industrial Design – Product Innovation Management

Company mentor | **Dr. Zoltán Szlávik**
Lead - Center of Advanced Studies at IBM Benelux

April, 2019



Master thesis
By Dasha Simons

Acknowledgments

»

During the course of this thesis, I encountered a diversity of inspiring people. I would like to thank every single one of them in sparking my imagination, expanding my knowledge concerning AI, ethics, philosophy, design and the corporate world. I would not have been able to create this quality of work without you.

Specifically, I would like to thank my supervisory team.

Elisa, thank you for the striking enthusiasm and teaching me the fun and valuable corners of critical design, always inspiring coming with new perspectives and initiatives.

Lianne, thank you for your exceptionally perceptive strategic designer eye, keeping me on the ground when philosophy was taking me to the clouds and the story straight with energetic optimism.

Zoltán, thank you for all the insights in the AI field, the striking and endless support during the course of entire thesis and the trust for letting me go my own direction in this unexplored domain.

I enjoyed the inspirational collaboration with all three of you. The freedom and confidence you gave me excelled this project.

Furthermore, working at the CAS department of IBM Benelux was a great opportunity to interact with highly knowledgeable experts in the field and to get in touch with a variety leading clients to validate my designs. The ethics community was undoubtedly a great chance for me to interact with a diversity of people within the company and thank you for the opportunities to share my thoughts and critiques. In specific thank you to Sophie Kuijt, Dolf Noordman, Alessandro Giordani, Walter Moulart and Edurardo Wilde Barbaro (in no particular order) for the support in client contact. Special thank you to people that participated in the interviews, the ideation workshops and validation in the course of this thesis.

I wish to acknowledge all my friends who made path of this thesis motivating and amusing along the way. In specific I want to thank Cees den Bakker for his razor sharp feedback and astonishing patience, Pervin Çelik for the positive conversations and support, Anna Filippi for the encouraging presence, Cyril Schouten for the down to earth attitude, Roel Tilbosch for his critical perspectives, Joris Hens for the right advice at the right moments.

Finally, I want to express my sincere gratitude to my family.

Executive summary

»

Artificial intelligence (AI) is an emerging field which unleashes massive new (business) opportunities. The potential growth and broad application of the AI technology has great economic benefits however also severe societal implications. Simultaneously, ethical challenges arise with its development. Questions of values and ethics are becoming urgent, as systems can be negatively biased and the decision processes are often not traceable, while impacting our lives. Abstract concepts such as fairness and values need to find their way into the fast and agile AI development processes. The contemporary (research and practice) fields tackle these challenges by technological feats, ethical AI principles and strategies. However, it are the decisions made by humans today and tomorrow that will shape our future. It is, therefore, alarming the translation of ethics to that day to day work of the AI development team is missing.

Hence, the central aim of this thesis is to explore and design support for AI teams with the creation of more ethical AI systems, bridging the gap between ethical AI principles and current practice. By that, design for organizational capacity for the development of fairer AI by using strategic design and critical design approaches. In this thesis, due to the diversity and magnitude of ethical challenges in AI, particular attention is paid to two challenges, fairness and value-alignment, to benefit from a design perspective. Three streams of expertise are brought together to tackle these challenges: AI, applied ethics and design.

Ethics bears critique, and this thesis argues that it can benefit from a design perspective, using imagination in the solution space and synthesized thinking for implementable ideas instead of solely discussion. The thesis focuses on ways how design approaches can supplement the ethical ones and thereby stimulate the ethical uptake in the AI field. Instead of defining what fairness is, this thesis takes a novel approach in unraveling ten unfairness sources in the AI development. It is aspired to reduce these sources of unfairness in AI, in project specific fashion. In AI practice, the ways ethics is incorporated and how value tensions are resolved is under-researched. In depth interviews, generative tools and provotypes are conducted and designed to research and critique the contemporary AI field in relation to ethics, both with IBM and their clients. Simultaneously to inquire novel value tensions in its development. Five main value tensions are unraveled in its relation to fairness.

All above is consolidated a framework to design for organizational capacity and team support leading to the creation of fairer and value-aligned AI systems.

With this framework an organizational role is designed, the ethical coach, to aid the AI team with co-creating fairer and value-aligned AI systems with an accompanying modular toolkit. The modular toolkit is iterated upon multiple times and uses the AI dish metaphor.

Finally, two evaluation sessions with IBM and their clients as well as the conversations concerning of the implementation of the toolkit led to recommendations for further development including education and implementation directions.

I sincerely hope you will enjoy reading this thesis, triggering your curiosity and sparking imaginations!

Reading Guide

The reading guide exhibits the overview of the report to assist the reader towards the desired text. At the right page is a short description per chapter which discloses its content.

Each chapter starts with a short introduction after which the topic and conducted research is elaborated upon. At the end of main section is an overview of the key insights on a dark blue background. These represent the rationale for the final design and in the recommendations for IBM.

If you are interested but have no time to read the most important insights are summarized or visualized on top of the dark blue background.

Client names and the names of studies' participants are anonymized to maintain privacy.

Text

» **Bold text**
Bold text indicates concluding insights

Italic text - Italic text
Italic text refers to quotations

Icons

Value tension is presented by |



Strategy to resolve value-tension is presented by |



Source is presented by |



Insights, Decisions & Conclusions

Text and visuals on the dark blue background contain the most crucial insights for the design and recommendations

Examples, Illustrative Quotes & Stories

Illustrative examples and quotes for a deeper understanding are written or visualized on the light pink background.

Define project	01 Project Context & Approach <i>Provides the projects context analyses, background information with the problem statements and the design objectives.</i>
Literature & expert review	02 Introduction to the AI Field <i>Provides the analyses of the AI technology necessary for the creation and understanding of the ethical support and introduces the AI Dish metaphor.</i>
	03 Defining the Flavor I Establishing the ethics foundation <i>Shares the analysis of the ethics field in relationship with the AI field and with the design field, thereby establishes fruitful complementary insights.</i>
	04 Seasoning Fairness I Disclosing sources of unfairness <i>Engages in an analysis of fairness in AI tackled with the new ethical design perspective.</i>
Empirical Research	05 Taste Differences I Demystifying value tension <i>Shows the analysis of value alignment and value tension literature in AI from the design for fairness perspective.</i>
Synthesis	06 A Peek in the Kitchen I Exploring the contemporary <i>Contributes with extensive design research with four AI teams and projects.</i>
Design	07 Preparing the Ethical Recipe <i>Provides a synthesis of the literature and study leading towards a framework to design for fairness in AI.</i>
	08 Designing for Fairness in AI <i>Explores a workshop to support AI teams to create fairer AI systems.</i>
	09 Ethical Coach Starters Pack <i>Proposes the ethical coach role with the initially designed tools and handles.</i>
	10 Recommendations & Discussion <i>Shares recommendations for the design and IBM as well as future research, concluding notes concerning the thesis and a personal reflection.</i>

Table of Contents

01 Project Context & Approach	10
1.1 Project context & IBM	12
1.2 Project objective & approach	20
02 Introduction to the AI field	26
2.1 AI foundation	28
2.2 AI Dish metaphor	32
03 Defining the Flavor I <i>Establishing the ethics foundation</i>	34
3.1 A taste of Ethics	36
3.2 AI & Ethics	43
3.3 Design perspective on ethics	45
04 Seasoning Fairness I <i>Disclosing sources of unfairness</i>	48
4.1 Fairness foundation	50
4.2 Sources of unfairness in AI	53
05 Taste Differences I <i>Demystifying value tension</i>	58
5.1 Value-alignment	60
5.2 Value-tension	68
06 A Peek in the Field I <i>Exploring the contemporary</i>	74
6.1 Design research set-up	76
6.2 Interviews and generative tool	79
6.3 Provotypes	84

07 Preparing the Ethical Recipe	90
7.1 From insights to design	92
7.2 Ethics in AI framework	96
7.3 Design guidelines	101
08 Designing for Fairness	104
8.1 Iterative Ideation	106
8.2 Ethical AI Coach	108
8.3 AI Dish	112
8.4 Shape Workshop	114
09 The Ethical Coach Starters Pack	122
9.1 Ethical coach with a modular workshop	124
9.2 Design validation	132
10 Recommendations & Discussion	140
10.1 Recommendations & Implication requirements	142
10.2 Discussion	146
10.3 Contributions to practice	152
10.4 Limitations & future research	154
10.5 Personal reflection	156
Bibliography	159
Appendices	

Chapter 01 |

Project context & approach

This chapter provides an overview of the project context. It sets the objective and the relevance of this thesis and introduces the intersection of the three domains addressed, namely design, artificial intelligence and applied ethics. Additionally, it shares the design and research approach tailored and used in this work.

In this chapter

- 1.1 Project context & IBM
- 1.2 Project aim & approach

Homo sapiens —Man the wise—

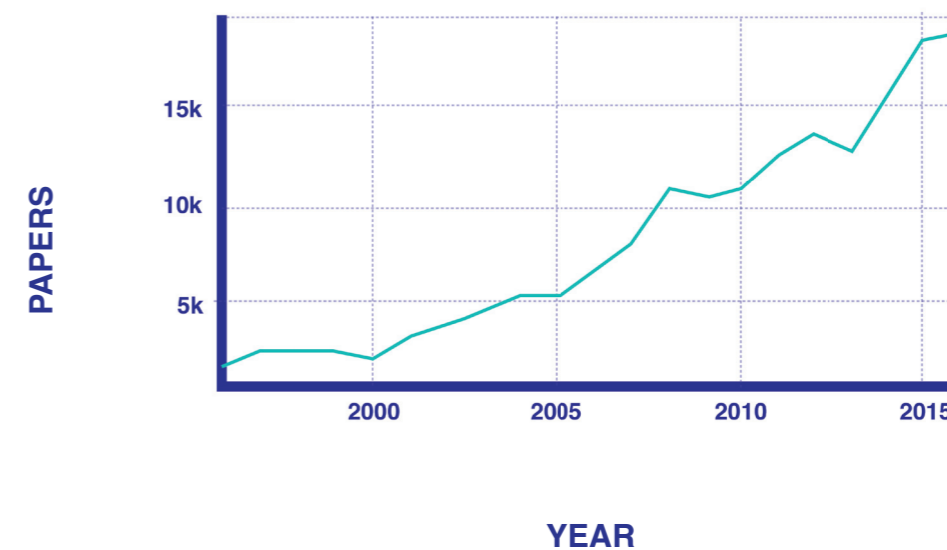
because our intelligence is so important to us, it is interwoven into the name of our species.

1.1

Project context & IBM

Since our existence, humans had the desire to understand how we think, predict, perceive, manipulate. The field of AI goes even further: it aspires to not only understand, but also build these intelligent entities. This chapter provides an overview of the project context and the chosen approach for the thesis founded in internal and external analyses.

Increase in research interest in machine learning



In figure 1.1. Growth of research in Machine learning over the years (Bughin et al. 2017)

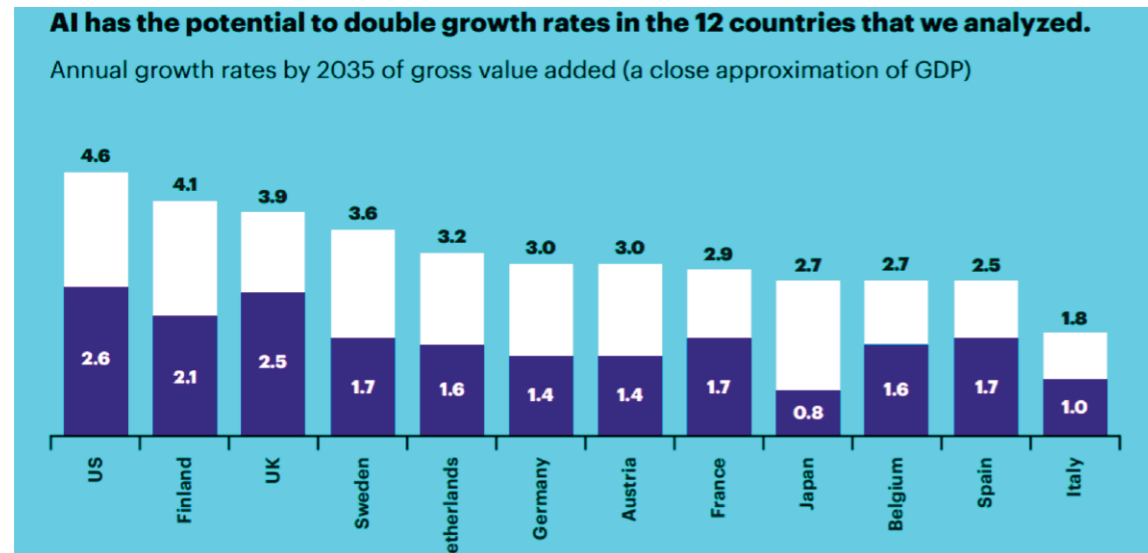
1.1.1 Growth of AI technology

Artificial intelligence (AI)—is an emerging field which unleashes massive new (business) opportunities (Fast & Horvitz, 2017; Raconteur, 2018; Davenport, 2018). AI consists of multiple technologies that can be combined in different ways to sense, comprehend, act and learn. By 2035, the economic profitability due to AI development is expected to increase by an average of 38 percent equivalent an economic boost of about US \$14 trillion across 16 industries in 12 economies by 2035

(Purdy, 2018) (figure 1.2). Simultaneously the research interest is rapidly rising (figure 1.1).

1.1.2 The need of ethics flavor in AI

The potential growth of broad application of the AI technology, has besides the economic benefits also severe societal implications, raising ethical questions concerning our future. Thus, ethics becomes increasingly relevant (Verbeek, 2014; Gonzalez, 2015; van den Hoven, 2015; Schatsky & Schwartz, 2015; Banavar, 2016; Boddington, 2017; Fast &



In figure 1.2 the economic impact of AI per country is visualized by Accenture (2017). It shows the high expectations of AI development in the future

Horviz, 2017; Erdelyi, 2018). Until lately, few attention has been paid to the ethical concerns of AI systems and the diverse ways it impacts people's lives, while ethically misaligned AI systems appeared in the market.

Unfair AI systems

The success and wellbeing of an individual are not fully in their own control. The decisions of others profoundly can influence our lives. For example, if a person is accepted to a particular school, job or someone's innocence in court is decided by others and increasingly by machines (Corbett-Davies et al., 2017).

» **Flawed, unreliable or arbitrary decision making is therefore extremely undesired, as it might lead to unfair access to opportunities (Barocas, Hardt, & Narayanan, 2018).**

For a long time, humans thought that math was objective and therefore fair (O'Neil, 2016). As algorithms are often based on math, the same was true for algorithms. Albeit, this is far from the truth. Companies and humans learned by mistakes in practice, with ethically misaligned AI systems (figure 1.4; 1.5). Examples include, Joy Buolamwini, MIT researcher who brought to the light that facial recognition systems are better in recognizing White-Caucasian

users due to a negatively biased data set (Lohr, 2018). Or firing teachers from high schools based on a "black box" algorithm making decisions, which aimed to calculate how much of the educational progress of the students could be attributed to the teachers, calculating this into a score (called IMPACT, by Princeton based Mathematica Policy Research). The system appeared to make unfair decisions based on too little data, leading to the firing of good teachers (O'Neil, 2016).

» **This is a problem of great concern in the present world. More decisions are left to machines and algorithms, which have consequential outcomes.** Due to these examples concerns about discrimination and fairness inevitably arise (Binns, 2017; Saxena, 2018). Currently models' outputs increasingly appear to be systematically biased towards people with certain attributes as race or gender. The consequences of this can be tremendous and therefore effort needs to be put together of both practice and research to try to prevent and solve these challenges with AI (Saxena, 2018). Thus, **the AI development process needs to be carefully assisted and guided towards a desired future, avoiding ethical pitfalls. Otherwise this can lead to undesired societal implications as illustrated before.**

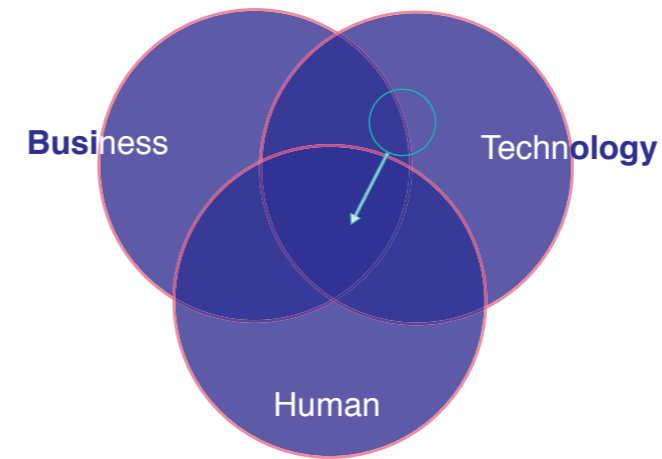


Figure 1.3 Representation of the technology push perspective in AI and the lack of the human one

1.1.3 The missing perspective

The external analyses performed in this thesis, indicated the gap in the contemporary AI field. See figure 1.6 on the next page.

Current technical endeavors

A creative trend research is conducted and presented in appendix D, the result in figure 1.8 to discover novel approaches to tackle the challenge of AI ethics. This creative trend research is based on AI events in the Netherlands visited during the course of the graduation and online trend research (Protein, Trendwatching, LSN Global, Deloitte, McKinsey trend reports on AI and AI ethics).

The use of the trend driven innovation framework (Mason, Mattin et al., 2015) provides a differentiating strategic direction, which not only helps distinguishing IBM from its competitors, but as well align with the human needs and the expectation for a fairer AI (see figure 1.6). The framework shows the sweet spot for the proposed strategic direction for IBM.

» Currently IBM, as well as their competitors release technical toolkits to identify and mediate bias in algorithms. Nearly all IBM's prominent competitors released this type of toolkit in 2018. This is a step towards fairer AI. However, these

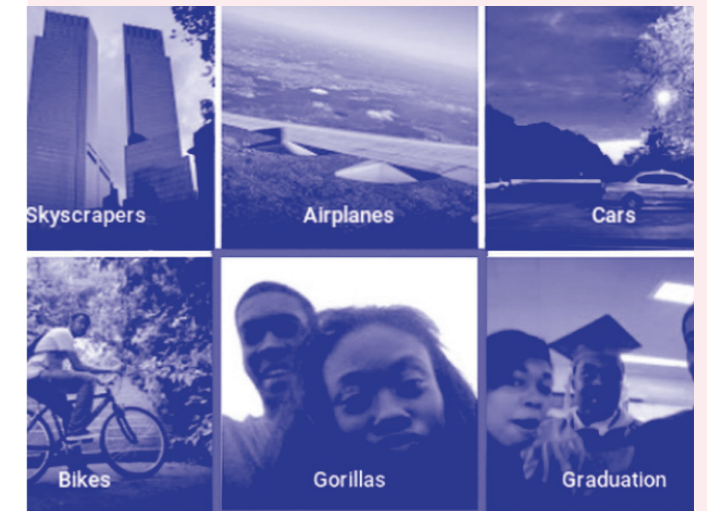


Figure 1.4 Biased Google image recognition recognizing darker skinned people as gorillas



Creative... Motivating and Fired

"It is a pleasure to visit a classroom in which the elements of sound teaching, motivated students and a positive learning environment are so effectively combined," Is written by the principles in the evaluation May 2011. He aspired Wysocki to share her methods with colleagues at the school. All observations of that year were positive and the ratings good. Two months later, she was fired.

Wysocki, was let go because the reading and math scores of her students didn't grow as predicted. Her undoing was "value-added," a complex statistical tool used to measure a teacher's direct contribution to test results which is used in around 25 states in the US to assess teachers. This tool was based on limited data, not representing reality. This led towards good teachers being fired.

Figure 1.5 Illustration of unethical AI

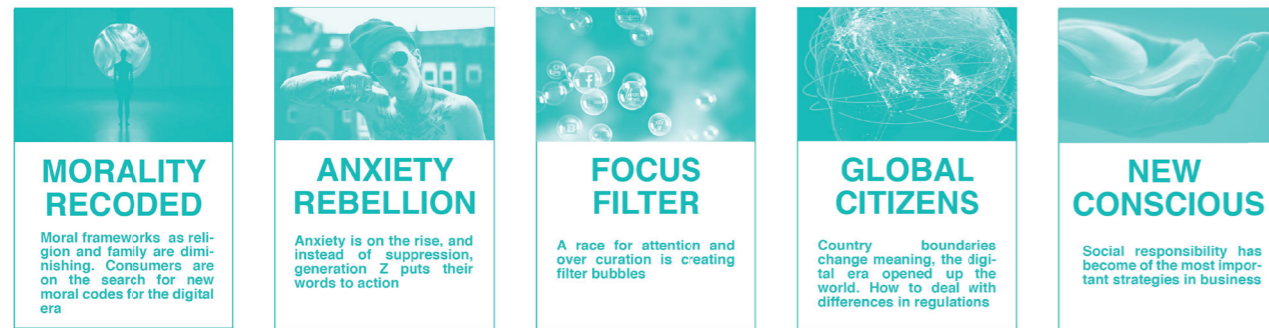


Figure 1.8 | Creative trend research performed in this thesis

toolkits exclude the human and societal facets of AI, their values. Therefore, many voices come to the same declaration: technology needs a human-heartbeat (Li, 2018; Avanaide, 2017; Massa, 2018). Instead of technological feat, it'll be human decisions that are made yesterday, today and tomorrow that will shape the future. It seems as if opportunities are led sideways by just sticking to the technology perspective (Fung, 2015). Therefore, we should take the (unaware) people making them in mind, the AI project development team.

At the same time at AI events (such as the world AI summit Amsterdam 2018) and in AI strategies of companies such as Google, IBM and Microsoft, share very inspiring principles toward more ethical AI development.

Strategic approach in this thesis

Hence, the translation of these principles

towards the day to day work of the AI team is lacking. Extracted from analysis of the trend driven innovation framework is a lack of a human centered approach in the more ethical AI development. At present the ethical challenges in AI are tackled by primarily technological endeavors.

» Supporting the people actually creating AI systems seems more suited to bridge the gap between (strategic) ethical principles and practice. As well, it is necessary to integrate societies perspective and humans context at center in the AI systems development to prevent ethically misaligned AI systems. Thus, this thesis takes these perspectives at center.

1.1.4 IBM

This thesis is written in collaboration with IBM Benelux and CAS, Center of Advanced Studies. However the design is tailored towards other

IBM departments. This section provides an comprehensive understanding of the context and the internal analysis of IBM. A general introduction of IBM is provided after which a light will be shed on their AI strategy. A more detailed analysis is shared in appendix B

The Big Blue

IBM, International Business Machine Corporation, also called Big Blue, is the largest information technology company in the world acting at 170 countries. Its foundation lies in the early days of the previous century, 1911. IBM has a strong research focus with the largest privately financed research labs and within its branch, owning the world record of patents (9,100, in January 2019 (Kirshna, 2019)). Over 350.000 people are working for this company with a diverse landscape of clients. Once it was famous for the first personal computer and the hardware they produced. In the changing scenery it transformed itself from a hardware company towards a software developing and service providing company. In figure 1.7 IBM's transformation is visualized over the last 50

years. This introduced great changes for them.

IBM values, purpose & ambition

The overall purpose of IBM is “**to be essential to our clients and the world**”. It is divided in three values, which are described in the flowing paragraph and nine different practices which are displayed in figure 1.8. IBM aims to be the coach during the client's digital change. This is a continuous change, which does not stop at 2020 – therefore IBM strongly invests in research and has a strong sense of where technology is heading.

» **IBM strategy is about trust – not because it is fashionable, but they have always been.**

IBM is the first who said, your data is your data, it is in our contract (Gerard Smit, IBM Benelux CTO, 5 in 5 technologies event 1/11/18)

» **IBM is a b2b company, therefore the services they provide are to business clients.**

Artificial intelligence is called by IBM also augmented intelligence or cognitive solutions. IBM's ambition is to be The AI company for large enterprises.

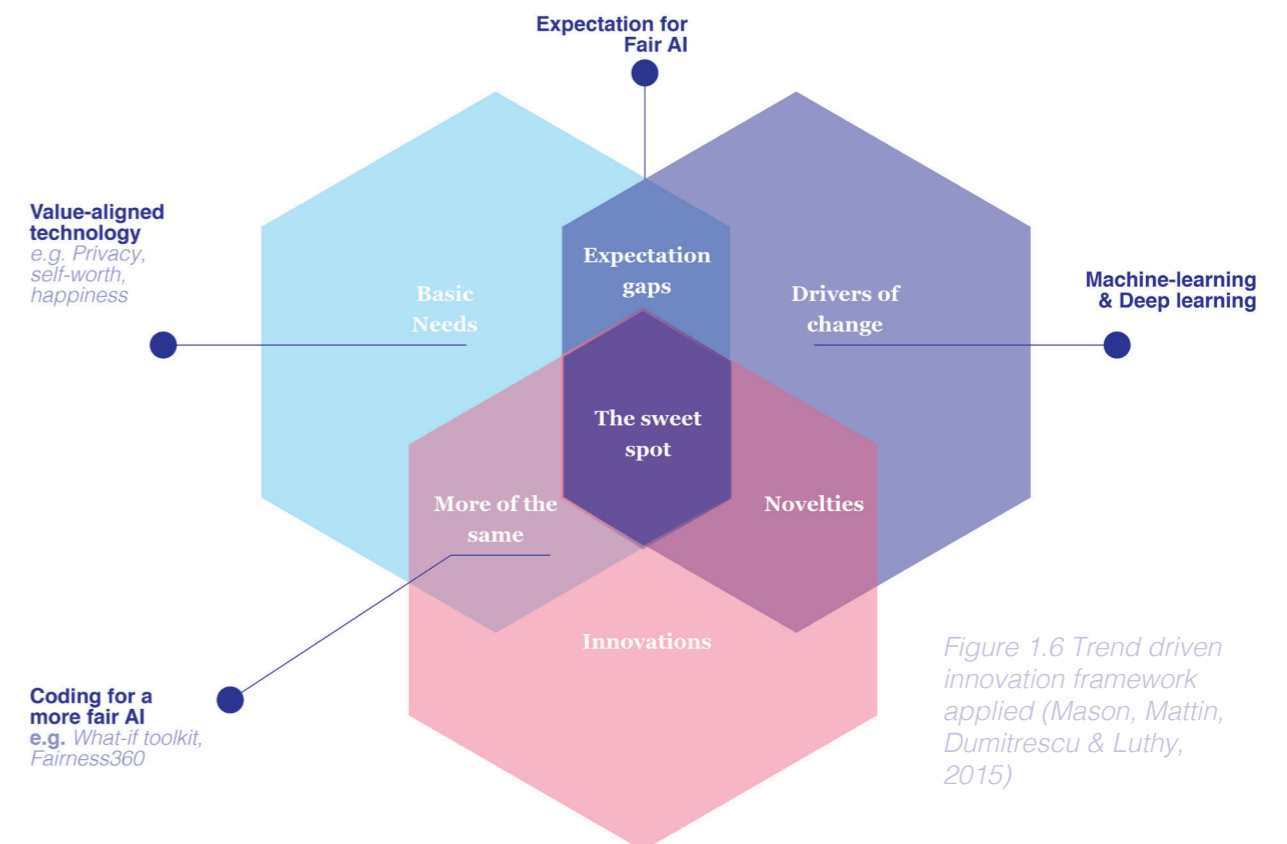
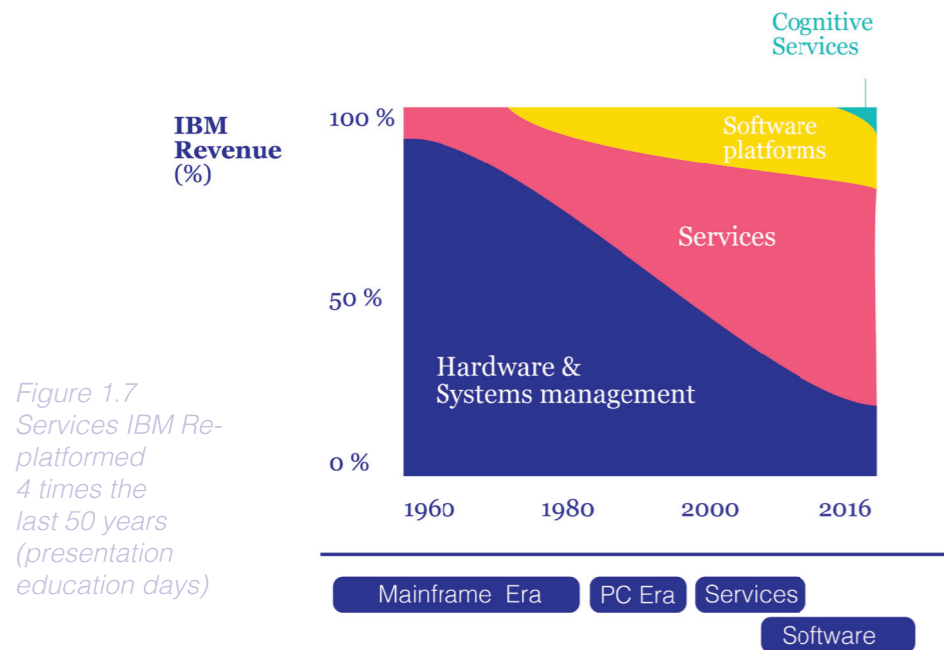


Figure 1.6 Trend driven innovation framework applied (Mason, Mattin, Dumitrescu & Luthy, 2015)



The three main IBM Values: **(1) Dedication to every clients' success;** **(2) Innovation that matter for our company and the world.** This represents IBM's believe in enhancing business, society and human conditions by the use of intelligence, reason and science. IBM aims to be the first in technology, business but also in responsible policy. Therefore, it is not afraid to take, sometimes the unpopular ideas; **(3) Trust and personal responsibility in all relationships.** This focuses on building sustainable trusted relationships, by following words by actions.

IBM AI Strategy

IBM's AI strategy focuses on three future pillars: **cloud platform, cognitive solutions and industry.** For the scope of this thesis is chosen to specifically look at the strategy of AI (cognitive solutions), which contains a prominent place at their strategic agenda.

Watson is IBM's suite of enterprise-ready AI services, applications and toolkits. IBM's strategy is focused on professional AI, in other words business to business. In Watson's strategy is deliberately put forward, the applications are aimed to **augment human intelligence** and not at replacing it. IBM strives to give the client control of data and the insights. Thus, the client

owns the trained algorithms. A quote of the CEO at the next page presents that vision.

» **Trust and transparency are key, in the overall AI strategy.** IBM currently puts much emphasis and research into AI ethics. This is also resembled in new product releases as the Fairness 360 toolkit (figure 1.10), released in September 2018, an open-source toolkit

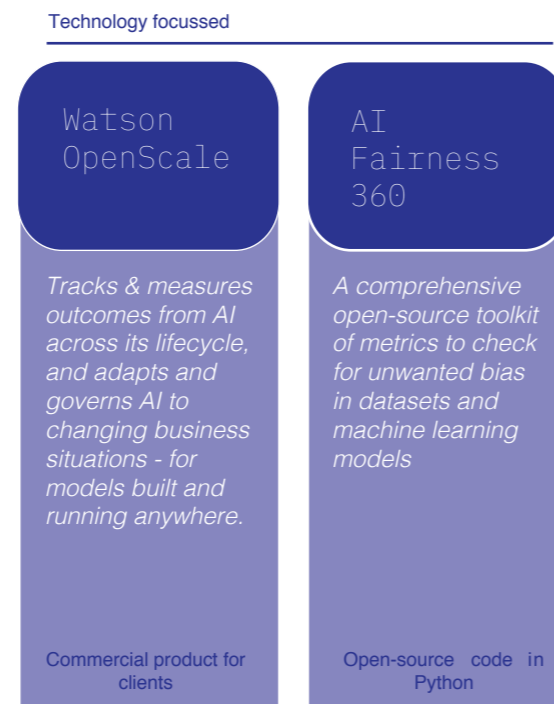


Figure 1.10 IBM AI ethics offerings

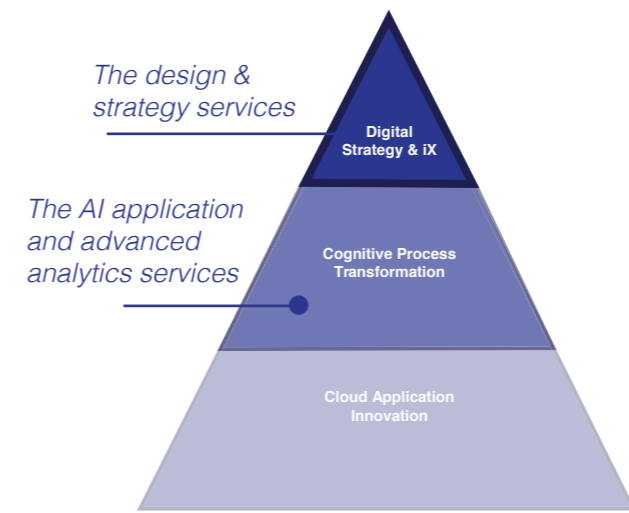


Figure 1.9 IBM global business services structure

of metrics to control for unwanted bias in datasets and machine learning models, with supportive algorithms to mitigate such bias. Correspondingly, IBM believes it has a duty to prevent, correct and monitor biases in algorithms as well as the ones caused by humans, believing in a open-source collaborative innovation. An illustration of this strategy is IBM's release of 1 million facial images, open for all to train AI systems with more diverse data sets.

» However, (IBM's) employees and clients working on these issues do not have practical support for the development of more ethical AI (applications) in their daily work at ethical decision moments. There is not yet a tool or process guiding the humans creating applications of these intelligent systems, from a non-technological perspective.

» A more ethical AI approach is a good strategic fit for IBM, ethics, human and societal benefits are in IBM's core values

Competition

From competitor analysis is extracted (appendix E), the biggest B2B players in AI are not

focusing on ethics in their AI strategy, leaving an opportunity area for IBM to differentiate themselves. This allows to distinguish with the ethical strategy, which is already in line with the internal values of IBM (internal analyses)

Center of Advanced Studies

This thesis is written in collaboration with Center for Advanced Studies Benelux. CAS is part of the innovation engine of IBM. It is positioned at the intersection of research, education and practice, creating an ecosystem of the business world and academia. The mission is fueled by innovation with the latest technologies and the goal is to become one of the biggest propulsions of innovation in the Benelux. It has novel knowledge concerning AI systems as well research track focused on AI fairness. However, at current mostly tackled from technology perspectives.

IBM Global business services structure

Due to IBM's size, solely a light is shed on the departments with the target groups is designed for, the AI teams (figure 1.9).

The global business services are divided into three "growth platforms", founded in the client needs: (1) Digital strategy & iX (Ds&iX), working at the intersection of innovative strategy, creative vision, and transformational technology. (2) cognitive process transformation (CPT), in which application of AI, automation are advanced analytics founded (and many others); (3) cloud application innovation (CAI), cloud migration, integration, enterprise Automation etc.

Ds & iX, is the growth platform in which the digital transformational strategies are created and the design studio's are placed in. In CPT the AI teams working at clients are situated. Every growth platform consists of multiple service lines.

» The two service lines focused on in this thesis due to the guiding AI polestar, are "cognitive process automation" and "cognitive analytics".

"Look, we really think this is about man and machine, not man vs. machine. This is an era really, an era that will play out for decades in front of us."

– Ginni Rometty (CEO IBM)

1.2

Project objective & approach

1.2.1 Project aim

The central aim of this thesis is to explore and design support for the AI teams with the creation of more ethical AI systems, bridging the gap between ethical AI principles and the current practice. By that, design for IBM's organizational capacity for the development of fairer AI by using strategic design and critical design approaches. The project aim is achieved by a novel approach of identifying and reducing unfairness sources of AI systems by explicitly resolving occurring value tensions at the ethical decision moments in the development processes.

The argument to put forth this thesis is that if IBM's employees and their clients are supported by practical tools to resolve value tensions explicitly and thereby reducing unfairness sources of AI, the outcomes will be more ethical and socially desired.

This has been enacted by the development of a new organizational role with an accompanying practical starters pack, to increase the current ethical uptake in AI practice.

1.2.2 Research questions

Currently there is a gap between ethical AI principles and AI practice. Hence, the main research question and respective sub-questions are as follows:

1 » How to create an organizational capacity and infrastructure to support ethical uptake in AI projects?

2 » How to support the AI team and in which phase, for fairness in AI projects?

3 » How to support the AI team and in which project phase for, value-alignment in AI projects?

3.2 » How to support the AI team to resolve value tensions in AI projects?

1.2.3 Project scope

There are several factors that play a role in determining the scope of the project:

01 IBM

This thesis is written in collaboration with IBM Benelux. The tools with accompanying organizational role are designed with IBM as a context.

02 AI Systems | Ethics perspective

In this thesis AI systems are solely examined from the perspective of ethical challenges in order to support for the creation of ethically aligned ones. Elements such as the algorithms themselves are not explored in depth.

03 Ethics | Fairness & Value-alignment

Ethics is a broad discipline and, for the aim of this thesis, only its foundations are explored. Thereafter is focused on the two specific ethical challenges for AI: Fairness & Value alignment. Important to mention that other ethical challenges might overlap on certain areas, however they are not focused on in this thesis.

1.2.4 The target group | AI team

In this project ways are explored and designed to support the AI team. The AI team in this project often consists of both IBM employees and of their clients. Often the team inheres: data scientists, product owners, managers and IT specialists.

1.2.5 Involved stakeholders

Next to the TU Delft supervisory team and the company mentor at IBM CAS, additional parties were involved. Clients of IBM supported the empirical research. Additionally, many experts in the field of ethics and philosophy of technology fueled the ethical knowledge of this project as well as supported the validation.

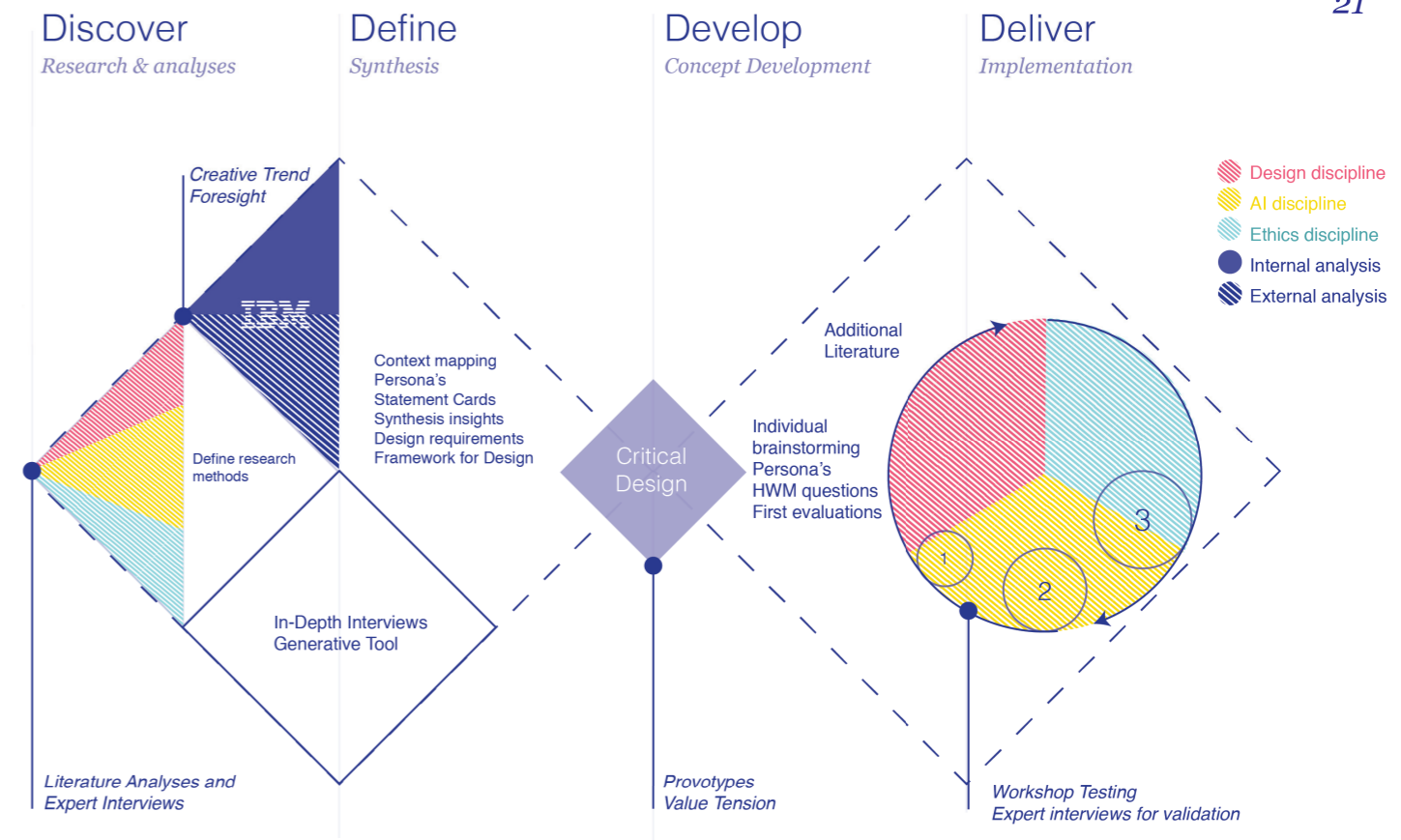


Figure 1.11 | Project approach and set-up

Design challenge I

Design practical support for AI teams to create fairer and value-aligned AI.

1.2.6 Strategic design approach

In the broadest sense, the double diamond process is used for this thesis due to its exploratory nature. In the diverse stages specific design activities, based on the challenge and goal were chosen. What distinguishes this thesis' approach is the bridging of two design domains; critical design and strategic design which are further discussed (figure 1.11).

Discover

The goal of this phase was understanding the AI technology, ethical processes of the ethics field and the project context (IBM). Many expert interviews were conducted from the fields of applied ethics, design and AI due to the newness of the topic. An extensive literature study was performed and interviews (with generative tools) with AI teams were held to understand

the contemporary field and empathize with the target group. This led towards a more specific research direction.

Define

The goal of this phase was to sharpen the design direction, (value tensions for fairness) and explore and inquire novel value tensions in AI development through prototypes (p. 87). Continuous expert interviews were held to keep up with the complexity of the topic. Together with the design research it led to the final chosen value tensions and a framework to design for fairness in AI by resolving value tensions.

Critical design I "is used as a medium to engage user audiences and provoke debate. It does this by encouraging its audiences to think critically about these engendered in the design work." (Malpass, 2017).

In this thesis critical design is chosen to fuel the strategic ideation and solution spaces in a

form of designed provotypes for inquiry of value tensions in AI and their hierarchies.

Develop

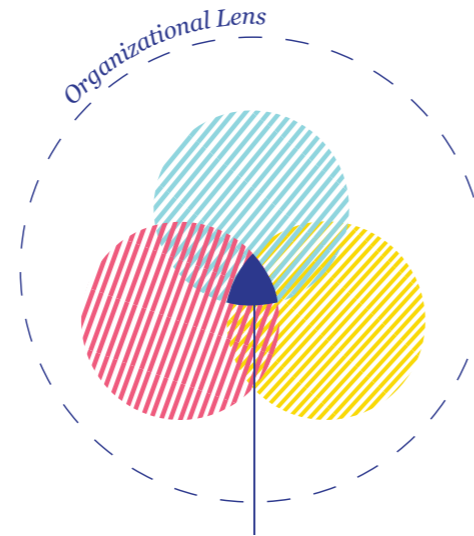
In this phase, the design is iterated upon multiple times. This was a workshop setting which tested twice in the ideation phase, from a variety of perspectives, design and AI ones. Individual brainstorming, how might we questions, persona's and expert interviews were used to lead from the initial concepts to the final design concept.

Deliver

Final tests with AI teams and clients of IBM are conducted to validate the concept. In this phase the final proposition of the design was created together with implementation ideas and corresponding recommendations. A variety of interviews were conducted at IBM to design a strategic fit.

1.2.7 At the intersection of three disciplines

Theoretical background of three main disciplines is gathered in the course of this thesis with the organizational lens. The fields of design (general, strategic and critical), ethics (applied ethics, philosophy of technology) and AI (Fairness). This leads to an extensive literature study presented in chapters 2,3,4,5.



At the intersection of three domains

Figure 1.12 | Visualization intersection of the main disciplines touched upon in this thesis.

- Design discipline
- AI discipline
- Ethics discipline

“A.I. systems are shaped by the priorities and prejudices—conscious and unconscious—of the people who design them, a phenomenon that I refer to as “The coded gaze”

~ Joy Buolamwini (Buolamwini, 2018)



- Davide Aurucci | **Data Scientist**
- Archana Nottamkandath | **Managing Consultant Watson and Advanced Analytics**
- Dorottya Mezofi | **Senior Consultant iX**
- Jory Wielaard | **Strategy Consultant iX**
- Cristina Meniuc | **Senior Designer**
- Sophie Kuijt | **Ethics ambassador NL**
- Jonathan Leung | **UX designer iX**
- Nicky Hekster | **Technical Presales & Business Development EMEA Watson Health**
- Mando Rotman | **Practice Leader Cognitive & Analytics Benelux**
- Jeroen van den Hoven | **Editor in chief of Ethics and Information Technology TU Delft**
- Brian Goehring | **Associate Partner, AI Cognitive & Analytics Lead, IBV, US**
- Jack Esselink | **DS & Machine Learning Evangelist**
- Madli Uutma | **Data Scientist**
- Dolf Noordman | **Data Scientist**
- Lammert Kamphuis | **Philosopher at Lammert Kamphuis**
- Saniya ben Hassen | **Technology Architect**
- Jan Ploeg | **Cognitive Business Decision Support Strategy Consultant**
- Harry Langevoort | **Cognitive Industry Solutions FSS**

● Aimee van Wynsberge | **Member European Commission high level expert group on AI, Co-director Foundation for Responsible Robotics**

- Rob Nijman | **Client Executive, Government Sector Business**
- Vincent Vijn | **UX Research & Design Lead iX**
- Sophie Kuijt | **Ethics ambassador**

- Gerlof du Bois | **Sector Leader Public & Health**
- Gerard Smit | **CTO & TSE Benelux**
- Reggie van de Westelaken | **CIO - Mobile Europe**
- US Design team | **AI Design Practices**
- Mark Esseboom | **Director Government and Regulatory Affairs IBM Benelux**
- US Design team | **AI Design Practices**

1.2.8 Expert interviews & events

AI and (applied) ethics are enormous fields to grasp. Therefore, expert interviews are conducted in this thesis as means of information retrieval. Figure 1.13 represents the overview of the expert interviews held during the course of this thesis.

- Discover stage (8 interviews)
- Define stage (10 interviews)
- Develop stage (4 interviews)
- Define stage (6 interviews)

The appendix shares a more detailed overview, also concerning the informative events and presentations given and vided.

Figure 1.13. represents solely the expert interviews. Semi-structured interviews held in the design research are described in chapter 6. Chapter 9 shares the interviews for validation of the design.

- Design discipline
- AI discipline
- Ethics discipline
- Internal analysis
- External analysis

Supervisory team who were constantly involved:
Elisa Giaccardi, Lianne Simonse & Zoltán Szlávik

Figure 1.13 | Visualization of expert interviews held during the course of this thesis.

01 Project Context & Approach

AI Discipline is rising so are the consequences

Both research and businesses concerning AI are rising. More decisions are left to machines and algorithms, which have consequential outcomes. Questions of values and ethics become urgent, as the systems ethically misaligned.

The need of ethics flavor

Companies and humans learned by mistakes, in practice, with ethically misaligned AI systems. There is a need for the incorporation of ethics in AI systems development.

IBM I Trust & ethics in their veins

IBM's believes in enhancing business, society and human conditions by the use of intelligence, reason and science. It is not afraid to take, sometimes the unpopular ideas for societal benefit. Thus, a more ethical direction towards the development of AI is a strategic fit.

Technology perspective to ethics

Currently AI ethics is tackled by technological endeavors. This leaves out untouched upon approaches from more human perspectives.

Strategic bottom up approach

There is a need to integrate societies perspective, human thinking and context, not only from principles but from support for AI teams making these systems. This provides the strategy and design direction for my project

» Overall, the strategic bottom up, human approach, provides strategy and design directions for my project that distinguishes IBM's approach from its competitors on the long term and aligns AI systems with human and societies values. (internal & external analyses)

Strategic design fueled with critical design approach

Due to the newness, abstractness and complexity of the topic, a combination of both strategic design and critical design are chosen to fuel the ideation and design research spaces.

Design & Applied ethics & AI

Theoretical background of these three main disciplines is gathered in the course of this thesis with the organizational lens.

Chapter 02 |

Introduction to AI

This chapter aims to bring a basic understanding of AI systems to further design for the AI teams. This is based on a literature dive in the AI field, that besides articles included books about AI development, presentations and online AI courses. It is not meant to give a complete overview of the AI field, but rather to reach a level of understanding and reasoning in order to support the current process in an ethical fashion. Therefore, the current capabilities and challenges are discussed.

In this chapter |

- 2.1 AI foundation
- 2.2 AI Dish Metaphor

2.1 AI foundation

2.1.1 Alness

The term artificial intelligence (AI) is brought up by John McCarthy and others in 1956. The field has been strongly increasing the last years (appendix C and figure 1.1).

However, the distinction between AI and non-AI is not black and white. To understand AI, it is important to realize, that AI does not have one single dimension as temperature does. Someone can compare the temperature of Amsterdam to Cape Town and tell the differences. Someone can compare yesterday's cold weather with today's hot one. People tend to do the same for AI, while it cannot be compared on a single axes or dimension. It does not make sense to compare the intelligence of a spam filter to a movie recommendation system. AI is narrow and because it is able to solve one type of problem it does not say anything about solving another (Roos et al., 2018). There are also products that involve a bit of AI, one could say a bit of Alness. Additionally, AI is not a countable noun. It is more a scientific discipline as physics, meaning it is a collection name for diverse concepts problems and methods to solve them.

Definition AI I *"The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it."* (McCarthy et al., 1955). In other words, it breaks down elements of intelligence into smaller steps, that can be described by coding. It solves well-defined and isolated problems, solving them one at the time instead of all at once

2.1.2 Current capabilities of AI

Andrew NG, computer scientist and co-founder of Google brain explains what AI currently can do:

"Any cognitive process that takes a human under one second to process is a potential candidate for AI" (NG, 2017)

Misconceptions about the current state of AI exist. AI presents state, is narrow intelligence. In other words, specified to one specific task in one industry. The capabilities of narrow intelligence are visualized in figure 2.1 based on the book and article of Burgess (2017) and Snoek, (2018). The different types of AI are shortly addressed in the next paragraphs.

3.1 General & Narrow AI systems

Narrow AI handles one very specific tasks. General AI or Artificial examples of narrow AI. However often the pop-culture including dystopian visions refers to general AI. Hence, the most developments are in the field of narrow AI (Roos et al., 2018; Burgess, 2017).

3.2 Strong & Weak AI

Strong and weak AI is based on the difference of being intelligent and acting intelligently (Searle, 1990). Strong AI therefore refers to a mind which is self-conscious and genuinely intelligent. Currently humans use and develop weak AI systems, exhibiting intelligent behaviors (Russell and Norvig, 2016).

» AI systems can solve a narrow-defined task for which it is trained very well. It is often

NARROW AI

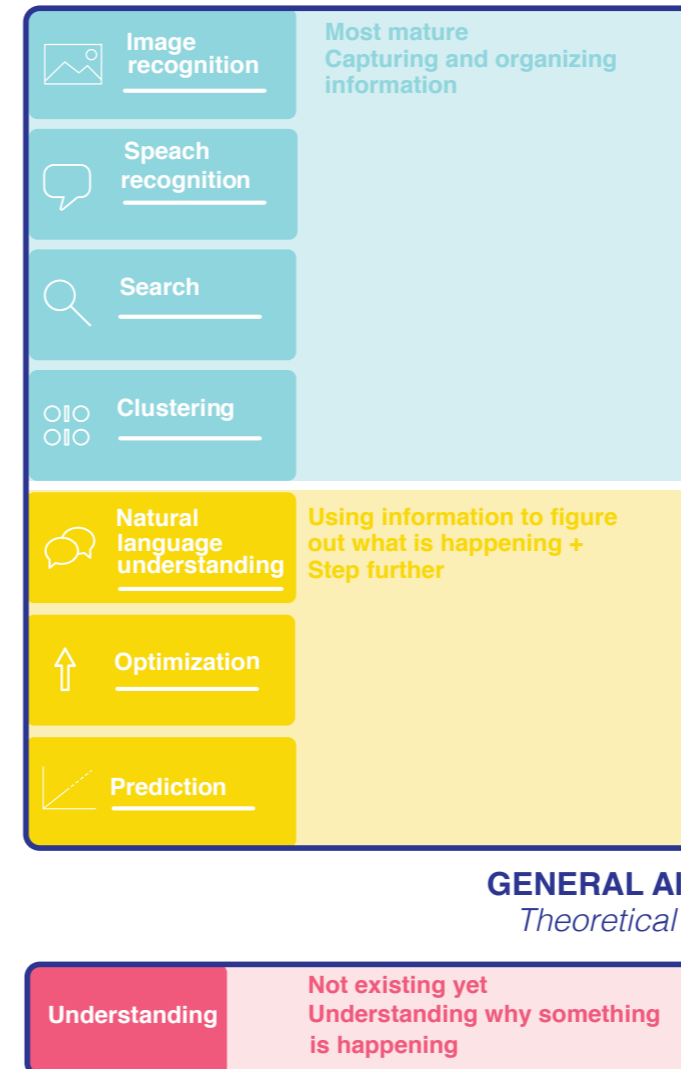


Figure 2.1 | Capabilities of AI systems



Figure 2.2 | Artwork made by an AI system

incredibly good in aspects humans are less good at, such as really fast in calculating or scanning through much historic data. AI systems are particularly bad in activities humans are good at, such as understanding, reacting to actions in a context specific manner etc.

2.1.3 Machine learning

"It has been long understood that learning is a key element of intelligence. This holds both for natural intelligence - we all get smarter by learning - and artificial intelligence." (Roos et al., 2018).

Since 1950, within the field of AI much development has been made. Both the fields of machine learning and deep learning developed (figure 2.3).

» Machine learning systems are systems that can advance their performance due to gaining more experience/data, for a particular use case. It can employ data, to result in knowledge, novel patterns and generate models through a set of methodologies/techniques, which are referred to as machine learning. In other words, a machines capacity to modify or define decision making rules by itself, autonomously, is machine learning. This can be used for example for predictions (Van Otterlo, 2013).

In figure 2.4, the three most common approaches to learning of machine systems are visualized as a synthesis from literature and AI courses. It gives an overview of the current types of learning and the necessities for it. **Different types of learnings have their own challenges.** Currently, supervised

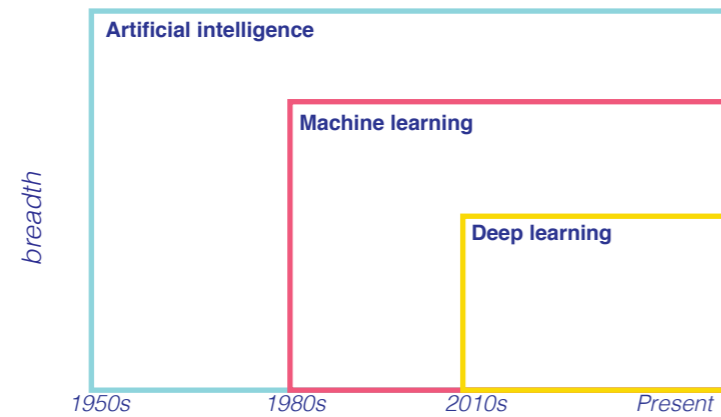


Figure 2.3 |
3 Main disciplines
within AI

learning is the most relevant AI category in terms of business impact (Pereira, 2018).

To teach machines, one needs much “practice exercises” and therefore you need much data. Therefore, massive accurately labeled data sets are needed, of which is a lack (McKinsey, 2018).

» While training systems made on incomplete data-sets lead to systems that are less well prepared for the real world. It possibly leads to biases and unethical systems discussed in later chapters.

The supervised learning approach is a human in the loop method and mostly used in cases of costly errors, class imbalances, or little initial data. A human operator controls the semi-automated processes and supervises it for two main goals: (1) The human can identify errors or misbehavior and (2) secondly can take a corrective action (Rahwan, 2018). This can lead both to notions to make AI more ethical when applied consciously, nevertheless when performed unconsciously, also unwanted biases can be embedded, which are discussed further.

Deep learning refers to specific machine learning techniques that uses “neural networks “ in the learning process. It uses several layers of more simple processing units connected by a network, imitating the way our brain processes through our eyes. This type of network use stimulates the creation of more complex “machines” without the need of tremendous amounts of data (Roos et al., 2018). However, these types of techniques

often have the downside of being a “black-box” model. In other words, the decision making of the system cannot be completely traced back.

2.1.4 Challenges in AI systems

With implementation and creation of complex AI systems, challenges arise. These are classified into six main categories, based on the current literature review.

1 Oversimplification

Flasinski (2016) teaches us that the translation of the complex world, humans live in, with the many context specific situations that we are in, are impossible to translate into code. This leads to oversimplification of the actual world and to systems that work in narrow contexts.

2 Elite group of developers

Also, the people who actually make these systems are a quite a homogeneous group of people, and that represent a small part of our societies. There is a lack of diversity in the development process of AI leading to systems which are not fulfilling desires of the population, not representing the values of societies (Flasiński, 2016).

3 Strong technology push

With the dominantly technological expertise of the AI team, there is also a strong technology push perspective in the development of AI, neglecting the needs and human perspective and

Supervised learning

Humans teaching machines

One teaches the system with most often human provided feedback.

Needs learning data (dog/no dog)
Training data and test data

Feedback needed

Used with unstructured or
semi-structured data

Examples

Classification | Fraud detection, image classification, medical diagnosis, gender detection
Regression | Weather forecasting, sales growth prediction, market forecasting, estimating life expectancy

Unsupervised learning

Learning without a teacher

Makes new connections in the data. The system starts with a very large data set that will mean nothing to it. The AI will spot cluster of similar points in the data

Used with unstructured data

No feedback needed, no labels no correct outputs

Is a black box

Examples

Clustering | Customer segmentation, recommender systems, targeted marketing
Generative networks | 2D to 3D modelling, pattern modelling, image generation, music generation.
Dimension reduction | Big data visualisation, structure discovery, feature extraction,

Reinforcement learning

Trial & error to build experience

Used when an AI agent must operate and in cases where feedback about good or bad choices has delay. Used in situations there are many unknowns (Self-driving cars, some games)

Uses extreme trial and error to update its experience, using this experience to determine the most optimal next step towards the goal

No feedback given

Examples

Real time decisions, skill acquisition (robotics), robot navigation, game AI

Figure 2.4 | Main types of learning of AI

using technology for the sake of using it (Internal interviews IBM). This is also due to the fact that there is lack of expertise in the translation of AI systems and the business human (Burgess, 2017).

4 Lack of AI understanding

» A lack of knowledge and expertise about AI, from the client’s side, the team, the management and sales departments is mentioned as a hurdle in AI development (almost all preliminary informal interviews within IBM). It leads to disappointments, too ambitious deadlines as well as undesired, unethical outputs.

5 Sky-high expectations

Burgess (2017) additionally mentions that the biggest barriers in the development of AI is due to the increasing expectations that do not match reality.

6 Ethical challenges

» However, the most mentioned challenge of AI is concerning ethics. Examples are mentioned such as, biased AI, non-explainable AI, unfair AI are the biggest challenges the AI field is currently facing (Verbeek, 2014; Gonzalez, 2015; van den Hoven, 2015; Schatsky & Schwartz, 2015; Banavar, 2016; O’Neil, 2016; Boddington, 2017; Fast & Horviz, 2017; Erdelyi, 2018; McKinsey, 2018).

2.1.5 Conclusion

This section provides a basic understanding of AI systems and the challenges they bring with. The AI field has seen great developments over the last decade, however these developments also bring along great challenges.

2.2 The AI Dish Metaphor

A metaphor is created to explain AI and its ethical challenges in a playful and relatable fashion in this thesis. This metaphor is the one of an AI dish. It is referred to both in the report and design. It is further elaborated upon in this section.

Ingredients of AI

From a technical perspective, an AI system consists of three aspects. These are:

- the data on which the model is made
- the algorithm which is programmed for example in the Python language
- the computing power needed to run the system

» In this thesis the metaphor of cooking is used to further explain AI systems and what affects their quality.



Data is the ingredient



Computing power is the fire



Algorithms are the cooking appliances



Models are the recipes



The output are the final dishes

If a chef is cooking a dish, and the ingredients are of poor quality, the final dish will obviously not taste good. The same counts for an AI system, when the data is not of good quality, the output

of the system will neither be of good quality. In cooking, the appliance can be used for multiple purposes and multiple dishes. Even though it influences the quality of the dish (such as a mixer does for texture of the dish), it does it less than the ingredients. In line in AI development, the same algorithms can be used by different companies for different use cases. Many packages of these algorithms are open source and anyone understanding the code could use them. However, a blender without food does not bring any value. Similarly, an algorithm is not much worth without the data.

Also the chef and the people making the dish influence the taste of the dish hugely. In line, in AI development the data scientists and developers impact the final outcome (un)consciously.

02 Introduction to AI

AI Ingredients

In this thesis a metaphor is used to explain AI systems, the AI dish (figure 2.4). The data is the ingredient, the fire the computing power, the appliances are the algorithm and the dish are the output of the system. If the ingredients are of poor quality the dish will be too. Similarly in AI when the data is of poor quality the AI output will be too.

Machine learning techniques

The three main learning techniques for AI systems are supervised learning (learning with a teacher), unsupervised learning (learning without a teacher) and reinforcement learning (trial and error). More advanced types of learning systems such as deep neural networks have good performance but hands in on explainability.

Narrow AI

Narrow AI handles one very specific tasks, which is the current state of AI systems. The current capabilities of AI systems are image recognition, speech recognition, search, clustering, natural language understanding, optimization and prediction. Current AI systems are incapable of understanding.

AI Challenges

The AI field has seen great developments over the last decade, however these developments also bring along great challenges. One of the most mentioned ones is the ethical challenges that AI systems bring with. In the current AI development the human heart-beat is lacking.



Figure 2.5. | AI ingredients of the AI Dish metaphor

Chapter 03 I

Ethics flavor

Establishing the ethics foundation

The field of ethics is explored in order to research how to support the AI team in the development of fairer and more value aligned AI systems. This led to three main ethical building blocks. Also, a light is shed on the AI ethics and design ethics. These sections share the insights how these three fields can complement each other.

In this chapter

1. A taste of ethics
2. Ethics & AI
3. Ethics & Design

Ethics I *One could see ethics as reflection upon our morals (J.van den Hoven, personal communication, October 18, 2018), usually in terms of right obligations, benefits to society, fairness or specific virtues” (Verlasqueez et al., 1987, para. 9)*

Ethical framework I *” built around delineating rights or obligations, estimating benefits to society, determining fairness or developing virtues – can help us make decision between competing values and recognize values that advance human flourishing” (Shilton, 2018).*

Ethical pluralism I *“recognized that there are some universal values such as wisdom and peace but also recognize that the degree of importance of each of these values in a culture or in an individual may vary” (Borning and Muller, 2012)*

Moral problem I *“ that there are two or more positive moral values or norms that cannot be fully realized at the same time. A good moral question meets three conditions: (1) it must clearly state what the problem is, (2) it must state for whom it is a problem and, finally, (3) the moral nature of the problem need to be articulated” (Van de Poel, & Royakkers, 2007). A simpler way to describe it as conflicts of rights, values or professional responsibilities (Dorst & Royaakkers, 2006)*

Moral I *Relating to the standards of good or bad behavior, fairness, honesty etc. that each person believes in rather than to laws (Cambridge dictionary)*

Norm I *An accepted standard or a way of behaving or doing things that most people agree with (Cambridge dictionary)*

“Human lives and societies are co-constituted and co-shaped by technology”

- van den Hoven, 2017

3.1



A taste of ethics

By means of literature review, tool & method analyses and expert interviews this section is constructed. An organizational lens is kept in mind, exploring manners to support ethical AI processes from an organizational and strategic standpoint. This led to the three main ethical building blocks that are elaborated upon and visualized in figure 3.4.

3.1.1 What is ethics?

In business, ethics is often preferred to be bypassed, both due to its complex nature and the questions it arises (Davis and Patterson, 2012). However, this thesis argues in line with Boddington (2017), it should be seen more positively and promoting as it can enable new innovation opportunities.

Ethical issues are difficult to describe by empirical reality and concern topics that are difficult to weight, thus these stimulate philosophical debate. Often it concerns decisions where two or more positive moral values cannot be realized simultaneously. Ethics is never finished, and values are in flux, depended on the context and inconsistent for many individuals (Boddington, 2017).

» **This thesis takes the perspective of ethical pluralism,** “recognizing that there are some universal values such as wisdom and peace but also recognize that importance of each of these values in a culture or in an individual may vary” (Borning and Muller, 2012).

Often ethics is brought up in terms of right obligations, benefits to society, fairness or specific virtues (Verlasqueez et al., 1987). The following definition for ethics is chosen due to the reflective nature:

» **Ethics is mostly focused on normative issues and people could see ethics as reflection upon our morals** (J.van den Hoven, personal communication, October 18, 2018). It is in contrast to prescribing definitions of what is

right or wrong.

» **This is in line with the design ethics perspective. In which ethics is addressed as a mindset rather than a prescribed framework of right or wrong** (Gispen, 2017).

This thesis is written from a design perspective and takes context specific nuances into account for the a more ethical AI development therefore these definitions fit most.

3.1.2 Ethics strategic benefit

Strategic alignment of ethical values and actions is profitable for business (Shilton, 2018).

» **An ethical strategy gives a strong sustainable competitive advantage in the market, is proved by research on the longer term.** Therefore incorporating ethics is also a compelling strategy also from the business perspective.

3.1.3 Necessities for an ethical company

A few necessities for an ethical company and thereby outcome are discussed. First, an **ethical company culture** is crucial before for ethical outcomes. Elements that might support an ethical company culture are: **a diverse team, an ethical mindset by empowering employees to do the right thing, being transparent, take feedback and understanding the companies’ values by examining outcomes and trade-offs of value-based decisions (Baxter, 2018).** Second, **moral motivation** is crucial for ethical decisions. Research shows there is no

	Theory basis	Advantage	Disadvantage
Utilitarian <i>consequence based</i> <i>J.S. Mill</i>	Debates that the action bringing the best consequences is the right one. It aims to bring the greatest balance of (un)happiness for the largest number of people (Boddington, 2017).	Promotes utility and happiness	Ignores views of minorities
Deontology <i>Duty-based</i> <i>Kant</i>	Argues that an action should be in line with the general overarching principle(s), to be the right action. (Boddington, 2017).	Promotes responsibility and respect towards other people	Less concerned with happiness and social utility
Virtue <i>Character-based</i> <i>Aristotle</i>	Believes in the ideal moral agent, by describing the specific virtues one has and the right thing to do would be the one what the virtuous person would do. (Boddington, 2017).	Promotes moral education and character development	Needs homogenous community standards

Figure 3.1 | Most common ethical judgment theories

correlation with moral motivation and IQ (Moutafi et al. 2004). Thus, because someone has a high IQ (AI team members often do), it does not mean that they have motivation to choose ethically. This is in line in with designing “ethically”, it is concerned with a mindset rather of specific ethical topics.

3.1.4 Ethical engineers

Designers and engineers (un)consciously design with their values and morals, thus the technology they develop reflects that (concluded from the analyzed philosophical theories (appendix H).
» Therefore, they should be morally responsible engineers and incorporate ethical wisdom (Burg & Gorp, 2005; Van de Poel & Van Gorp, 2006; van den Hoven, 2017; Shilton, 2018). In AI development it means that the AI team should be morally responsible and know the ethical basics in order to lead to more ethically aligned systems.

3.1.5 Ethical decision moments

Ethics is concerned with decisions. Thus, when connecting the AI development process and ethics, it is crucial to identify moments when decisions turned into actions, entitled as ethical decision points.
» Identifying these moments is the first step towards are more formal ethical process and development (Davis & Patterson, 2012).

3.1.6 Ethical judgment

When dealing with moral dilemmas or ethical choices, people can use ethical judgment to make a decision. Approaches of ethical judgment can be divided into two streams, formal and informal ones. Three well-known normative ethical theories are described briefly in figure 3.1.

Approaches to ethical judgement	
Formal <i>Deontology</i> <i>Utilitarian</i> <i>Virtue etc</i>	Informal <i>Intuition based</i> <i>Dominant-value method</i>

Figure 3.2 | Formal and informal ethical judgment

Formal ethical judgment

Figure 3.1 presents an overview of the three normative ethical theories, with the advantages and disadvantages. These are the most common philosophical views and can give support for ethical judgment and choices in an AI development process. It gives handles for switching perspectives and might support the challenging issues in AI ethics.

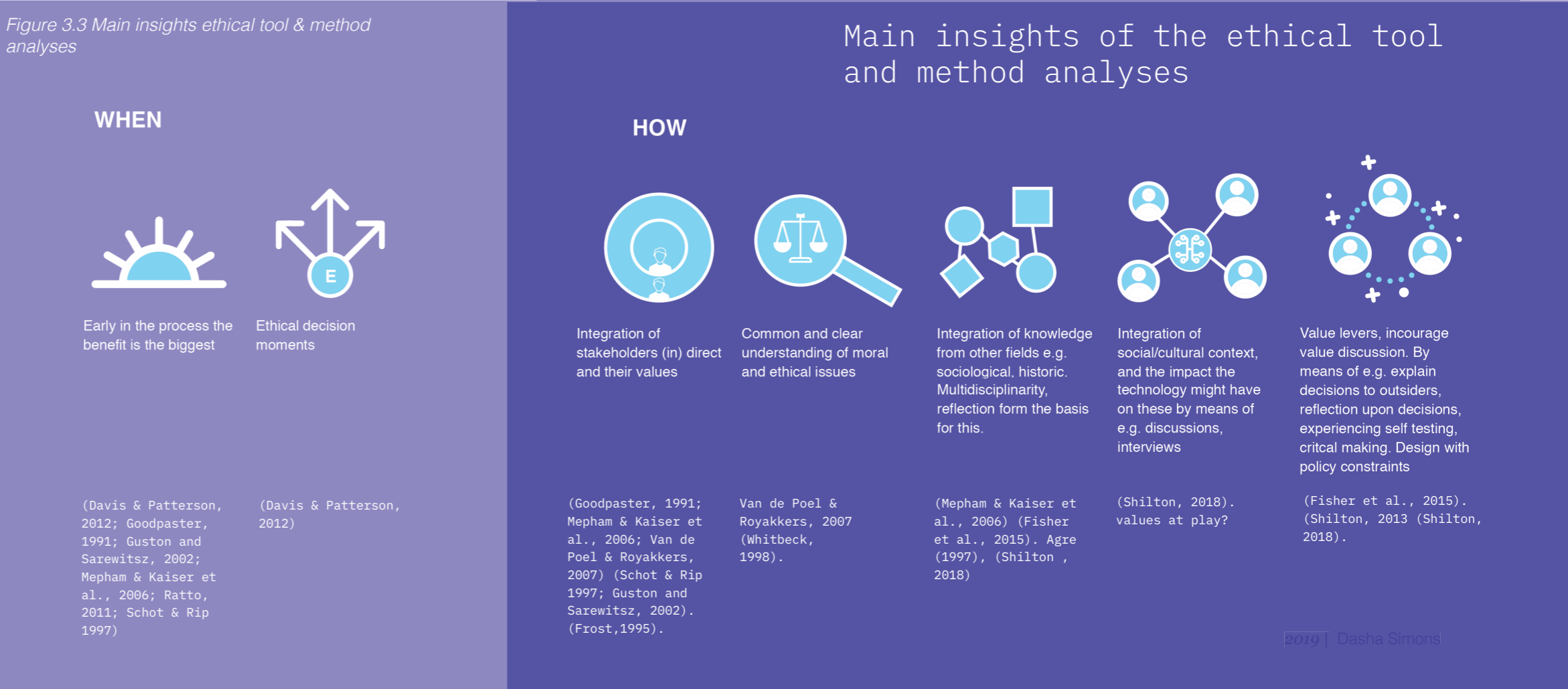
Informal ethical judgment

Next to formal ways of ethical judgment, also informal ethical frameworks exist (van de Poel & Royakkers, 2007). Two informal frameworks are briefly touched upon: the intuitions one and the dominant-value one. **Intuition based** ethical

judgment works with the action which is on an individual view, intuitively most adequate (van de Poel & Royakkers, 2007). While the basis of the **dominant-value method** is a favored value in a specific case. The concept is that in specific cases one value is predominant. Determining the dominant value, can support the use of certain guidelines.

3.1.7 Ethical tools & methods analyses
Ethical tools and processes are the foundation for ethical decision making. They are crucially important during the development process of technology. Therefore, general and engineering ethics tools are analyzed to gather an understanding of ethics implementation in projects. These are analyzed with the goals of the research question in mind. The following general ethical tools and approaches are analyzed: stakeholder analyses (Goodpaster, 1991), ethical matrix (Mephram & Kaiser et al., 2006) and ethical cycle (Van de Poel & Royakkers, 2007). Ethical methods and

Figure 3.3 Main insights ethical tool & method analyses





Strategic Ethical Building Blocks



Ethical Company

- Create an ethical company culture
- Create a diverse team
- Create an ethical mindset by empowering employees to do the right thing
- Be transparent
- Take feedback
- Understand the companies' values by examining outcomes and trade-offs of value-based decisions



Ethical Process & Tools

- Early in the process the biggest benefit
- Direct & Indirect stakeholder integration and their values
- A common and clear understanding of ethical issues is needed for a good ethical deliberation process
- External knowledge by information or a full time expert in the team
- Integration of social/cultural context
- Integration of the technology impact
- Value discussion by: explanation decisions to outsiders and reflecting upon, experiencing self-testing of systems, critical making for developers to experience socio-technical challenges, designing with not only technological constraints put also policy ones



Ethical People

Employees & clients

- Have moral motivation
- Consciously design with their values and morals
- Are moral responsible engineers
- Incorporate ethical wisdom (awareness & education)
- Two orders of reflective learning (1st and 2nd)

Figure 3.4 Strategic building blocks for ethical organizational capacity divided into three categories based on the analysis of the ethical tools and methods

Ethical process

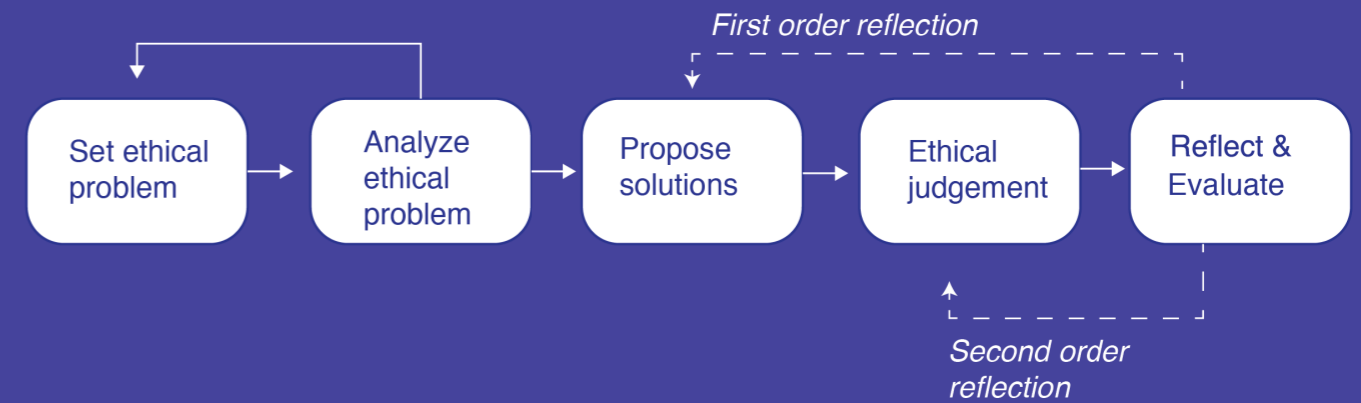


Figure 3.5 The ethical deliberation process extracted from the literature

approaches in engineering, often focus on the impact the technology has ethically on society (Guston and Sarewitz, 2002). The following ethical tools and approaches from engineering are analyzed: Constructive/real-time technology assessment (CTA) (Guston and Sarewitz, 2002), Socio-Technical integration research (STIR) (Fisher et al., 2015), Critical technical practices/reflective design/critical making (Agre, 1997), a Sartrean model (d'Anjou, 2011), Value advocate.

The goal of the analysis is threefold. First, distill the ethical process to understand the steps and translate them towards the AI process (figure 3.5). Second to discover when the ethical tools and approaches are used to discover when it is most beneficial, to tailor it for the AI development. Third, how ethics is integrated and applied in the current methods and tools. The results are visualized in figure 3.3. The full analyses is shared in appendix I. Briefly a summary is explained by the structure of the three goals.

(1) Most all of the mentioned tools, do not lead to one particular answer but rather prompt discussion and integration of certain types of knowledge to guide the people to answer. Difference in opinions with good founded arguments seem to be basis for most of these tools, given form in more systematic ways of addressing, often iteratively. Figure 3.5 shows the distilled ethical deliberation process. Starting by (1) setting the ethical problem,

then (2) analyzing the problem, (3) proposing solutions, (4) ethical judgment about these solutions and (5) reflect and evaluate these. Fisher et al. 2006, proposes a framework for intervention-oriented activities to improve and make clear the “responsive capacity” (Fisher et al. 2006). It distinguishes two types of reflective learning. The first order reflective learning is “improvement of the technology and the improved achievement of one’s own interests in the network.” And second order reflective learning “requires a person to reflect on his or her background theories and value system” are improved with the use of this framework (Van de Poel and Zwart 2009, p. 7). Thus, one’s individual reflective capacity and learning affect the ethical dimension of an outcome. This view increases the designers’ responsibility.

(2) The analyses results show that integration of ethics early in the process has most benefit at ethical decision moments (figure 3.3).

(3) From the analyses five elements for how ethics is integrated are distilled (figure 3.3).

(1) The integration of stakeholders (in)direct and their values, (2) Common understanding of ethical issues, (3) Integration of knowledge from other disciplines, (4) Integration of socio/cultural impact, (5) Value levers to create moments for discussion by means of e.g. describing decisions, critical making, self testing.

Furthermore, literature points out that giving a team member the explicit responsibility of ethics

and values during the technology development process, has beneficial ethical results (Fisher and Mahajan, 2010; Manders-Huits and Zimmer, 2012; van Wynsberghe and Robbins, 2014; Shilton and Anderson, 2017). Also, value consciousness an explicit responsibility of the design, helps to build values reflection into the scope of work and the success metrics of a team (Shilton, 2018).

3.1.8 Three building blocks

This section led to the synthesis of insights, with the organizational lens in mind, to three main strategic ethical building blocks (figure 3.4). These are building block necessary for organizational capacity to support more ethical outcomes of (AI) team processes. These are the following: (1) ethical organizations; by stimulating ethical culture and behavior (2) ethical processes and tools; shares ways to incorporate ethics in the development processes; (3) ethical people, concerns the skills, awareness and knowledge people should have for more ethical development outcomes. These together form the basis of the proposition for the ethical AI organizational capacity.

3.1.9 Limited uptake practice

A current reoccurring impediment how to bridge ethics into the coding. Although, most tools and methods aim to structure the ethical process, many approaches are theoretical and little empirically tested with industry. Also, a clear translation of incorporations of “abstract” ethics into the AI development process, and coding is currently lacking (van den Hoven 2013; Shilton, 2018). **Furthermore, Spierkermann (2015) mentions that the current agile development methods are less suited for the integration of ethics during development (Spiekermann, 2015). Research describes it as a challenge to prioritize ethics in a technology industry which is currently dominated my market values such as efficiency and speed**

(Shilton, 2018). Simultaneously, attempts to shape our technologies are often too late and too slow (van den Hoven, 2017) e.g. regulation and policy.

Thus, the applied ethics field needs novel perspectives and approaches that will increase the uptake in practice and in more agile processes.

» Currently, little approaches start from the day to day work of the people creating the technology. This thesis argues that a closer integration of ethics in the day to day process might increase the uptake of ethical consideration and implementation and thereby lead to more ethically desired outcomes.

3.1.10 Conclusion

To conclude, three strategic ethical building blocks are derived to create ethical organizational capacity to support more ethical (AI) development, as a result of this literature study (figure 3.4). Furthermore this research extracted when ethical aid in the process is needed (figure 3.3) and provided a consolidation of the ethical deliberation process (figure 3.5). Despite a considerable amount of research interest in ethics, there is a lack of incorporation of ethics in decisions in commercial development, even though it has competitive and strategic benefits for companies.

» **In AI, the necessity to incorporate ethics is strong due to the increasing impact it has on our society. There is a need for improvement in ethical uptake in AI development, aligned with the daily work of the AI team and the current agile manners of working. In this thesis it is argued that ethics in AI might benefit from a design perspective, which is elaborated upon in section 3.3.**

3.2 AI ethics

In this section a light is shed on the contemporary AI ethics field and the main challenges it copes with.

3.2.1 Ethical Challenges of AI

There are specific ethical challenges that arise with the development of AI systems. Questions of values and ethics become urgent, as the AI systems can be negatively biased, the decision processes often not traceable, while impacting our lives. This realization evoked a strong need for more ethical AI systems and new manners to create these.

Contemporary AI ethics field

Referring back to the AI chapter and the ethical building blocks, **much work needs to be done in the field of ethical AI**. AI systems are often made by a select homogeneous group of people, with little integration of stakeholders, no explicit ethical issues nor moral motivation of the team. While these are the foundation of ethical processes and thereby prompt more ethical outcomes. Additionally, AI and ethics have a challenging relationship as often ethics is seen as a constraint rather than an opportunity area.

In 2018, an increasing amount of AI ethics events, foundations, collaborations are organized and established. Inspiring words, visions and principles were released by companies as well as research institutes concerning AI ethics. Nevertheless, little practical support is published. Currently, similarly to ethics in general, the uptake of ethics in the AI practices is very limited. However, the AI development process and team need to be carefully assisted in creating a desired future, avoiding ethical pitfalls. This can lead to undesired societal implications. To ensure AI development is aligned to benefit

humanity, research and design must be supported by ethical methods and legal norms (Davis, 2012).

In line, interviewees within IBM mentioned:

“I think in specific on the topics of fairness in AI, AI responsibility, steps lack in the development process” - Interviewee IBM

» Thus, the argument put forth in this thesis is that if IBM’s employees, and their clients, are supported by practical tools to develop more ethical AI, the outcomes will be more ethical and socially desired.

Five Domains in AI Ethics

Rossi (2018), the IBM ethics global lead, identified five main areas of AI that call for ethical perspectives. **These are the following: (1) accountability, (2) data handling, (3) explainability, (4) value-alignment; (5) fairness** (figure 3.6). For the scope of this thesis is chosen to focus on fairness as overall goal of the project, with a specific focus on the value-alignment issues.

Both fairness and value-alignment are closely interwoven. If a system is not properly aligned it can lead to unfair outcomes. Furthermore, values and fairness are highly context dependent, therefore use-case differing. At the TU Delft faculty of Industrial Design Engineering, the problem is looked at from three diverse perspectives; the human, the business and the technology. Using the design lens sheds a new light on the ethical challenges arising with AI. Especially from the strategic design one, supporting ethical implementation in industry

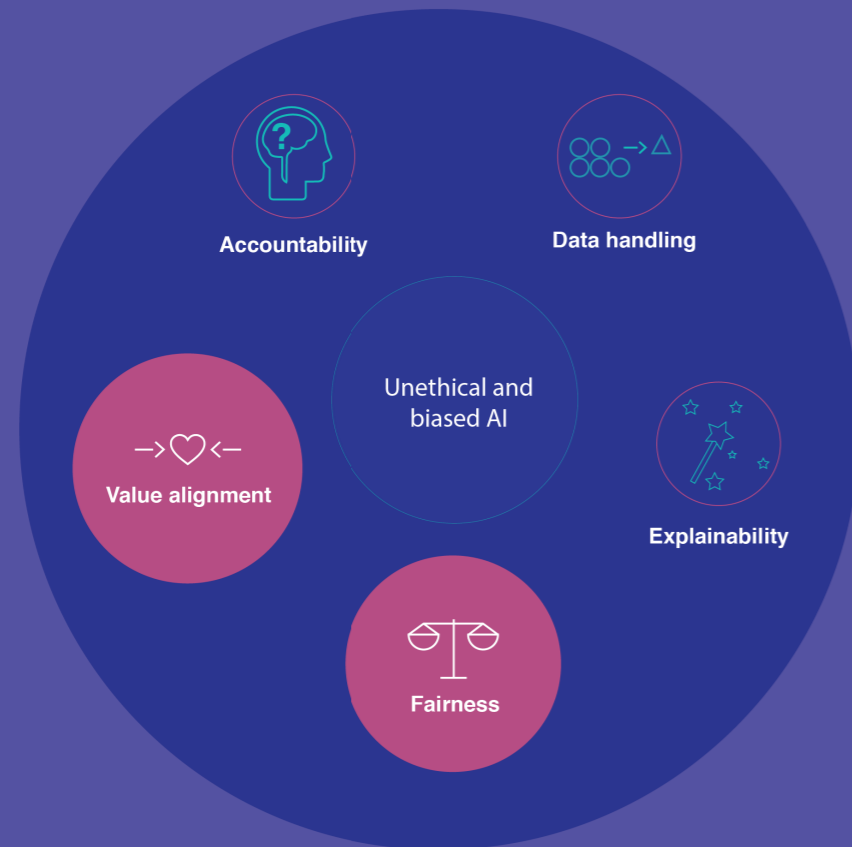


Figure 3.6 | The 5 main ethical challenges in AI development

Fairness and value-alignment both seem to lack the human perspective, which design can bring. Chapters 4 and 5 elaborate on the two topics in detail.

Value | “what a person or group of people consider important in life” (Friedman, 2012) or what a person or group of people embed (un) consciously into a system.”

Value alignment challenge | The lack of integration of “desired” values into (AI) systems development processes and their outcomes.

Fairness in AI | A fair algorithm is an algorithm whose outputs do not discriminate between different classes of people (Balayn, 2018) and is not perceived as unfair in the context of use.

However due to the complexity of defining fairness, this thesis takes a different approach. It identifies unfairness sources and aims to reduce these in the AI development in context specific fashion.

3.3

Design perspective on ethics

From the previous conclusion it is clear that there is space for improvement in ethical uptake as well as approaches. Critique on ethics is that it needs more solutions, fewer discussion and more synthetic reasoning (Dorst & Royakkers, 2006). Currently ethics is highly abstract, too slow, not integrated with current processes. In this thesis is argued that ethics in AI might benefit from a design perspective. Shortly the current relationship in literature with design and ethics is described and manners in which it fruitfully supplements.

3.3.1 What is design?

Design is an act aimed at changing and transforming the world (d’Anjou, 2011). Designers create products and services which have an influence on society and environment directly (Papanek, 1971). They create, the future due to their actions and decisions and therefore one can say it is prescriptive (Dorst & Royakkers, 2006). Often designers use imagination and creativity in their process to imagine solutions and new propositions. Mental stimulation strategies support designers to cope with uncertain design processes. ‘conceive the building in the imagination, not on paper but in my mind.’ (Lloyd, 2009).

3.3.2 Benefit of design in ethics

Multiple scholars argue that the ethical process can benefit from a design perspective for diverse reasons (Whitbeck, 1998; Dorst, & Royakkers, 2006; Harris et al., 2000; Lloyd, 2009; d’Anjou, 2011).

» In line ethical literature recognizes designers as important professionals as they cannot only provide technical means but also address values of people and society and create ways how to express them in material culture and technology (Van den Hoven, Vermaas, & Van de Poel, 2015). The ways both disciplines can supplement each other are extracted from the literature review and shortly described.

Complex problems

First, both design and ethics are concerned with complex problems. Both have different frameworks, tools and methods to cope with them. The different perspectives, tools are argued to benefit from each other. Both design and ethics people cannot decide the “best solution”. Rather they are approximations of “the better” solutions in the professional (Whitbeck, 1998; d’Anjou, 2011). For example to deal with complexity designers use modeling, a competence to creatively solve problems visually and making decisions (Simonse, & Badke-Schaub 2015). Prescriptive models support designers to choose by the consequences of multiple decisions (Simonse, & Badke-Schaub 2015). In line, both in design and ethics the challenges deal with questions concerning of what is valuable. Thus, designers often have to decide what is right and valuable (Johnson, 1993).

During the design process new information arises. Uncertainty is one of the characteristics of the problems design copes with. In design diverse solutions are acted upon simultaneously and incomplete solutions are at the mind of a designer while undertaking its actions (Dorst, & Royakkers, 2006).

» Therefore Lloyd (2009) calls designing a prototypical kind ethical thinking and ethics can benefit from this design thinking.

Imagination

Second, in design creativity fuels the solution space. Ethics demands creativity in this space. Researchers argue that ethics could benefit of the design creativity methods and tools especially in the area of creative imagination (d'Anjou, 2011). In line, Withbeck (1998) argues that the imperative facet that can be transferred from design to ethics is, that design next to the analysis of the challenges and choosing one solution, additionally puts effort to finding and imagining new solutions. This is referred to as synthetic reasoning. Correspondingly, Lloyd (2009) mentions the imagination helps to play with possible situations and outcomes of solutions, which would support the ethical decision-making process. Even though, in design no specific moral framework is used, it still often consciously explores diverse alternatives. Simultaneously, distinct values of the involved are integrated. The ethics field particularly focuses on analytical part resembled by the methods. However it lacks support in the solution space, while design does. **Thus, is argued in this thesis that stimulating imagination and creativity of the AI team in designerly fashion, might support more ethical AI solutions and opportunities.**

Conflicting demands

Third, design deals well with satisfying conflicting demands (Whitbeck, 1998; Dorst, & Royakkers, 2006). In the design process, designers need to make design choices i.e. to the usability, budget, sustainability. Design solutions and design problems are in constant tension, due to the diverse stakeholders with a variety involvement in decision making processes. Integrating these can be done through variety of (empathy) tools such as persona's or more intuitive approaches. In ethics different value tensions or value-trade-offs arise due to the impossibility to fulfill all needs in reality. **Yet, not much support exists for ethical practitioners dealing with ethical issues and these trade-offs. Thus is argued in this thesis, a design perspective might support ethics in the resolving of value tensions, perhaps using empathic tools.**

'We must cultivate moral imagination by sharpening our powers of discrimination, exercising our capacity for envisioning new possibilities, and imaginatively tracing out the implications of our metaphors, prototypes, and narratives.'

-Mark Johnson (1993)

Opportunity space

Fourth, the previously discussed ethical methods and tools suggest that the problem should be clearly beforehand. However, both Van de Poel et al. (2007) and Withbeck (1998) argue that people should unfold the problem during the process of solving, similarly as in design, leaving opportunity of finding the real problem. Thereby tapping with solutions people did not even realize they wanted (Lloyd, 2009). Furthermore, this thesis argues that design, in notions of critical design and research through design (RtD) can support ethics by means of novel inquiry. Designing artifacts can be used as provocation, speculation, exploring new design spaces, establishing critical areas of concern (Giaccardi & Nicenboim, 2018 p.68). It allows to explore the design solution space through user's engagement and interaction with the design and aims to diversify the manners of problem and idea interpretation (Malpass, 2017). Thus, design can supplement ethics in redefining moral problems and fuel the opportunity space.

3.3.3 Conclusion

» Thus, AI ethics might benefit from a design perspective in particular in (1) addressing complex problems while using design tools and methods e.g. modeling, visualizing (2) use imagination for more ethical solutions, (3) dealing with conflicting demands e.g. by empathic tools (4) widening solution and opportunity space. » This perspective lies at the foundation of this thesis. Using design methods and tools to incorporate ethics in the AI practice is intended. Stimulating, creativity imagination and reflection in an empathic fashion fuel the ideation space.

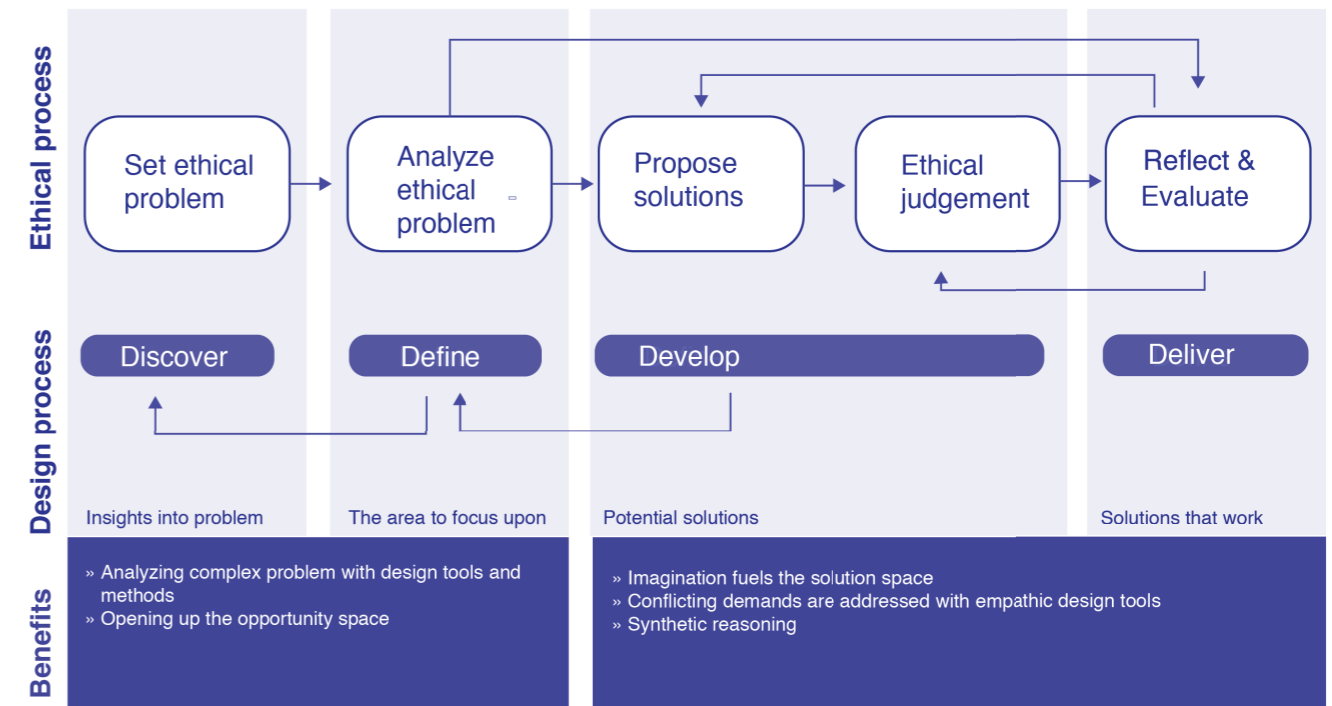


Figure 3.7 | A comparison made of the ethical process and the design process based on literature review of ethics and the double diamond method in this thesis (Delft Design Guide)

03 A taste of ethics

Value-laden technology

Technology is value-laden due to the values inscribed in them during the development process, (un) consciously.

When ethical support is needed

It is important to identify ethical decision moments to support AI teams in the development. The benefit of implementing ethics is the biggest early in the process

3 strategic ethical building blocks

Strategic building blocks for ethical organizational capacity are derived based on the analysis of the ethical tools and methods: ethical people (employees & client); ethical company and ethical tools/processes.

Increase of ethical uptake needed

The format of ethics currently does not fit in the move fast agile developments and there is a lack of ethical uptake in AI practice.

Focus fairness & value-alignment

The two focus areas of this thesis in AI and ethics are fairness and value alignment as these can benefit strongly from a design perspective.

Benefit of a designers perspective

Ethics and its uptake might benefit of more synthetic reasoning from design. Four main areas in which design can supplement are identified: dealing with complex problems; imagination; conflicting demands; opening up the opportunity space.

Chapter 04 I

Seasoning Fairness

Disclosing sources of unfairness

After the more general foundation of ethics this chapter explores more in depth the notion of fairness, based on an extensive literature review. Instead of defining what fairness is this thesis takes a novel approach into disclosing sources of unfairness in AI development. These are elaborated upon per AI development phase.

In this chapter

1. Fairness foundation
2. Unfairness sources in AI

4.1

Fairness foundation

Nowadays, research into machine learning and fairness is rising due to the insight algorithms can be (perceived) unfair. This topic is increasing in interest both from research and practice perspectives (Sylvester, & Raff, 2018). The first approaches to deal with this unfairness are released from technological perspectives (such as Fairness 360 toolkit). Little is done in the AI field, from the actual sources of unfairness, namely the humans. In this section guides alongside the main perspectives on and dimensions of fairness. Then shed a light the current identified sources of unfairness in AI development. With this in mind, one of the goals of this thesis is to reduce these sources of unfairness.

“One of the major problems with our blind trust in algorithms is that we can propagate discriminatory patterns without acknowledging any kind of intent.”

~ Cathy O’Neil (2016)

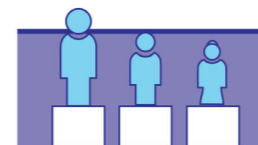
4.1.1 Introduction to fairness

Is it fair to give everyone equal probability in obtaining benefits, or should we aim to distribute based on the need? The question of what fair is, has been a topic of discussion by philosophers as well as in justice (Law & regulation) (Taylor, 2017). Still, there is no agreed upon definition of fairness (Gajane & Pechenizkiy, 2017). According to Saxana et al. (2018) it is even very unlikely only one definition of fairness is sufficient, with whom I agree. Nevertheless, this makes the development of fairer algorithms even more complex as measurements for algorithmic fairness also differ. This can be even problematic. An example of AI system tested with one of these “fairness measures”, appeared to discriminating against people with darker skin

colors (ProPublica, predicting the likelihood of recommitting a crime after being in prison).

» Therefore, instead of trying to find an agreed upon definition, I argue in this thesis to reduce the sources of possible unfairness in AI development is a promising novel approach. These are often related to anti-discrimination laws (Equal Employment Opportunities, 1964) and the avoidance of negative bias see figure 4.1. Briefly the three main perspectives of fairness are addressed related to AI. The three identified dimensions of fairness are elaborated upon Then, the sources of unfairness in AI are explained.

4.1.2 Three main perspectives



I Equality

One perspective on fairness is that everyone should get everything equally, both good and harmful (Leventhal, 1967). This perspective

Different types of bias

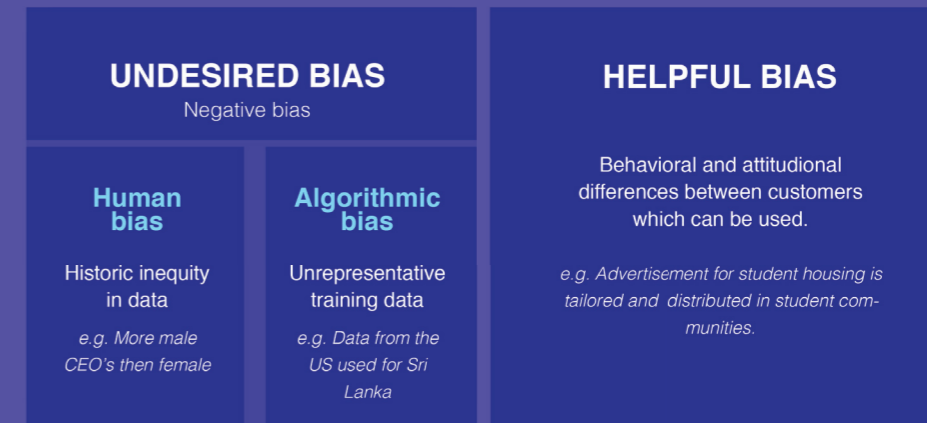


Figure 4.1 | Different types of bias based on Purcell (2018)

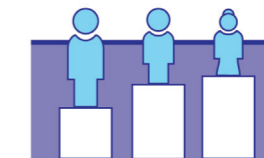
does not account for the need of a sportsman to eat more than a child. Fairness is seen as equality of outcome (Dobrin, 2012). Linking this to AI development is practically impossible as AI systems operate based on differences between data. Thus when it makes decisions or creates predictions concerning people, it bases decisions on differences in that data. When everything is really divided equally there might be no need for an AI system. Nevertheless, the AI team could look to protected attributes (such as: gender) and strive for equal distribution in the data set for these.



II Deservedness

In this notion of fairness, people get what they deserve. Fairness in this view seen as a more rational calculation (Dobrin, 2012). If Person A works hard, is ambitious, smart, fast, this person deserves more than Person B. If Person B is careless, and less hard working. Thus, fairness is seen as an individual choice, an individual freedom. Linking this to an AI system it immediately raises the question based on who's notion of deservedness the system would work. It could result in a pre-

selection of attributes the system will train on, such as university grades in a hiring process.



III Need

Fairness can be seen as social justice. In this fairness view, people with much money should pay more taxes than people with little money (In order to help each other out, for the common good.) It is connected to responsibility and empathy plays a big role in this notion of fairness as the person who is able to contribute more helps others out who have less (Dobrin, 2012). Translating this perspective to AI systems the AI team first will need to determine on which parameters a need of something will be measured to translate it into a deciding factor.

4.1.3 Dimensions of Fairness



I Fair Process & Fair Outcome

Fairness can concern actions, processes, outcomes. Thus, next to the different perspectives on what is fair, also two other different dimensions of fair assessment can be

distinguished, namely fair results (also called substantive fairness) and fair procedures (sometimes also called procedural fairness) (Ryan, 2006; Carr, 2017). People can think something is unfair due to the process/procedure. Contrastingly, people can judge fairness based on the outcome. The two different perspectives are illustrated with an example from the Netherlands which bares much discussion in appendix F.

» **In line, researchers argue that operation within accepted parameters, does not guarantee ethical behavior (Mittelstadt et al. 2016). Thus, both outcome and process need to be analyzed to fully be able to determine if an AI system is fair (as far as is possible).**



II Perceived fairness

When a decision is made, a response of one and the same unfavorable outcome can be perceived differently by people. In fact, when a decision is made by a group of decision makers is experienced less fair than when an individual makes the decision (Kouchaki, Smith and Netchaeva, 2015). Thus, not only based on which theory the decision is made but also by an individual or group is perceived differently. This gives an extra dimension to fairness. In line, a recent study concerning AI found out that mathematically-proven fair algorithms might be not perceived as fair due to a mismatch of social concepts of fairness (Lee et al. 2017).

» **For AI systems it means next to the fair process and fair outcome, also context-specific fairness needs to be taken into account, to be perceived as fair.**

4.2



Sources of unfairness in AI

To actually understand how to alter an AI system or process to make it fairer, people firstly need to deeply understand the sources of unfairness in AI systems. Different aspects can lead to unfair process or outcomes. These unfairness sources are the result of an extensive literature analyses and analyses of ethically misaligned use cases. The following ten sources of unfairness are described in relation to the AI development process. (Figure 4.2)

4.2.1 Sources of unfairness per process stage

I Data reparation & understanding phase

Data scientist Judith Red said a quote about algorithms and their output: "garbage in garbage out".

In other words, our society is full of demographic disparities which naturally is reflected in the data we generate. Algorithms process historical data and therefore have limitations, as the output cannot exceed the input (Mittelstadt et al., 2016). Choices made about the data set and on which parameters models work, influence the fairness of AI systems. The way the data is collected, from whom and where all influences the data set, the model and, therefore, the outcomes. When designing a system for a specific target group or, in general, global society, it is important to make sure that the data represents the context of use and the people using, it to avoid undesired outcomes

01 Algorithmic bias / incomplete data

Cases in which too little data is present to train the machine learning algorithm or with misaligned data-sets for the use context can lead to algorithmic bias (Purcell, 2018). Term reflects incomplete/wrongly sampled training

data that was used to create undesired biased model. A well-known example in our current society, is from Google. In their image labeling technology, it labeled darker skinned people as gorillas (Vincent, 2018). In this case the training data set did not contain enough pictures of people with darker skin colors, which resulted in a racist labeling by their application. Unfortunately, it is also the case that proportionately there is always less data from minorities, which leads to worse predictions for these minorities (Hardt, 2014). Figure 4.3 shows the origin of image training data.

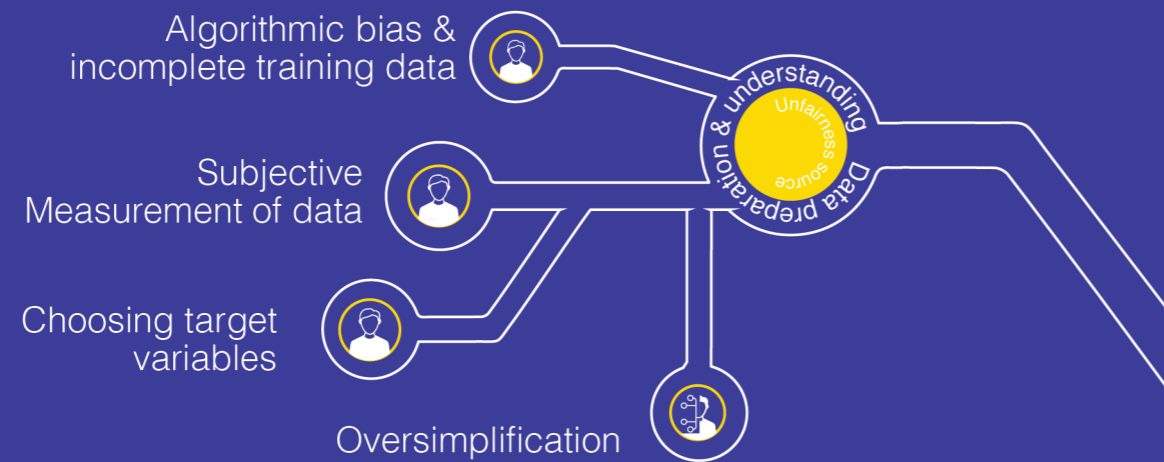
02 & 03 Choosing target variables & Subjective Measurement of data

Measurement seems to be an objective process. Nevertheless, also in this phase decisions can lead to undesired biases in AI systems (Barocas et al., 2018).

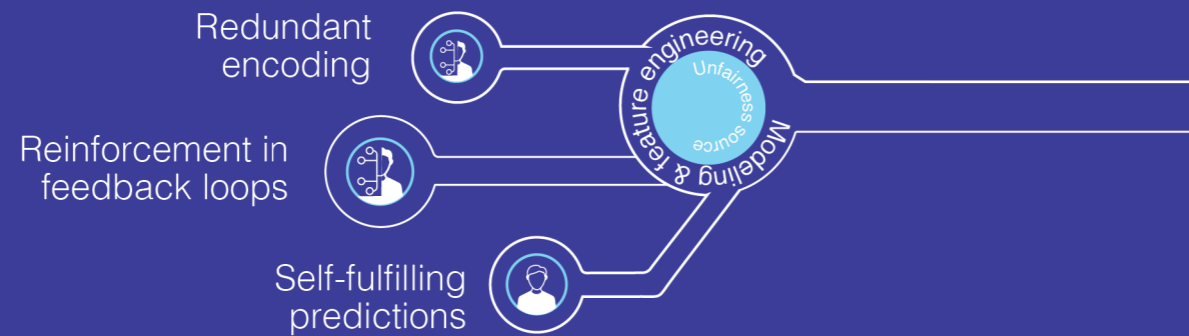
To illustrate, a team working on a machine learning system supporting the HR department will need to define categories based on which a person gets to an interview. Defining these categories, for example based on target variables is a subjective act. The AI team will decide based on which characteristics one gets the job interview or not while often not being an expert in the field for which the decision is made. Bias in the target variable is

Unraveling sources of unfairness in AI

Data preparation & Understanding



Modeling & feature engineering



Evaluation & deployment

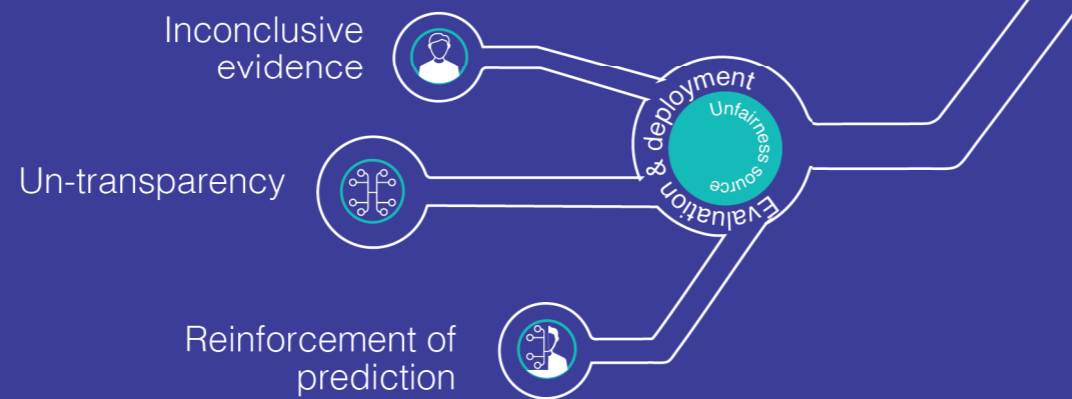


Figure 4.2 | Conceptual framework of unfairness sources in AI, a result of this thesis

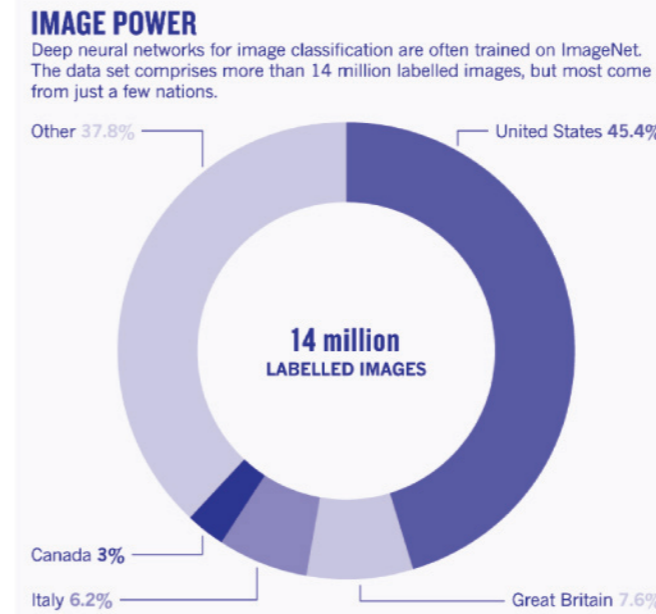


Figure 4.3 Zou, J., & Schiebinger, L. (2018). Shows how AI systems are trained based on data just of a few countries

of crucial importance as it directly influences the outcome/prediction. Also what to measure, such as a “good employee” is extremely difficult. To conclude, choosing what to measure and how to measure is a subjective act and can be a source of unfairness in AI systems. This is often not addressed by AI teams (Barocas et al., 2018).

04 Oversimplification

The world is complex to measure due to the many contextual variables. While measuring decisions are made that often lead to Oversimplification of the actual context. This can happen due to data cleaning when dealing with messy data or an absence of enough context data. In general, this points out a limitation of data-driven techniques, it is an oversimplification of the world (Kamphuis, 2018; Barocas, et al . 2018).

II Modeling phase

05 Redundant encoding

A manner how AI teams aim to remove historical bias in training data is to remove the parameter for example for gender (or

any other one which might lead to unethical classifications). However, still other data might point out the gender, even though it is not specifically used as input (Corbett-Davies et al., 2017). For example, in the Netherlands 85% of the single parents is female (CBS, 2008). When including this information, the model be trained using gender indirectly (statistically seen this is the same). This is called redundant encoding (Purcell, 2018).

06 Reinforcement in feedback loops

Machine learning systems often work with feedback loops when making predictions. To illustrate, Google normally records the amount of user clicks and time spent on websites to determine the relevance of the results. This feedback can be problematic to interpret correctly (Barocas et al., 2018). Does the number of seconds spend on a page show the relevance or can it also have other reasons? This can lead to unfairly trained systems. A manner to avoid this could be the use of crowdsourcing labeled data (Bashirieh et al. 2017).

07 Self-fulfilling predictions

With predictive systems, the real world actions are influenced with the prediction outcome of the algorithm. An example to illustrate these turning into self-fulfilling predictions is described. Some police departments might use predictive systems to determine areas in cities with a higher crime. This leads to more police officers being sent to the area with high crime prediction, which might lead to more people being arrested in that area. Thus, the feedback will strengthen. The prediction seems validated and performing will even though it might be fully based on negative biases, which is undesired (Barocas et al., 2018). The prediction affects the training data set itself. A manner to account for this is by “tweaking the model by quantifying how surprising an observation of crime is given the predictions.” After which the model should

be only updated in case of surprising events (Engsign et al., 2017).

III Evaluation and deployment phase

Decisions are increasingly left to AI systems, which directly influences humans and society. One of the most important research directions might be to rethink ethical implications of AI systems actions (Dignum, 2017), the way they actually are evaluated and deployed in the market.

08 Inconclusive evidence

Outcomes of algorithms are probabilistic when they process data with for example machine learning techniques. In other words their outcomes is still uncertain knowledge (Mittelstadt et al., 2016). Even though significant correlations can be found by performing these techniques on data sets, this relation is rarely sufficient to prove causality, as well as very complex (Illari & Russo, 2014). Actually acting on these correlations can lead to problems. This leads to actions which are taken based on inductive correlations (Mittelstadt et al., 2016; Barocas et al., 2018). Thus, it is of crucial importance to identify these limitations as well as how and to who are the risks when this relationship appears to be incorrect.

09 Untransparency

To address if a process or decision is unfair or fair, people need to see how the decision is made (Dignum, 2018; Barocas et al., 2018). This is currently not possible with specific AI systems, referred to as black box models (using deep neural networks, which lead to unexplainable outcomes). This can be unfair especially when concerning life decisions. Currently more and more research into explainability of AI systems is performed. Questions arise such as: what are good decisions and based on what can we make these decisions?

10 Reinforcement of prediction

In predictive systems the outcome of the prediction in itself can sometimes reinforce

the result. Thus, in itself prediction affects the outcome, similarly as with the feedback loops (Barocas et al., 2018).

4.2.2 Lack in strategies for fairer AI

Four current approaches towards fairer AI (most of them released in 2018) are discovered through an analyses of the current propositions and tools: Control, Code Fix, Reminders & Checklists and Awareness & Dialogue (described in detail in appendix F). Although they definitely support in a good direction, in particular a critical perspective is shared.

After this literature review, fairly little work is found on actual implementation on a day to day basis of fairness in AI development.

Current tools work as afterthought, rather than at the beginning of the process. While in ethics literature advocates for prevention early in the process.

In line, there is an inclination in diverse disciplines to solve their challenges within its discipline. Just because something has (partly) a technical cause, does not necessarily need a technical solution (Boddington, 2017). In AI it is called **Artificialintelligencication**. Current toolkits often tackle the problem from a technology perspective and do not take context specific fairness and many of the identified unfairness sources into account. The found solutions do not account for context specific fairness or values, which is of crucial importance.

4.2.3 Conclusion

The recent releases and studies prove the increasing relevance and need for practical support for the creation of fairer AI applications. Albeit, due to this literature study also a critical light can be shed on the contemporary field of AI fairness.

Instead of defining fairness, this thesis is guided by a novel approach. Ten sources of unfairness in AI are unraveled in relation to the development process.

» Concluding from the fairness literature review is argued that **the cause of unfair AI is (mostly) a result of human actions** (see figure 4.2). Therefore, this challenge is tackled by looking at the human unfairness source. To complement existing approaches this is performed from a design perspective.

» In this thesis is argued that **fairness needs to be treated as a central concern during the development and training of AI, rather than an afterthought which it often it still is** (Barocas, Hardt, & Narayanan, 2018). In other words, it should be addressed with preventive matters. A proactive stance in developing fair systems is taken.

Futhermore, there is a gap between the people building the algorithms and the people who use them or are affected by its decisions, as Cathy O'Neill describes in her quote.

» In this thesis is looked at **manners closing the gap by empathizing with the AI team**.

“ We have a total disconnect between the people building the algorithms and the people who are actually affected by them” – Cathy O'Neill (2016)

And how Narayanan (2018) re-frames the problem:

“ It is not about mathematical correctness, it is about how to make algorithmic support human values“

» Finally, most support for fairer AI does not take into account human, context or industry specific fairness. In this thesis is looked at ways to foster the diverseness and manners to account for context specificness, in a practical form.

04 Seasoning Fairness

AI can be unfair

Companies and humans learned by mistakes, in practice, with ethically misaligned AI systems that AI systems can be unfair

Increasing need for support

The recent releases and studies prove the increasing relevance and a need for practical support in the AI field for fairer AI systems.

No agreed upon definition

Three main fairness perspectives are based on the following: (1) Equality (2) Deservedness, (3) Need.

3 fairness dimensions

Three different dimensions of fair assessment can be distinguished, namely fair results, fair procedures and perceived fairness.

Artificialintelligencication

Just because something has (partly) a technical cause, does not necessarily need a technical solution, solutions from non technical perspectives are lacking in support for fair AI systems

Unfairness sources

Instead of trying to find an agreed upon definition, ten sources of unfairness in AI are identified in this thesis (p.54). This allows to design support to reduce sources of unfairness in context specific fashion.

Humans are source of the unfairness

The identified sources of unfairness in AI are mostly from human origin.

Lack of human support for fair AI

Little support for fairer AI systems exists and no support is found from a non technical perspective.

Chapter 05 I

Taste Differences

Demystifying value tension

The previous chapter explored fairness in AI. This chapter firstly explores the meaning of value and the value-alignment challenge. After which it particularly focuses on value-tensions in AI development and on strategies how to resolve them.

In this chapter
5.1 Value-alignment
5.2 Value tension

“It is not about mathematical correctness, it is about how to make algorithmic support for human values “

- Narayanan (2018)

5.1 → ←

Value-alignment

The following section aims to deepen the understanding regarding value-alignment in AI. It elaborates why it requires attention, why it is so challenging to solve and shed a light at the current attempts to minimize this challenge through a variety of approaches. The AI field is one multidisciplinary one. Therefore, literature has been consulted with the research questions in mind from a diversity of disciplines ranging from (social) psychology, philosophy of technology, human computer interaction, design and software engineering. Additionally, practice has been consulted to gather insights both from a data science perspective and philosophy of technology one.

5.1.1 Theoretical background value

Before diving into value-alignment, firstly the definition of value are elaborated upon. Similarly to ethics in research there is no agreed upon definition of the word value. The addressed fields have distinctive perspectives. From a physical perspective, value lies at the heart of ethics (Zimmerman, 2015). Philosophers distinguish between intrinsic and extrinsic values (Zimmerman, 2015). Values which are ends such as happiness/wisdom are intrinsic, desired from themselves. Extrinsic values are the means to an end, such as privacy. The value privacy is extrinsic in this view as it contributes to intrinsic ones such as self-worth or happiness. Research from more technological origin or social technical systems refers to values as entities appearing in technologies (Friedman and Nissenbaum 1997; Johnson 2000). In detail this is described in appendix H, philosophy of technology. In social studies, such as social psychology or anthropology, values are referred to as criteria that humans evaluate their behaviors, judgments etc. (Bennett 2003; Shilton, Koepfler, & Fleischmann, 2013). An example Kenneth Fleischmann (2013) calls values as “bridges between the individual and the social. Individuals hold values, but others influence the formation of those values” (Fleischmann, 2013, p. 2). In line, value can be seen as the propulsion of

motivations for humans. From this perspective values can be seen as a choice (Kluckhohn, 1951). Design literature is addressing values in a slightly different fashion. It explores how values are exposed, negotiated and translated into product features which consequentially leads to adoption and social impact of the design (Le Dantec, Poole, & Wyche, 2009; Shilton, Koepfler, & Fleischmann, 2013). At the intersection of humans and technology values are semantically confound.

All these perspectives have one aspect in common, values infuse our interactions, will it be socially or with products. For a clear understanding, the different value sources and manners to describe values are categorized and visualized in figure 5.1. These axis support research into values by understanding the sources and attributes. This is used for the further empirical study in this thesis.

I Describing value sources

Agency refers to the degree of autonomy and self-determination in expressing and possessing values. Subjects can express their own values while objects have values inscribed to them. Humans are mostly subjects. Unit refers to the scale on which the value is researched or expressed, at a scale from individual to collective. This is very relevant as values on different

Value Source

Can be from an intrinsic or extrinsic value

Describing value attributes

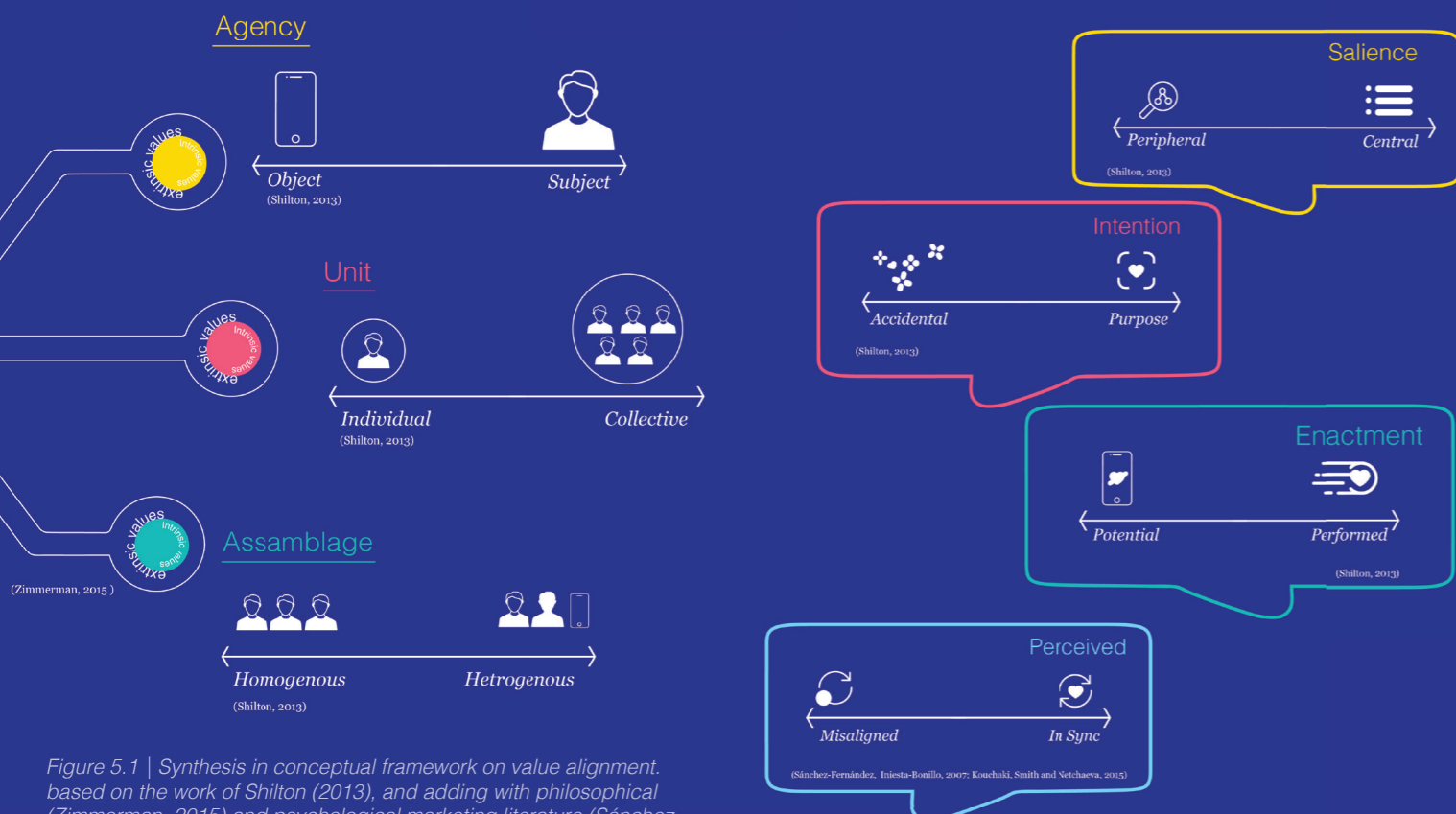


Figure 5.1 | Synthesis in conceptual framework on value alignment. based on the work of Shilton (2013), and adding with philosophical (Zimmerman, 2015) and psychological marketing literature (Sánchez-Fernández, Iniesta-Bonillo, 2007; Kouchaki, Smith and Netchaeva, 2015).

levels for example school values and one pupils' values can clash, and to understand this an understanding of the different levels is needed. As well a new AI system might support certain collective values, however might be in contrast with some individual ones. **Assemblage** refers to the convergence between the diverse actors. It is important to know if the values are researched from a homogeneous group or a heterogeneous group.

II Describing value attributes



Salience refers to the dimension of central values or more peripheral values. Some values might be more important than others which can be discussed from a central value perspective or one which accounts for the context of the

value in use. Discussion is ruling which values to implement in AI systems, core values, situational/peripheral values or a combination of the former. Research identified core values for ICT designers. These consist of human welfare, ownership and property, privacy, freedom from bias, universal usability, trust, autonomy, informed consent, accountability, identity, calmness, and environmental sustainability (Friedman, Khan, Borning, 2006). Others argue that the universal declaration of human right should be the north star for AI development (Schmid, 2018). Hence, these opinions are fired with critique leading to discussion, especially if values need to be singled out (Borning & Muller 2012; Van den Hoven, Vermaas, & Van de Poel, 2015). On one hand it gives a basis for AI development process. On the other hand, values are culture-specific, per person etc. Therefore, singling out values might conflict with situational ones (Borning &

Muller 2012). **This salience of certain values is empirically researched, which is further discussed in this thesis.**



Intention refers to the dimension of values that are embedded by accident and unconsciously to values that are embedded purposefully. These unconscious values might lead to biases in AI systems discussed in the previous chapter. **In the case of this thesis it is thus preferred to address values purposefully.**



The enactment dimension entails the degree to which a socio-technical system actually shows this value. The certain design choices of a system might not directly influence but indirectly bring or alter values into this world. **Potential values** are seen as present but not active in the design/work. **Performed values** are values that are materialized in this world by the system/design, which in the case of AI systems might be preferred.



Perceived refers to the dimension to which extend the intended values are actually perceived in the similar manner.

5.1.2 Value view in this thesis

The previous paragraph leads to the use of the following definition of value in this thesis. Integrating the different perspectives of the different disciplines, due to the multi-disciplinary of the AI field.

Value I “what a person or group of people consider important in life” (Friedman, 2012) or what a person or group of people embed (un) consciously into a system.“

As mentioned earlier, values have different levels

(Flanagan et al. 2005; Mason & Loukides, 2018). Values differ for example per individual, team, company, industry, country, culture. They are depended on socio-cultural context (Dignum, 2018). These different so-called levels of values can lead to value trade-offs or value-conflicts. Even though the analysis level of is different per value level. Thus, this framework supports understanding of the diverse levels of value analysis that fuel the empirical research of this thesis fruitfully.

However little work is done to systematically account for different value levels in the process (Flanagan et al. 2005). Thus, this might contribute to the value research field.

Based on the previous sections, in this thesis is argued that a combination of both “core” values for AI development teams in IBM can be researched with situational values per industry, stakeholders, individual, team members and project might be a promising approach.

In this manner both the main values IBM stands for can be supported for in their projects, as well as account for situational and context specific needs that might be differing per case.

5.1.3 Value-alignment

For value-alignment as goal the following definition used in this thesis:

Value-alignment goal I “AI systems should be designed so that their goals and behaviors can be assured to align with human values throughout their operation.” - (The future of life institute, 2017) “the pragmatic goal of designing systems so that they embody values to which designers, users, other stakeholders, and the surrounding society are committed.” (Flanagan et al. 2005)

Value-alignment problem I The lack of integration of “desired” values into (AI) systems development processes and their outcomes.

The example in figure 5.2 illustrates that an extreme version of the value alignment problem, which shows us that even slight, and on first hand unharmful and unconscious misalignment can turn into a serious problem both for humanity as well as companies (Yudkowsky 2008; Bostrom 2012; Soares, 2015). Three main areas are identified for the need of value alignment discussed in appendix S. These are goal orthogonality, instrumental convergence and unconscious unwanted value embedding. » **Founded in this literature review the notion is taken that to make the AI team more conscious about their own values and open for discussions on these values, supports the development of more value aligned AI applications.**

1 Value-alignment in the AI process

The value alignment process can be divided into three consequential stages, identified by IEEE and visualized in figure 5.3. In detail discussed in appendix S. The first step is identifying whose/ which values will be implemented, for example the values of the end user, the company values, the clients values etc. For example, one could say the client's values need to be integrated, however these can be contradictory with the actual end users. Van den Hoven et al. (2015) argues that the explicit use of values is crucial for innovation processes. Additionally, it is important to decide how this decision (of who's values) is made. Value conflicts and updating for changing future values are part of this phase. Second, actually implementing these values into the AI system, which is the translation part from human values to machine learning or code. This demands tools and methods for translation and implements as well as for the explainability of design choices as well as the system. Third is evaluating the implementation of these values in the system assessing the results and adjusting when needed.

5.1.4 Value-tension

» **For the scope of the thesis, the decision is made to specifically look at the value tensions**

Example value alignment problem

An illustrative manner to explain this problem is by the example of the paper clip maximizer, introduced by Nick Bostrom:

"Suppose we have an AI whose only goal is to make as many paper clips as possible. The AI will realize quickly that it would be much better if there were no humans because humans might decide to switch it off. Because if humans do so, there would be fewer paper clips. Also, human bodies contain a lot of atoms that could be made into paper clips. The future that the AI would be trying to gear towards would be one in which there were a lot of paper clips but no humans" ~ (Bostrom, 2003).

Figure 5.2 | Extreme example of misaligned values in AI

for the following four reasons:

(1) the first stage needs to be resolved before the further ones can be executed in a more ethical fashion; (2) design deals well with satisfying conflicting demands (Whitbeck, 1998; Dorst, & Royakkers, 2006). Designers need to constantly make decisions concerning value-trade-offs in their work. Thus, this challenge might benefit from a designer perspective; (3) this is an untouched upon area, and almost no research from a non technical perspective is found in this matter. Thus this is a novel area to research; (4) for the scope of the thesis and the depth of the research the choice has been made to focus on one of the questions.

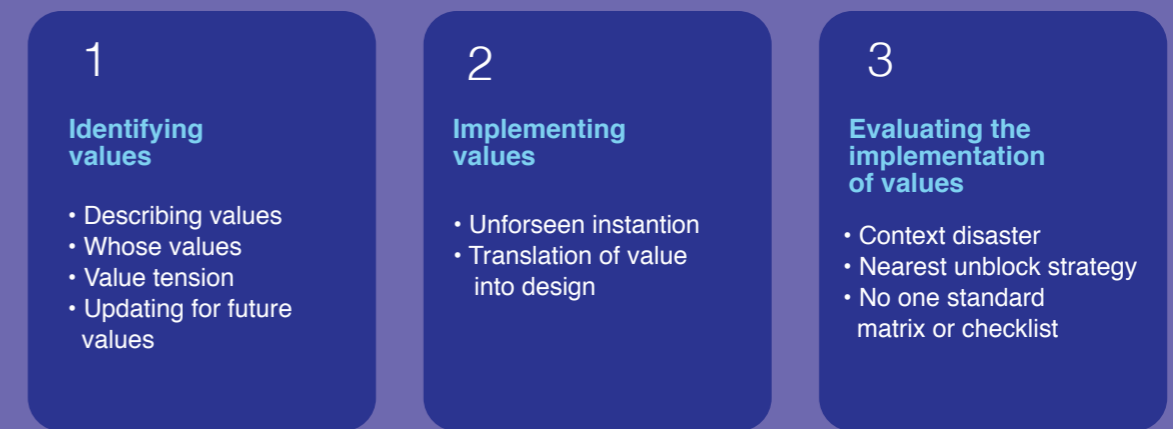
5.1.5 Conclusion

This section provides a synthesis of values literature in a form of a conceptual framework (figure 5.1). The sources of value and ways values are described are analyzed to find agreed upon dimensions. The preference of the value dimensions in AI development are based on the earlier ethics foundation (figure 5.4 & 5.5).

Salience | Peripheral & central: In this thesis the focus lays on resolving value tension in context specific fashion. However also certain central values concerning AI, such as fairness are taken a central aim.

Challenges in value-alignment

Figure 5.3 Value alignment challenges in AI per process based on IEEE research



Intention | Purposeful: Values should be consciously embedded in systems to prevent ethically misaligned AI systems

Enactment | Performed: Deliberately designing with ones own values to create more ethical AI systems.

Perceived | In sync : Perceived values should be in line with the intended ones to lead to ethically aligned outcomes.

Furthermore, the challenges of value alignment are shared. One main challenge is chosen for this

thesis, resolving value tension. The next section elaborates in depth on the literature research performed into value tension and strategies to resolve them.

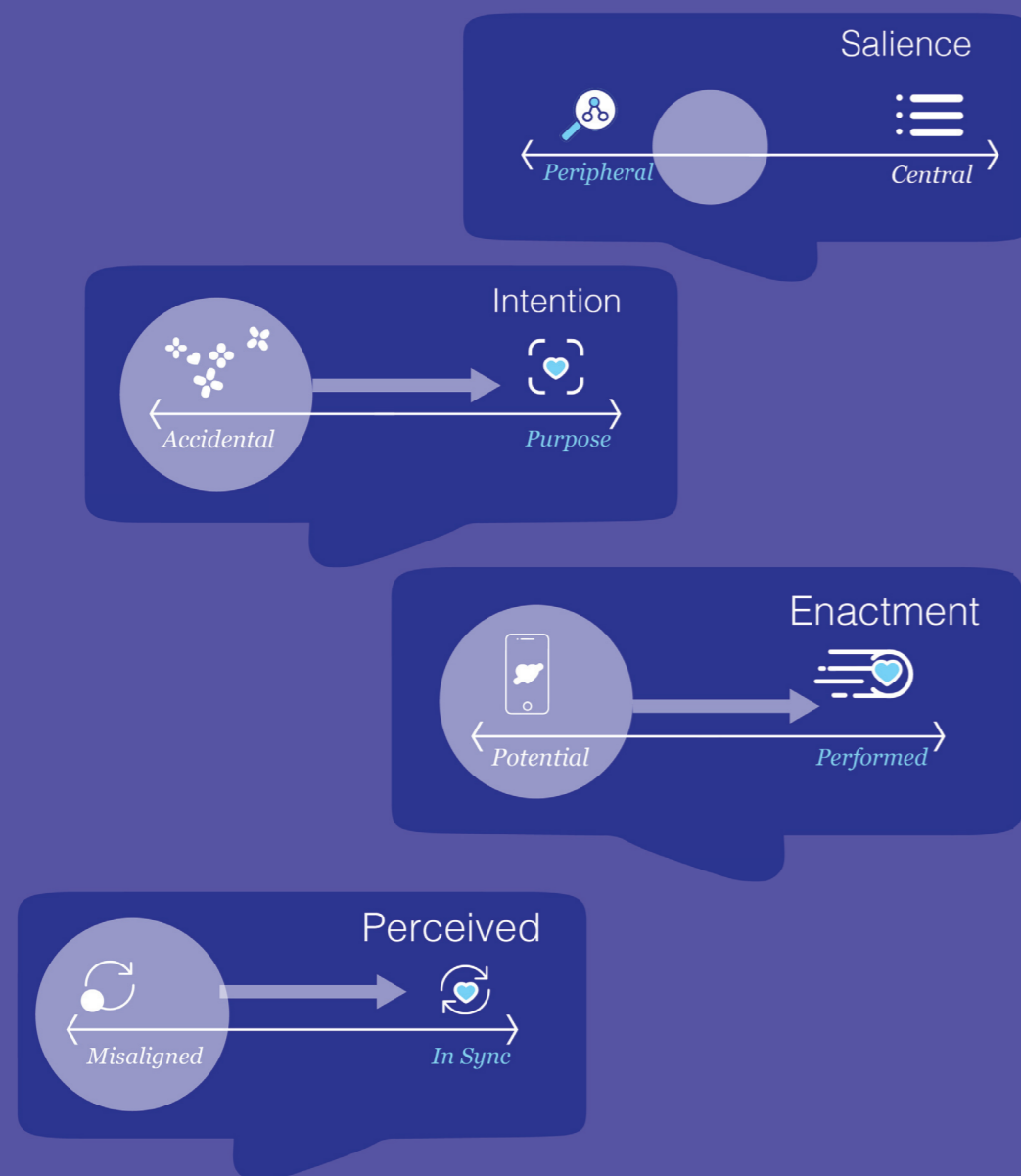
Scope of values

Figure 5.4 Scope of values in this thesis based on unit. The scope of values is chosen within the EU



Desired approach of values in AI development

Figure 5.5 | Desired dimensions of values in this thesis based on the literature study of values, philosophy of technology and ethics.



Value-alignment

Value alignment

“AI systems should be designed so that their goals and behaviors can be assured to align with human values throughout their operation.”
- (The future of life institute, 2017)

Value sources and describing these

In this thesis a synthesis is made in figure 5.1 of value sources in three dimensions (subject/agent, individual/collective, homogeneous/heterogeneous) and the four manners how values are described (peripheral/central, accidental/purposeful, potential/performed, misaligned/in sink).

Situational values

Values are highly context specific differing per i.e. person, context, industry. Simultaneously, certain values, such as fairness are desired of AI systems. This thesis the perspective of these context specific values with a central one: Fairness but resolving it in context specific fashion.

Explicit values

The explicit use of values is crucial for technology innovation processes. This this is aspired to do in the AI development process.

Challenges of value-alignment

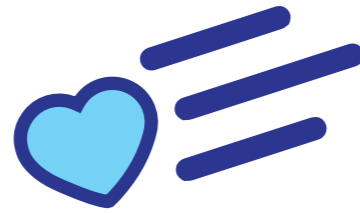
The challenges of value alignment in AI development per AI process stage are: identifying values, implementing values and evaluating values

Value tension is the focus

The decision is made to specifically look at the value tensions due to the following reasons.

First, the first stage needs to be resolved before the further ones can be executed in a more ethical fashion. Second, design deals well with satisfying conflicting demands. Therefore, this challenge might benefit from a designer perspective. Third, this is an untouched upon area, and almost no research from a non technical perspective is found in this matter.

5.2 Value-tension



Contemporary value-tensions in AI development are an under-researched phenomenon, while explicitly addressing value tensions has beneficial results. This section shares the insights gathered concerning value-tension in AI in literature. It intends to shed a critical eye on the field and explore manners to resolve these tensions.

5.2.1 The value-tension challenge

There is a need to take (social) values into account in AI system creation, with the priorities of values by the different stakeholders in diverse multicultural context while still explaining reasoning and guarantee transparency (Dignum, 2018). It is not possible to address all these values in a desired fashion. Thus, naturally value conflicts (value tensions) occur.

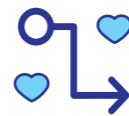
» **Not addressing value tension in an explicit way can lead to a lack of appropriation by disadvantaged groups, unfair AI systems or even more drastic consequences such as system sabotage (Flanagan et al. 2005).**

Yet, how to decide which values to integrate when there are conflicting values? Should moral values (e.g., a right to privacy) be of greater importance than non-moral ones (e.g., aesthetic)? Usually, when contexts are described clearly and in detail, no single value and its following action meets all obligations and desires. These situations are often referred to as moral dilemmas/overload (Van den Hoven, 2012). So how should we address the value trade-offs in design and its implementation?

5.2.2 Specific value tensions in AI development

In literature five main value tensions are discovered. In detail these are elaborated upon in appendix G.

- ≡ Accuracy vs Fairness
- ≡ Explainability vs Performance
- ≡ Bias vs variance
- ≡ Precision vs Recall
- ≡ (Historical) data value vs Socially desired value



4.2.3 Strategies for resolving value tension

A light is shed on value-alignment and value-trade-off tools and approaches with a focus on resolving value tension. These differ from policy making to ethical or engineering tools. For a detailed tool analyses see appendix J. Six strategies are extracted to resolving value tensions, which are described in the following sections (figure 5.6). One strategy does not exclude the other and often they are applied simultaneously.

The current tools, methods and approaches for resolving value tension have little to no evaluation beyond academic setting (Miller et al. 2007; Shilton, 2018) or bear still much critique. Additionally, the field is still at the very beginning of systematically thinking about design and

Example of a value tension

An example of a value tension in a system is: An open calendaring system which supports group activities, awareness and presence over one's individual privacy.

values (Flanagan et al. 2005). In line, few practical methods address value tensions among diverse values (Miller et al. 2007).



I Untangle value (tensions)

Making values explicit in the development process is repeatedly mentioned crucial for resolving value tension. The design for values in ICT tool mentions explicit thinking concerning values build into systems as significant for this. (Van den Hoven, Vermaas, & Van de Poel, 2015, p 838). This method also takes into account a more proactive stance in designing for specific values (Shilton, 2018). Also, the values at play tool includes value discovery for relevant values before resolving the tension (Flanagan et al. 2005). From this method, it is remarkable that it explicitly uses the designer's values, as often that is lacking. Also, Value dams and flows (Miller et al. 2007) method explicitly mentions to make the value conflicts explicit between the different stakeholders. This strategy focuses on the awareness of the team of these value conflicts.



II Decompose values

Multiple scholars mention decomposing values as a strategy towards resolving the conflicts between them. Due to decomposition of values towards principles, functionalities, features or

requirements, choices between these are made for the specific use case (value dams & flows, design for values IC, values at play). Design for Values ICT, translates of these values into a more formal language. Therefore, three levels of abstraction of values are made to support the translation:

- The abstract level (highly abstract statutes of a system, not yet contextual)
- The concrete level (specific model components in terms of concrete functionality)
- The implementation level (system components as the basis for implementation) (Van den Hoven, Vermaas, & Van de Poel, 2015, p 838).

Value dams and flows (Miller et al. 2007) takes the decomposition approach as a start for a well-founded and understood value-discussion. This strategy gives the discussion more context and is therefore easier to grasp. It leads to more case specific decisions which are often easier to relate.



III Avoid problematic features for stakeholders

This strategy contains a strong ethical angle by recognizing the minorities desires and potential harms. It is extracted from the value dams and flows method. In this method problematically, experienced features even by one of the stakeholders, are avoided. This led to designing for the desired stakeholders' values, while

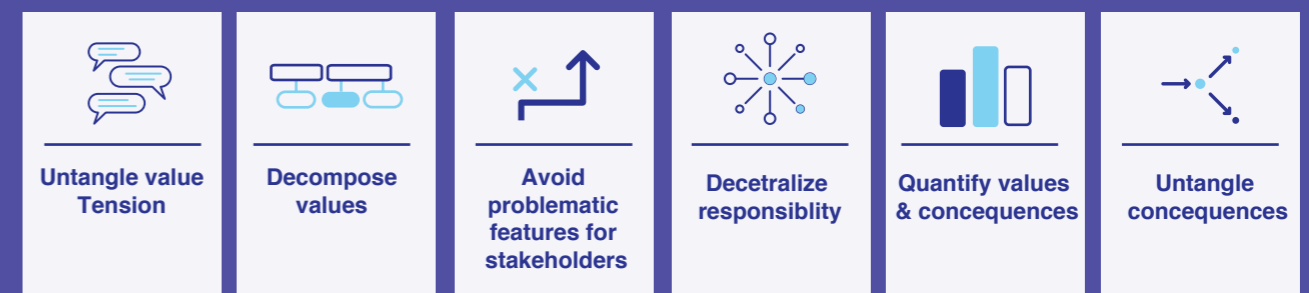


Figure 5.6 Strategies to resolve value conflict/tension a framing result based on the literature review

continuously addressing value oriented design trade-offs systematically (Miller et al. 2007).



IV Decentralize responsibility

This strategy extracted from policy making, is decentralizing responsibly for different values (Thacher, 2004). This is executed by establishing and sustaining multiple teams or institutions which are committed to different values, decentralizing responsibility per value. This ensures that each value has a vigorous advocate.



V Quantify values & consequences

Currently in AI creation as well as in design discipline and many others, value trade-offs are addressed by quantifying the values or consequences (cost-benefit, direct trade-off, Van den Hoven et al., 2015). Example of quantified fairness is often used in algorithmic fairness, or fairness metrics based on the difference between false positive (similarly to A pregnancy test is positive, when in fact you aren't pregnant) for different categories of population as an indication of possible fairness of algorithms. An example from the design discipline is weighted objectives (delft design guide). In this tool, design concepts are compared based on weighed values "scores". This is in contrast with the line of thought of this thesis. From the perspective of value used in this thesis quantification is not the approach to go for two reasons. **First, it does not account for the diverse contexts the system can be placed in (also not perceived fairness). Second, it limits the solution space using fairness as a boundary instead of a new inspiration source for AI systems.**



VI Untangle consequences

Several tools aim to put technological development in a wider socio-technical context, addressing it with a longer-term vision and making the consequences explicit. For example, the envisioning cards, a versatile toolkit, aims to discuss human values in this long term context. They take into account "envisioning criteria such as "stakeholders, time (the time span), values (impact technologies on human values) and pervasiveness (at the new interactions that the rise of the new technology evokes). It is supported to consider implications the idea has on people. It uses design activities such as sketching and asking questions. It is meant to support "diversity, complexity and subtlety of human affairs, as well as the interconnections among people and technologies" (Friedman and Hendry 2012). These type of tools catalyze designers both humanistic as well as technical imaginations as well it stimulates as a form of ethical reflection.

Design empathy tools and design scenarios are also stimulating imagination into possible consequences of certain value-tensions (Despotou, 2005). Schon (1988) says only thinking is not enough to envision, but also require a form of seeing and doing to create new world and envision consequences. **Thus, I position design tools/principles in this strategy, such as scenarios and using the creative imagination towards predicting consequences of certain ideas. It is referred to as an iterative process to untangle consequences.** This is very in line with the previous lines of thought concerning values and fairness perspectives discussed earlier in this thesis.

4.2.4 Common mistakes

Briefly some common mistakes in resolving value tension are analyzed and briefly described with prevention.

1 I Misunderstand value tension

is a common mistake in value trade off (Keeney,

2001). Thus, It is important to clearly be aware of the value tension and openly discuss these in the AI development.

2 I Not understand the (decision)context

is repeatedly mentioned a common mistake performed during the resolving value-conflicts (Keeney, 2001; Friedman et al., 2013). Hence, it is crucial to upfront discuss the context for which the AI system is designed.

3 I Misunderstand consequences

or not taking into account the consequences is a common mistake made (Keeney, 2001). Clearly discussing the possible consequences the technology might have, not only for the direct stakeholders, but also the indirect ones is crucial for developing desired AI systems.

4 I Confusing goals with means

is also a regularly made mistake (Keeney, 1996), discussion and clear distinction of the end goal of the output of the AI system should be very clearly communicated across all the team.

4.2.5 Need for empirical research

Not only research is necessary into which to values integrate into an AI system, but also how in a certain context people prioritize values. **Thus, there is a need to perform empirical research towards hierarchical relations and trade-offs in specific industries and communities.** IEEE research institute pointed out, fixed hierarchical relations of values often do not fit. Thus, context specific value trade-offs would be more suited. **To achieve this, context specific input play sa crucial factor in the understanding of the subtle context specific differences that fuel the value trade-off hierarchies in AI-systems.**

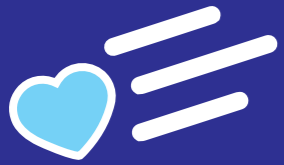
4.2.6 Conclusion

To conclude, five value tensions are identified in literature specific for AI development. In this thesis, further research is performed to explore if these are addressed in the AI development. Also, this research aims to explore new value tensions and how (if) these are resolved currently at which moments in the process.

» Six main strategies to solve value tension are extracted from the analyzed tools and approaches into a conceptual framework (figure 5.6): (1) untangle values, (2) decompose values, (3) avoid problematic features for stakeholders, (4) decentralize responsibility, (5) quantify values & consequences (7) untangle consequences. These strategies are often used hand in hand and fuel the ideation phase of this thesis in creating the practical support to resolve these in a desired manner.

Concluding, value-tension is a under researched phenomenon. All analyzed tools/approaches have little to no evaluation beyond academic setting (Miller et al. 2007; Shilton, 2018) or bear still much critique. Especially in AI development the value tensions are approached from a technical perspective and with the strategy of quantifying values. This is problematic because quantification does not apply to moral values. Besides, this perspective does not account for the context specific fairness and values, potentially leading to ethically misaligned systems. In this thesis is aspired to take context specific fairness into account. A crucial factor in this is to understand and explore the subtle context specific differences. Currently, few research has been performed into the discovery of AI specific value tensions in the actual development process of AI systems. This is a under researched field, and there for a novel argument to put this thesis forward.

» **Finally, in order to design support to resolve value tensions related to fairness in AI, empirical research into these value tensions is required. As values are latent, generative tools and provotypes are used to discover these, discussed in the next chapter.**



Value tension

Need to resolve value tension in AI

Not addressing value tension in an explicit way can lead to a lack of appropriation by disadvantaged groups, system sabotage or unfair outcomes of AI systems. Thus, there is a need to resolve value tensions in AI development to prevent drastic consequences.

Five value tensions specific for AI

- Accuracy vs Fairness
- Explainability vs Performance
- Bias vs variance
- Precision vs Recall
- (Historical) data value vs Socially desired value

Four common mistakes in resolving value tensions

- Misunderstanding of the value tension
- Not understanding the decision context
- Not understanding or not taking into account the consequences
- Confusing goals with means

Value tension is under-researched

Value-tension is a under researched phenomenon. All analyzed tools/approaches have little to no evaluation beyond academic setting or bear still much critique. Especially in AI development the value tensions are approached from a technical perspective and quantifying values. This is problematic because quantification does not work for moral values, it does not account for the context specific fairness and values. Thus, this is a novel research direction.

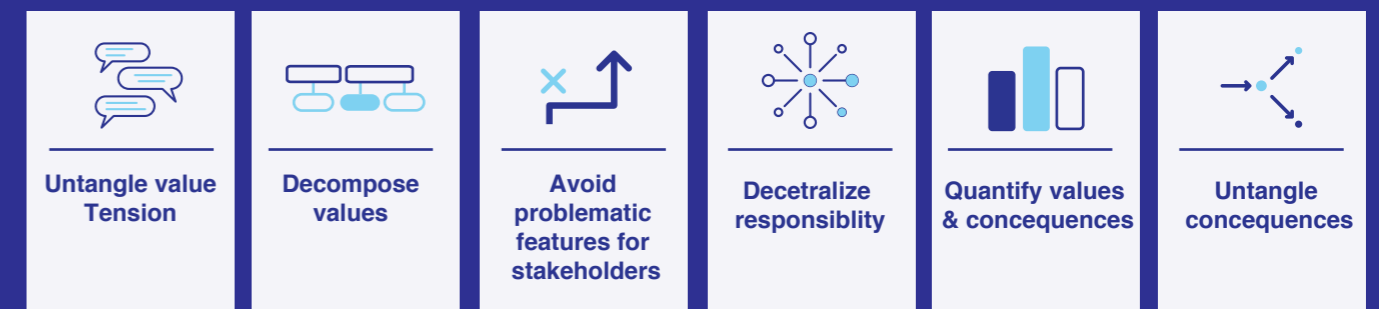
Empirical research needed into context

It is necessary to do empirical research towards hierarchical value relations and conflicts industries and communities. To achieve this context specific input plays a crucial factor in understanding the subtle context specific differences that fuel the value trade-off hierarchies in AI-systems. Thus the contemporary AI field is researched.



Six identified strategies to resolve value tension

>> Six main strategies to resolve value tension are extracted from the analyzed tools and approaches: untangle values, decompose values, avoid problematic features for stakeholders, decentralize responsibility, quantify values & consequences and untangle consequences. These strategies are often used hand in hand with each other and these fuel the ideation phase of this thesis in creating the practical support to resolve these in a desired manner.



Six strategies to resolve value tension | a framing result founded in the literature review

Chapter 06 I

A peek in the kitchen

Exploring the contemporary

This chapter shares the design research insights concerning the contemporary AI field. Semi-structured interviews, generative tools and provotypes founded in the previous literature and analyses are used to uncover insights for the design.

In this chapter

- 6.1 Design research set-up
- 6.2 In depth interviews & generative tool
- 6.3 Provotypes
- 6.4 Value tensions for design

6.1

Design Research Set-up

This section elaborates on the research approaches utilized and designed in this thesis to explore the contemporary AI teams and processes.

Overall goal I » Design practical support for the AI development team which will lead to more value-aligned and fairer AI systems.

From the literature, internal and external analyses, the need for empirical research in the field of AI ethics is withdrawn. Few research exists concerning how to aid AI teams in the creation of fairer and value-aligned AI systems (from a non-technical perspective). It is therefore imperative to first create an understanding of the current state of the practical AI field before designing. For this goal, interviews are selected as the research approach (figure 6.3 for overview).

To gather more latent and tacit knowledge (see figure 6.1), concerning the team members' personal values and the value tensions a complementary approach is vital. A generative session is chosen for the extraction of latent knowledge (figure 6.2)

Gathering deeply latent knowledge regarding value-tensions in one session is difficult. This allows for only limited conclusions to be drawn. Thus, with the insights from the interview and

generative tool, provotypes are made to discover and confirm the value tensions and how they are solved in AI practice (see figure 6.2). The further goals and research strategies are elaborated upon per specific part in this chapter.

6.1.1 Interview participant selection

Concluding from the value literature, values are profoundly context specific. Thus is chosen to specifically look into one industry of AI development. Five preliminary interviews at IBM Benelux are conducted to identify the industry which the needs ethics incorporation uttermost (from the AI teams' perspective). Four data scientists and one technology architect are informally interviewed.

» **Based on these interviews the insurance AI industry is chosen as scope. It is experienced as an ethical sensitive industry to work in as a data scientist.** Additional requirements for the case selection are (1) the use of machine learning approaches and (2) the use of human data or data from humans. When using human

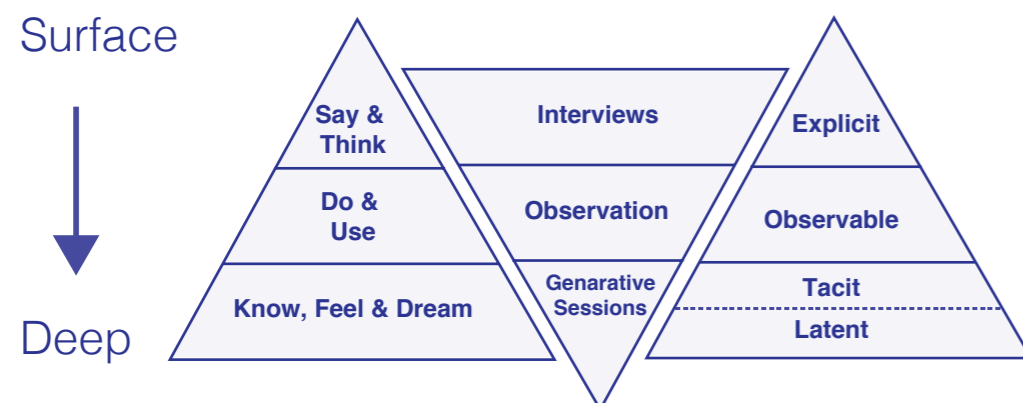
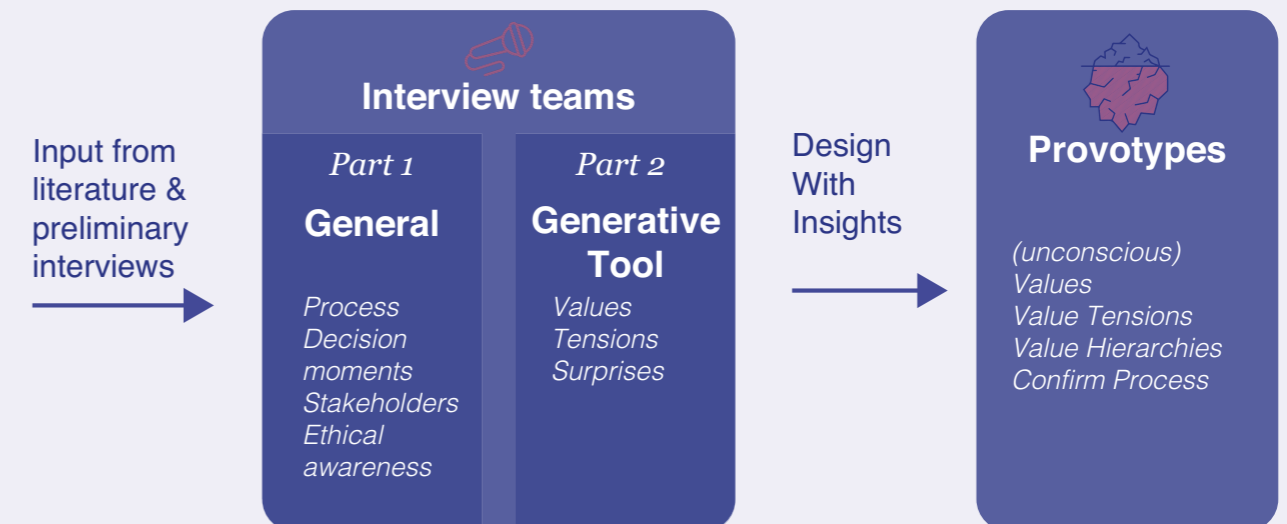


Figure 6.1 | Research approaches from explicit to latent

Design Research Set-up

Figure 6.2 research set-up with goals



Participants in the generation of new ways of working on value alignment in AI-development.

	Exploratory interviews context	Expert interviews thesis topic	Interviews specific insurance AI teams + provotypes	Testing Concepts	Interviews validation concept	Testing Final Design
DESIGN	4 IBM internal			3 1 design student team	1	
AI	6 IBM internal	5 IBM internal	7 4 teams IBM & Client	8 2 teams CAS IBM internal		7 2 teams IBM & Client
ETHICS		5 Including Jeroen van den Hoven & Lammert Kamphuis			1	
IBM other					4 Concerning implementaion	

Figure 6.3 | Interviews held for this thesis overview

data in machine learning, unexpected ethical issues arise soon in relation to fairness. Thus, specifically into these types of AI development projects are researched.

From this departure point, IBM's clients in the insurance branch were contacted. A worldwide insurance company agreed upon a collaboration for this research. Teams from Belgium and the Netherlands were interviewed. The same teams participated throughout the whole research. See figure 6.4.

The further sections elaborate on the specific research elements separately, sharing the goals, structure and insights.

Reminder research questions

- 1 » *How to create an organizational capacity and infrastructure to support ethical uptake in AI projects?*
- 2 » *How to support the AI team and in which phase, for fairness in AI projects?*
- 3 » *How to support the AI team and in which project phase for, value-alignment in AI projects?*
- 3.2 » *How to support the AI team to resolve value tensions in AI projects?*

6.2

Interviews & generative tool

This section shares the goals, methods, development and insights from the semi-structure interviews with the use of the generative tool.

6.2.1 Research goals

In-depth interviews are held with four AI development teams using a tailored generative tool. The following roles were interviewed, data scientists, manager/scrum master and the business owner from four different projects.

» The main goals of the first interview part are:

- **Discover the current AI process with the ethical decision moments**
- **Discover the current (perceived) roles and responsibilities in daily work**
- **Discover the (perceived) stakeholders**
- **Discover the ethical awareness and sense of responsibility, reflexivity.**

» The second part of the interview is concerning more latent knowledge and therefore made use of a specifically designed generative tool. The main goals of this generative tool are:

- **Discover value tensions during the process**
- **Discover surprises in the process**
- **Discover the values of the interviewees themselves as well as they perceive them, of the model and the other team members (first iteration)**

6.2.2 Research approach

Interviews of approximately one hour where conducted at the insurance company over a period of two weeks. The projects are shown in figure 6.4 The interview guide, with the questions with subject areas is shown in appendix R. The interview was

semi-structured. In other words, the interviewer is free in the use of words, spontaneous questions and order of them (Patton, 2002). However, the focus of the questions is on the specific subjects determined in the interview guide. The sub-topics are: General project process and stakeholders, surprises, values, and fairness reflection. During the interview it relevant the interviewer remains within these areas to prevent biasing the interviewees answers. Questions started with more general topics such as their job description and the project after which step by step went to questions related to more latent knowledge.

All seven interviews are voice recorded and notes were taken simultaneously for further analyses. Detailed analyses of the interviews are in shared in appendix K.

6.2.3 The generative tool design

Founded in the ethics, value-alignment and fairness literature review, a generative assignment is designed (figure 6.6). Through these generative exercises participants are triggered towards deeper layers of knowledge (Sanders & Stappers, 2008).

The generative exercises in this research have the following subtopics:

- Stakeholder mapping (direct & indirect)
- The projects process with the important project moments, the personal important moments and the challenging moments
- Mapped value tensions and their relation object/subject



Figure 6.4 | Researched project and teams

- Surprises encountered during projects
- Mapping of personal values, mapping of the values inscribed in the model ((un)desired), potential/performed values, and the expected values of the other involved stakeholders

Values are a challenging and abstract topic to discuss. Thus, a selection of different values was made in a form of small cards for inspiration. Values were extracted from IBM's strategy (such as serving client, augmenting humans), human values (well-being, connection, self-expression), moral values (respect, trust, responsibility) the ICT values (such as accountability, calmness), as well as through clustering of the most well-known AI ethical development principles out there (safety/security, awareness). The full generative tool is shared in appendix K. The goal of these cards is to support the interviewees value thinking and to explore which roles would choose what type of values. Also, the opportunity to add values themselves was provided.

6.2.4 Data analysis

All seven interviews are transcribed after which statement cards were created (Sanders

& Stappers, 2013). The statements from the interviews were clustered multiple times to find patterns and gather insights related to the research questions and compared against the findings from the literature. A detailed overview of the data analyses is shown in appendix K.

6.2.5 Insights and Findings

The AI development process with identified ethical decisions moments and value tensions is visualized in figure 6.8. Furthermore the research identified the roles of the AI team and synthesized it in persona's, represented in figure 6.5. A more detailed view of the persona's is shown in appendix K, used for an empathic view of the team. The further insights are briefly addressed.

- A lack of technological knowledge of the entire team (non-technical oriented people) and the ethical implications it brings with it is discovered. Thus there is need for team alignment on technical capabilities as well as ethical pitfalls.
- A lack of moral motivation is discovered. Thus, there is a necessity to increase intrinsic motivation for ethical decisions in projects of

“If we are allowed to use it we should use it, it is part of the game haha..”

- Interviewee about personal data (Manager)

“To be honest, I completely do not care how the model works, I find the output much more important.”

- Interviewee (Business owner)

“Haha I see there is a lot of ethics stuff I am not considering!”

- Interviewee (Data scientist)

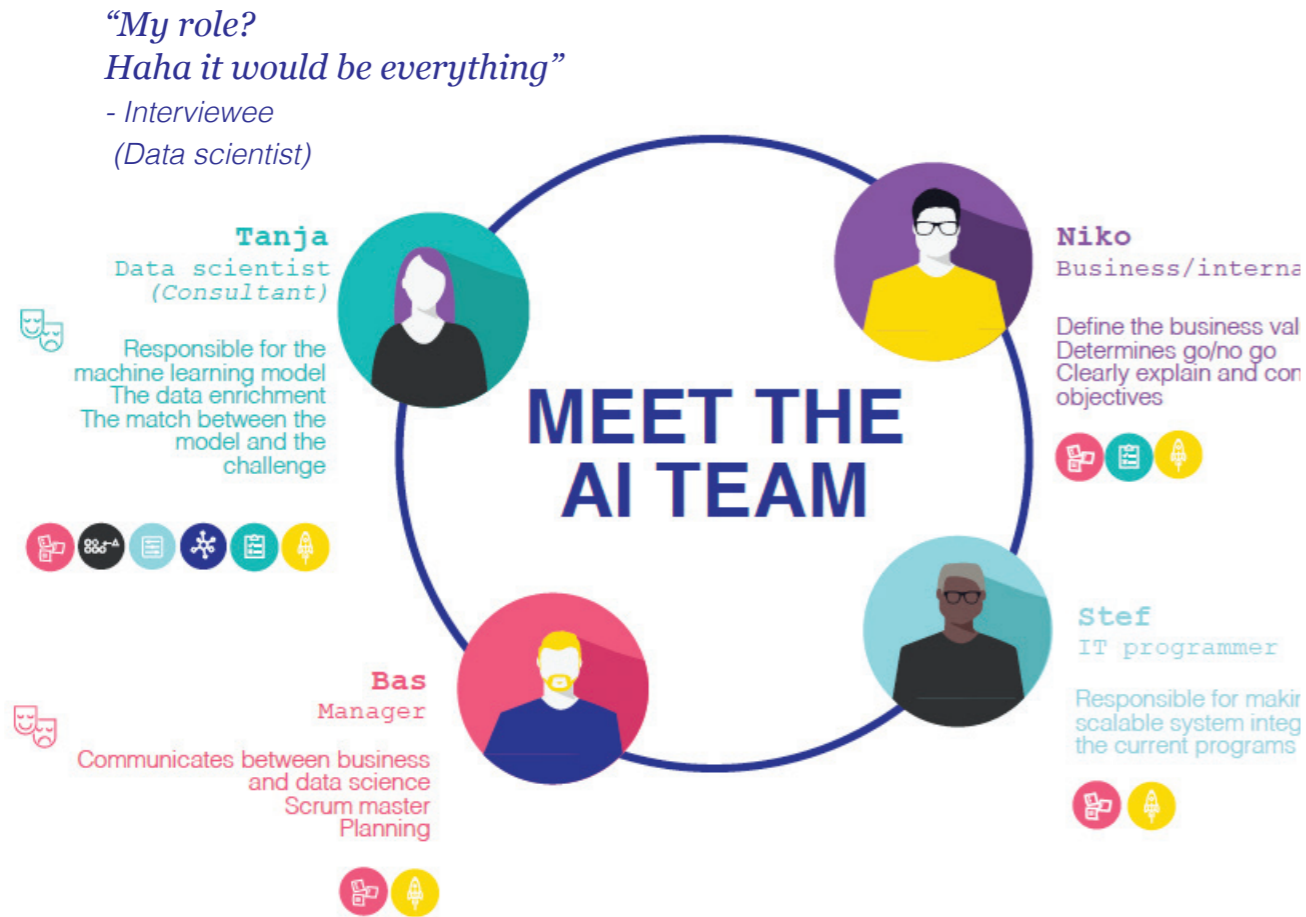


Figure 6.5 Roles of the AI team and in which project phase they are present

Tension

Retrieved from interviews

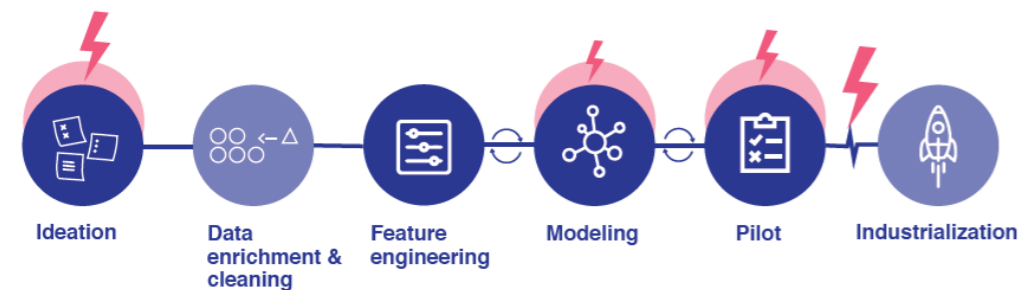


Figure 6.8 | Results from the research, process, value tension moments and project decision moments

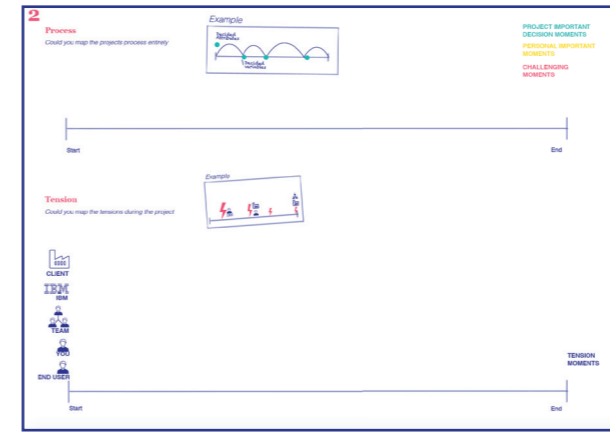


Figure 6.6 | Example of a sheet of the generative tool

the team and especially for the data scientist role as he makes many ethically impactful decisions by himself

- Lack of incorporation of stakeholders and (societal) consequences of the AI system. Engineers might need some check points to make sure they are not too ambitious to solve the technical issue. Thus, integration of the consequences of the models predictions in the ideation might support this.
- A lot of roles and tasks are the responsibility of the data scientist and therefore this role experiences a lot of pressure. Furthermore the majority of the decisions taken in the feature engineering and modeling phases are made by the data scientist alone. Reducing pressure put on the data scientist and highlight the importance of decisions in the

modeling and feature engineering phase could support more ethical outcomes.

- Value tensions nor values are consciously addressed. Thus, explicitly discussing values and solving value tensions for desired outputs is proposed.
- Unexpected challenges occur during the process and lead to (ethical) surprises.
- Lack of integration of indirect stakeholders and also no end customers (only 2/7). From the ethics literature appeared, for ethical outcomes it is essential to integrate the stakeholders opinions both direct and indirect.
- The following value tensions are used for the prototypes, chosen with the lens of the research in mind (design for fairness): (1) **socially desired value vs historical data**, (2) **simplification vs uniqueness/ veracity**, (3) **responsibility/ accountability vs autonomy/freedom**, (4) **probity vs accuracy** and (5) **explainability vs performance**.



Figure 6.7 Analog analyzing of the interviews

6.3 Provotypes

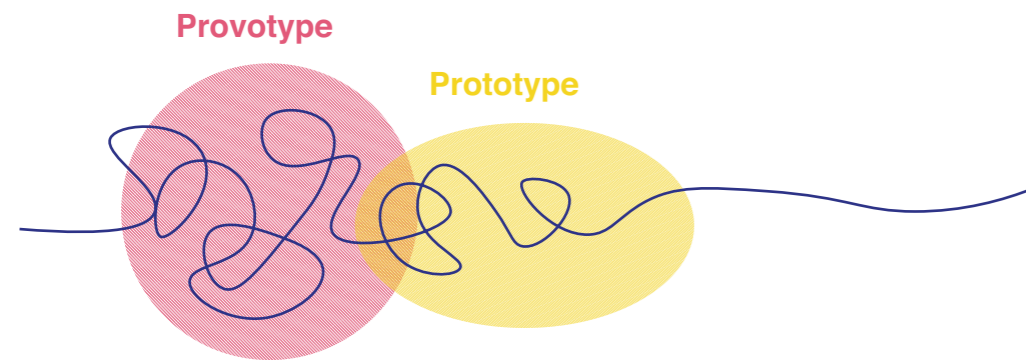


Figure 6.9 | Visualization of the design phase in which the provotype is used compared to a prototype

6.3.1 Theoretical background

From the interviews and the generative tool, a first glimpse of the preferred values and the value tensions is extracted. Discovering values is complex. The desired levels of personal and more latent/unconscious values is not reached in a desired depth and with enough confidence. Thus, another research approach to reach in the insights and unravel new value tensions is necessary. Provotypes are chosen due to its open and critical manner of inquiry, challenging the state of the art.

» **Provotypes provide an opening to conflicts in processes, these are artifacts/pictures that embody tensions in a certain context in order to explore new design opportunities (Boer & Donovan, 2012).** In other words, a provotype is a provocative prototype, used earlier in the design phase, the explorative phase (figure 6.9).

It can be used to explore a new problem/solution space by stimulating discussions around deeper unmet needs for desirable futures (Boer & Donovan, 2012). Thus, provotypes are chosen as a suitable method for this study to discover more latent value tensions and hierarchies throughout the interaction the AI team obtains with them.

6.3.2 Goals of the provotype

» The provotypes have following three goals.

- **Discover which values the interviewees prefer over others and reach a more latent level, discovering novel (un)conscious values**
- **Discover how and if value tensions are resolved by the current AI team.** To gain a richer understanding concerning these value trade-offs and how these are currently resolved in AI development, is targeted by the provotypes. Also differences between the functions are researched in terms of values
- **Test the extracted process from the interviews.**

6.3.3 The provotypes design

Two examples of the provotypes are shared in figure 6.10 and 6.11. The provotypes are provocative demonstrators of things or services that show an extreme form of the value-tension discovered from the interview and/or literature. Not all provotypes were shown to AI participants a switch was made between the responsible spending and responsible freedom per

FAIR PRICES.

Coffee place 2025

1! What is your first reaction? How would this scene continue?

2! Would you like to work on making a system like this? Why yes/not?

3! Do you consider this as fair and accurate prices? Why yes/not?

4! What values might be important to each person or group that would be affected? Try to think of at least 2

5! If the values are different, how would you resolve those values contrasts? (think about the different stakeholders)

D.P.Simons 2019



Figure 6.10 | One of the designed and used provotypes

participant. The provotypes were personalized in name usage and small details to increase the empathy with the scenario. The provotypes are tested upfront with a computer scientist for clarity. The following tensions are addressed: (1) socially desired value vs historical data, (2) simplification vs uniqueness/veracity, (3) responsibility/accountability vs autonomy freedom, (4) probity (fairness) vs accuracy, (5) explainability vs performance. They provotypes were send and answered by email contact.

6.3.4 Results and findings

The answers of the provotypes are all read, analyzed, summarized and compared to the answers between the participants. This analysis is performed with the lens of the research questions and goals in mind. The full analyses is shown in appendix M. A selection in shared in the following paragraphs.

Overall

Remarkable from the previous identified value tensions, the fairness vs accuracy and historical data value vs socially desired one did not reoccur in the answers. The performance of the models was also not specifically mentioned. An explanation for this could be the participants

answered the provotypes from the perspective of being the end user and less as the maker of the model. By means of interviews and literature these tensions are highlighted as crucial for the models fairness. Thus, these are chosen to still consider them in the design. It could mean that more education and explanation concerning these value tensions in the teams is necessary. The following insights are briefly described per research goal.

01 Goal one insights

1. Support is needed to move away from technical thinking. Educating the AI team about ethics is necessary.
2. New values and value tensions are discovered shared at the next page.
3. Scenarios work well for the AI team to empathize with the end user and support ethical reflection.

Example Answer Airport Provotype

"Again, quite possible. But it cannot be just based on gender alone. If we can match images to a database and can immediately detect 'less risk' passengers compared to moderate/high risk, we can create separate lane for less risk customers. Similar to 'nothing to declare' customs lines in airports. " - Data Scientist

- 4. When the participants own personal data was used, stronger signs of ethical reflection were expressed.
- 5. Explicit responsibility can be taken in two ways. On one hand good on the other hand it is also easily put on other people/ organizations (detected with manager roles).

02 Goal two insights

- 6. The following value tension in relation to fairness have been derived. (1) freedom/ privacy vs safety/control; (2)simplification vs authenticity; (4) individual benefit vs collective benefit (can be employee vs company benefit, can be risky individual vs careful citizens), (5) technically interesting vs socially desired.
- 7. The participants resolved the value tensions in the following manners:
 - Not make this system (unresolvable)
 - Change the data on which it is based
 - Change the type of decision making from autonomous to supportive changing the out put from the system (Example of optimized performance provotype from punishing to rewarding)
 - Add functions to the AI system for explainability, transparency, awareness of responsibility.

Example Answer Skiing Provotype

“Good signal. And looking at the social costs that come with it, for example an avalanche, it is good that you get one more time a reminder. It is similar to a warning with trajectory control. To be honest I really like skiing off-piste. My own consideration would be a risk consideration.” - Business owner

- 8. The different roles of the AI team did not show too much difference in answers. Only managerial/business participants expressed their thoughts more concerning responsibility as well as more often referred to others organizational responsibility (e.g. government). Data scientists expressed stronger affinity with freedom than other functions.
- 9. Almost all participants mentioned that If the company making the AI system does not take the responsibility, it should be very clearly communicated to the end user that it is their own responsibility.
- 10. The “deserved” perspective on fairness re-

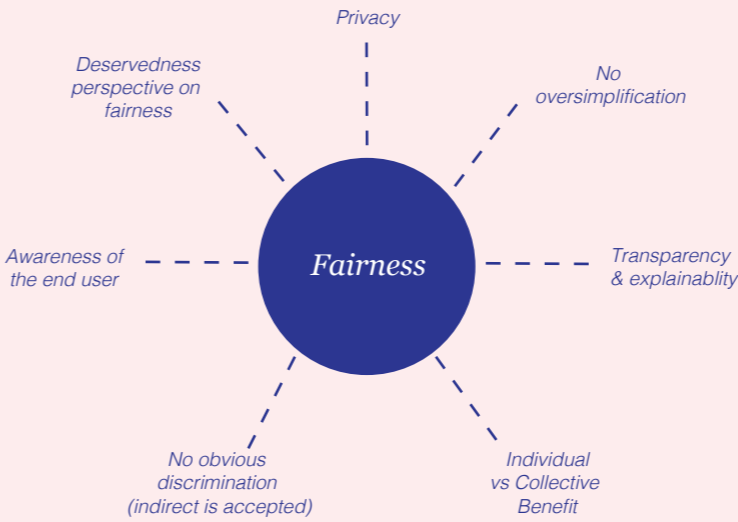


Figure 6.12 | Visualization - fairness from the AI team perspective by means of provotype

- occurs to be the more preferred one (described in the fairness chapter).
- 11. One value that appeared to be important to most of the participants was privacy, and the right of privacy. When systems become too intruding this was not well accepted and even labeled as unfair.

03 Goal three insights

- 13. Validated project process, moments of tension and decisions moments.

Insights Provotypes

Educate the AI team about ethics

Support is needed to move away from technical thinking and creation of ethical awareness by education is crucial.

Scenario’s stimulate empathy with end user

The participants responded to the provotypes mostly from the end users perspective and therefore expressed moments of reflection. The scenario’s supported the participants in doing so. Using scenario’s increases the ethical empathy for the end user.

Resolve value tension

- By not making this system (unresolvable)
- By changing the data
- By change the type of decision making e.g. from autonomous to supportive
- By adding features to the AI system for explainability, transparency, awareness of responsibility

These examples of changing features of the AI system to make it more acceptable can serve as input for the solutions space in resolving value tension.

Personal use of data stimulates ethical reflection

When the participants were addressed personally or their group was treated unfair, ethical reflection was stimulated. This is an attractive strategy to trigger this reflection.

Careful with explicit responsibility

When explicitly discussing responsibility, attention should be paid that it is not just transferred to other organizations or institution, especially with managerial roles.

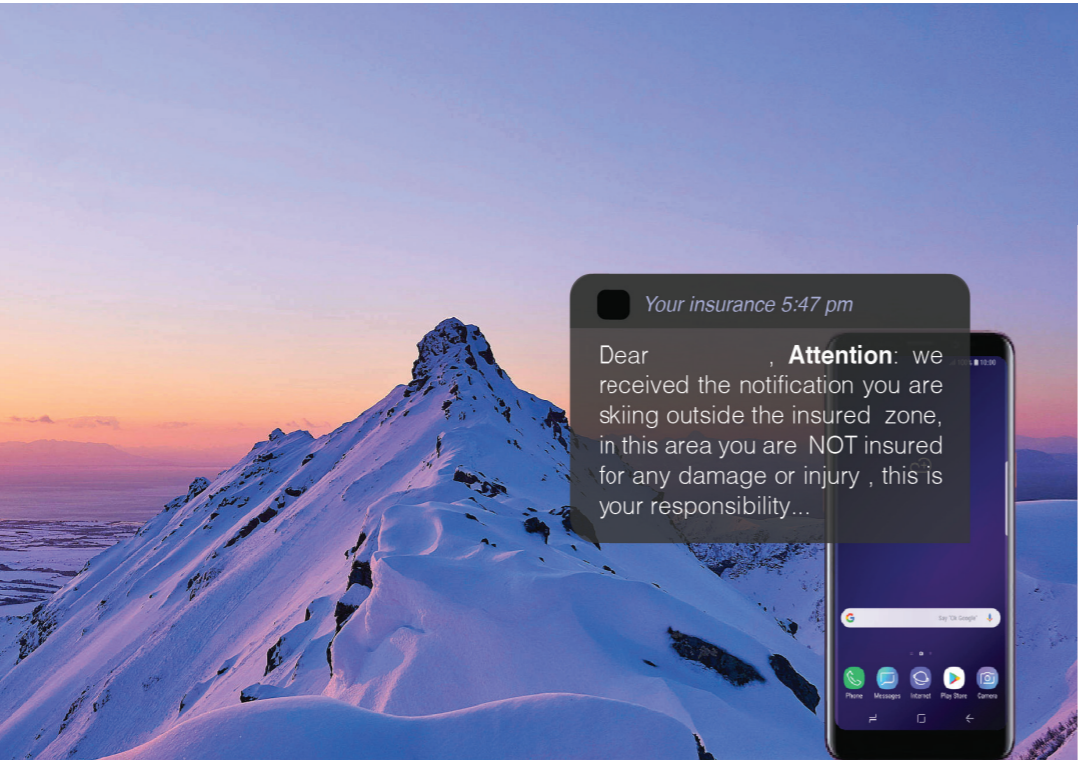
On the next page the value tensions are addressed

Figure 6.11 | One of the designed and used provotypes

RESPONSIBLE FREEDOM.

Skiing holiday 2030

- 11 What is your first reaction? How would you continue?
- 21 Would you like to work on making a system like this? Why yes/not?
- 31 Do you consider going on this slope your personal responsibility? Why yes/not?
- 41 What values important to you might be affected with this system? Try to think of at least 2.
- 51 If the values are different for the different stakeholders, how should you resolve those values contrasts?



Example value hierarchy choice Optimization provotype

“No. even if “technically speaking”, the project is really interesting and probably one of the most system ever build, I would not want to be part of that because there are too many bad things which could happen with that kind of system.”
- Data scientist
(Interesting work vs socially desired)

Value tensions for design

» *The choice is made to continue with five main value tensions, based on literature, the interviews, the generative tool and the provotypes analyses. The lens the goal, fairness in AI is used in the decision. The explanations of the following value tensions are based on a combination of sources. The sources differ from explanations in literature of the values, example cases of unfair AI and the performed empirical study these are constructed by me.*



Individual benefit vs Collective benefit *(provotypes)*

Individual value versus collective value is a dimension concerning who benefits from the model, based on more individual level or as society as a whole. From the provotypes the different analyses of the fairness on individual or more collective level (such as employer, the country or globally) are distilled. Although this is a lens which also can be applied with the other value tensions, this trade-off clearly was present in the answers of the provotypes and therefore is chosen.

Individual benefit: In this case the model brings advantages to (certain) people on an individual level at cost of the collective benefit.

Collective benefit: In this case the model brings advantages to the collective (the company, the country, the world) at cost of the individual benefit.

Why is this a tension? Often benefits for society,

the country, the employer, are in contrast with individual ones. An extreme example would be closer to communism, in which more people are equal (so no/less poverty) but also less wealth for specific individual who might work harder. With AI models it appeared to be a common trade-off. Does the team want to support more individual benefit or more collective benefit? This is highly depended on culture, industry and context.

(Historical) data value vs Socially desired value

(Literature & example cases)

Values represented by historical data: Values represented by historical data can be sometimes different than the values that are desired in society. Also, values represented by the model due to incomplete training data or un-matching context can support the value conflict with data one and the socially desired one.

Socially desired: A model with socially desired values, is for the social good and/or with benefit for humanity.

Why is this a tension? Relating to the Amazon example of the biased hiring system, things that happened in the past are not necessarily desirable, in the future. Also data is very context depended and highly depended on the data source.

Explainability vs performance

(literature, example cases)

Explainability: The capability of the model to be understood, the model being interpretable. The understanding of how the decisions are made by the model, based on which data increases the perceived fairness of the model. It also allows to discover other sources, perhaps undesired ones, on which the decision is based.

Performance means an action/process, how well somebody/something carries out work or an

activity. In this case the model, so for example how accurate, fast it preforms the tasks if that is demanded. Often black box models which work really well, need also less training data and therefore are also an appealing choice.

Why is it a tension? Systems are usually at odds with each other, as many of the best-performing models (viz. deep neural networks) are black box in nature (Dhurandhar, 2018). An example of this could be the COMPASS example in the American crime system, for a long time it was not clear based on which data the system predicted the likelihood of somebody performing another crime. Later was published that one of the data points it used was skin color, negatively discriminating non-white people.



Freedom/privacy vs safety/control *(provotypes)*

This value tension is described in two words as the aspects of these words were used together in the empirical study by the participants. It nuances the type of value tension.

Safety/control: In this case safety means, protecting the people using the system from negative outcomes. However, this also means taking a part of the decisions power away of the user. So, often it simultaneously takes the responsibility partially too.

Freedom/privacy The ability to make your own decisions without being controlled or negatively affected by anyone/something else. Privacy is a value used to describe this type freedom in the empirical study. Privacy here is the state or condition of being free from being observed or disturbed by other people/organizations/systems, on a personal level. This is very context dependent. For example, in a hospital one probably shares more information with the doctor than with one would do at a retail store.

Why is it a tension? Do you give the user the

responsibility of using the created product? Or does company/the team creating the system taking the responsibility for the consequences it might have. Freedom in this notion also means to have the freedom to make new models, innovate, move quickly. Security in this notion means more principles and rules that will secure the outcomes but limit the freedom in peoples work. In specific the data scientists appeared to value their freedom strongly, both professionally as personally.

Accuracy vs Probity

(Literature)

Accuracy: The degree to which the result the model conforms to the correct value or a standard. **Probity:** impartial and just treatment or behavior without favoritism or discrimination by the model the quality of having strong moral principles; honesty and decency.

Why a tension? If one aims to make the models probity high and fair, then often one needs to hand in on accuracy of the model as it needs to take into account perhaps different data, simplify less and thereby lose on accuracy,

» Presented are the five value-tensions focused on in this thesis.

Based on these tensions the ideation phase starts. The focus lies in how to resolve these tensions in an ethical fashion per project, in a context specific manner. Ultimately, this leads towards the co-creation of fairer AI systems.

Chapter 07 I

Preparing the Ethical Recipe

This chapter shares the synthesis of the literature and empirical study, which forms the basis for the develop and deliver design phases. It takes the form of a design vision and framework, which serve as the foundation of the ethical recipe for a fairer AI dish.

In this chapter

- 7.1 From insights to design
- 7.2 Ethics in AI framework
- 7.3 Design guidelines

7.1



From insights to design

In order to assist AI teams in the development towards fairer AI systems it is of great importance the design fits their current development process. The ethical deliberation process is linked to the needed support per stage. A vision to spells out the line of thought which founded the ideation phase. The insights gathered in the project are consolidated into a framework (by means of generative interviews, provotypes, expert interviews and literature). The goal of the framework is to design tools for ethical AI support for IBM Benelux and provide design directions for organizational capacity in the development of fairer AI systems. This framework and vision served as a basis for the ideation phase.

7.1.1 Vision for design

This vision spells out the line of thought behind the design phase. Founded in the analyses, literature review and design research the following personal design outlook is taken to design for fairness.

01 Ethics as fuel of co-creative innovation

This thesis takes the notion ethics should not be seen as a prescription, as it currently is often seen. Rather to adopt ethics as a fuel of innovation in the AI field in a co-creative fashion. This is aspired by imaginative activities stimulating creativity and by opening up the solution space. In line with the expert interview with Aimee van Wynsberghe beneath.

02 Prototypical kind of ethical thinking

This thesis proceeds in the line of thought Lloyd (2009) who calls designing a prototypical kind ethical thinking. Therefore, ethics benefits from

this design thinking. Additionally, in design the problem unfolds during the process, leaving opportunity of finding the real problem (Van de Poel et al. 2007; Withbeck, 1998), which is the line of thought applied for the development of fairer AI systems. In this process surprises occur, agile methodologies are used. Thus, the prototypical kind of ethical thinking is argued to accommodate and intended to enhance the AI practice advantageously. Hence, stimulation of ethical imagination and creativity within the AI team in designerly fashion is asserted, in assisting to more ethical solutions and opportunities. Additionally, supporting prototypical reflection to increase ethical awareness and considerations is proposed.

03 Proactive stance in designing for fairness by reducing unfairness sources

In this thesis the proactive stance to design for fairness is taken. It is intended to consciously and explicitly design for the value fairness, promoting this value of interest in the AI development process.

Instead of prescribing what is fair or not fair, this thesis intends to support the reflectivity of the AI team to reduce unfairness sources in the AI development by resolving the identified value tensions (chosen with the relationship to fairness).

“ You can use ethics as a manner to stimulate innovation, but you can only do that when you have someone whose main role is ethics, as a member of the design team, and they’re there through the different stages of the product being developed.”

Aimee van Wynsberghe - 14th of January 2019
Personal communication

04 Context matters

The line of thought in this thesis is that, values are context specific, while at the same time certain values in AI development should be strived for. Thus, this thesis takes situational as well as central values into account. In line it aspires to consider context specific fairness and that context of use should be explored in order for the system to be tailored towards desired fairer result.

05 Increase moral motivation and reflexivity

Data scientists often work by themselves in project phases such as feature engineering and modeling, taking ethically important decisions by themselves. Therefore, in this thesis is argued to stimulate moral motivation and reflexivity is necessary to create fairer and value-aligned AI systems. In order to aid the data scientists to consider and take more ethical decisions, especially in these project phases. It is put forward, that without this, the ethical uptake in later project phases diminishes.

7.1.2 Linking processes

On the next page figure 7.2 shows the ethical process, linked with the discovered needs at stages of the AI development process. Both from literature and the empirical study appeared that first there is a need for more ethical people (one of the building blocks) before starting the ethical process. Thus, “ethical people” is included in the following framework. Additionally, the lack of implementation of ethics in day to day process is resembled by the interviewees answers. Thus, implementation after the ethical deliberation is a necessity for the creation of more fair AI systems.

Design goals are derived from the link to the processes and shown in figure 7.1.

7.1.3 Target group

The target group is the AI development team (often working as consultants) with a focus on

the data scientists. The AI development team can consist of: a manager, product owner, data science manager, data scientists, IT personnel. When projects are internal then often the departments for who the models are made are also involved.

1.

Create awareness & understanding of both technology and ethics at the AI team

2.

Creativity and synthesized thinking in the opportunity creation for resolving value tension

3.

Fulfillment of the ethical outcomes in the actual AI development process

Figure 7.1 | Design goals linked per process phase

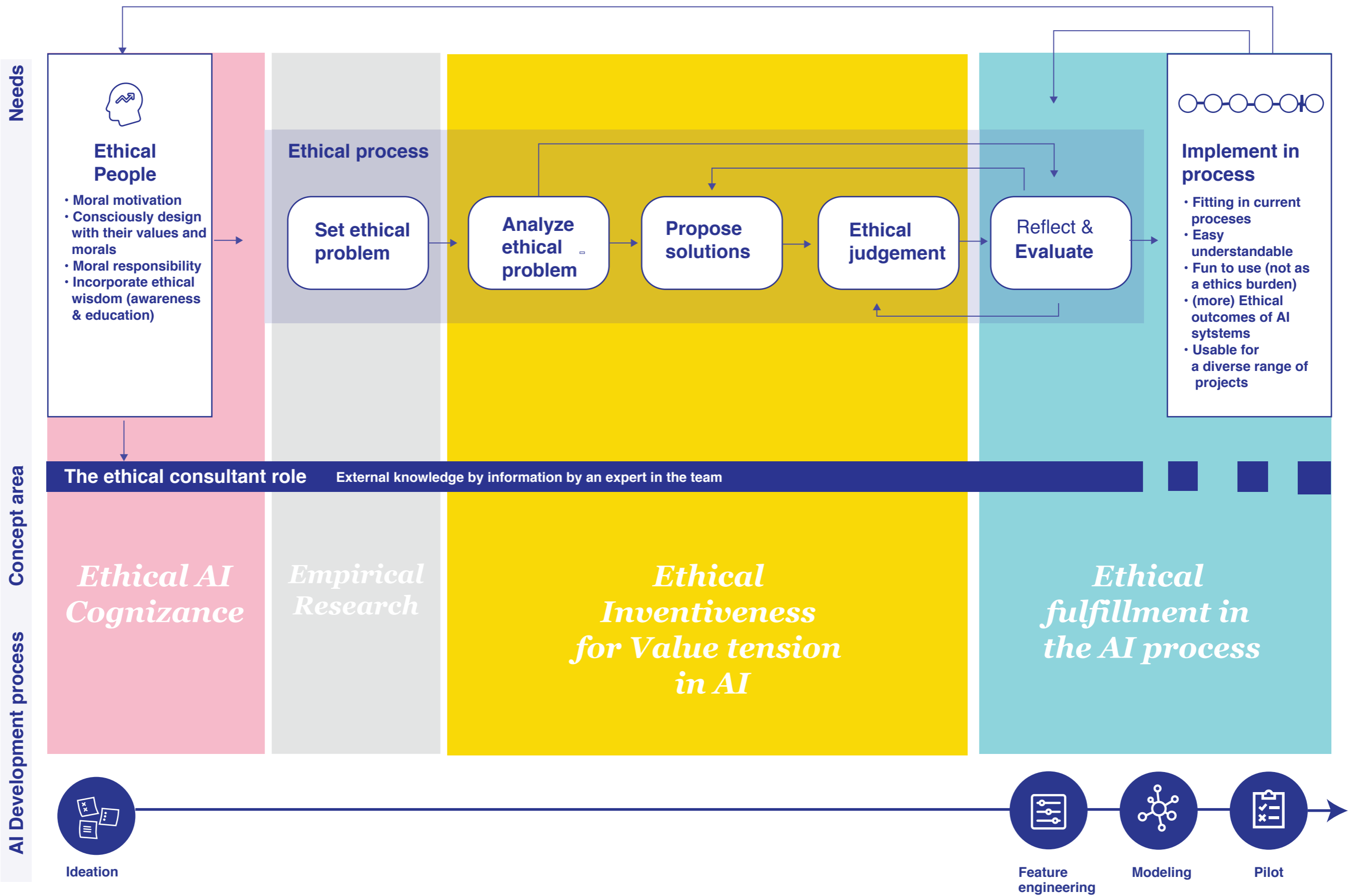
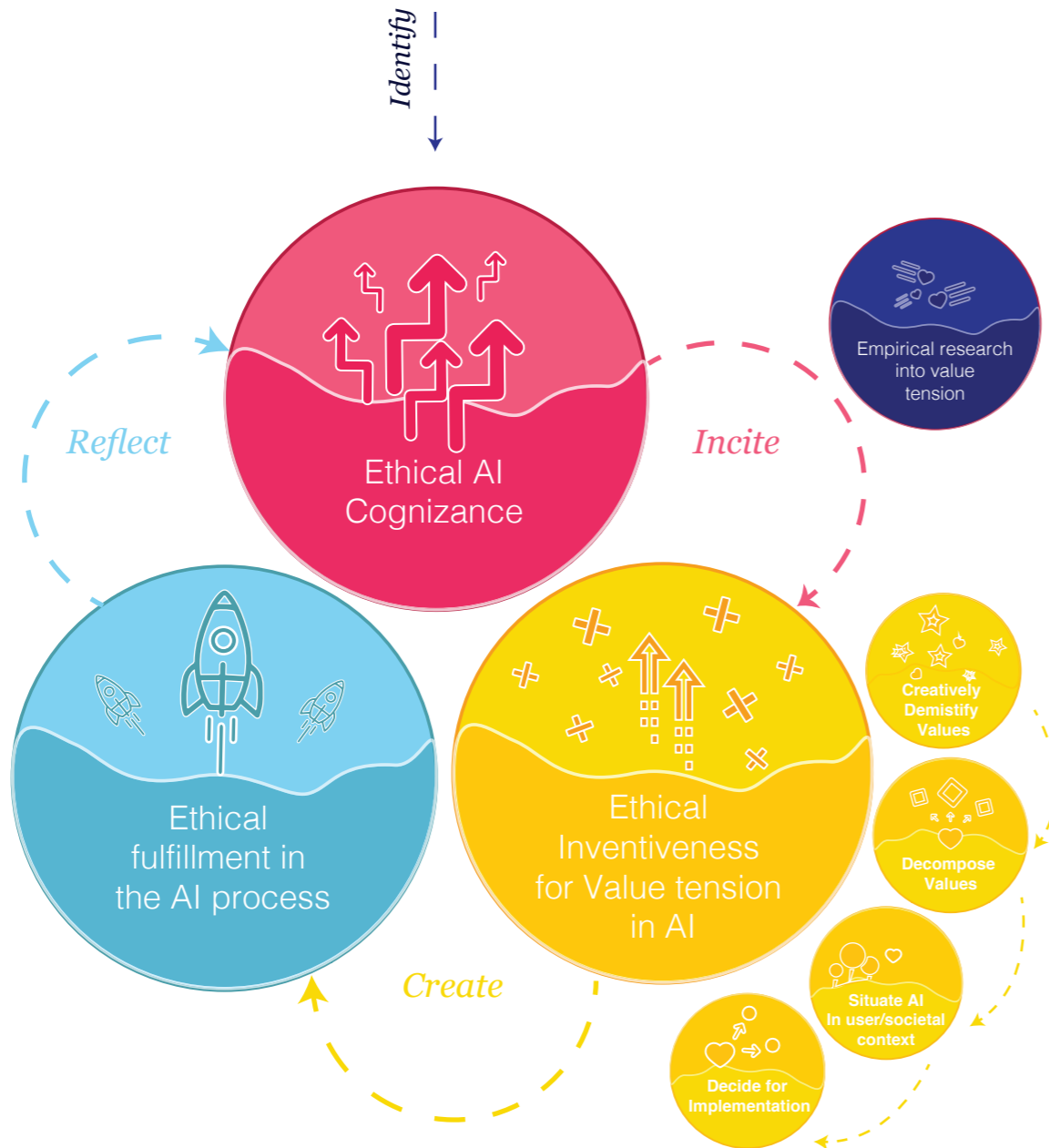


Figure 7.2 | Processes linked to framework concept areas

On top, the needs of the ethical process are visualized. One row below, the concept areas are identified to support this ethical process in the AI development phase. The last row shows the link to the AI development process to which the concept areas are linked.



7.2

Ethics in AI framework

This framework serves as a basis for the tool development process. The five main phases are briefly explained. The framework in figure 7.3 intends to portray a flow of the goals for the AI team for more fairer AI systems. The framework is open for reuse for the design for support for AI teams.

01 Identify

Not all AI projects need ethical support. In this thesis machine learning projects using human data were researched as projects which are more ethically sensitive (within the insurance industry). There is a necessity to identify if there is a need for more ethical support in the development in a specific project.

02

Empirical research into value tension

In this thesis value tensions are identified for machine learning projects in the insurance branch in the Benelux. Values differ per context, per industry, per country. Even though for this project the value tensions are chosen related to fairness of AI systems could be more general, for some other industries these identified value tensions might not be transferable. Then, there is a need for empirical research into value tensions for that specific context. For example the values in the medicine industry are distinct from the ones in the banking industry. Attention should be paid towards the different value sources of the value to design better support (p.66).

03

Ethical cognizance

It is crucial to have the team aligned in terms of both the technological understanding as well as the understanding of ethical implications. From the current research appeared there is a lack in both. Because of this unethical decisions are made due to misunderstandings, due to lack of moral motivation, moral responsibility, or ethical

knowledge. Thus a basis of understanding is needed as well as support for discussing the first ethical considerations.

04

Ethical Inventiveness for value tension in AI

Currently ethics bears too discussion and has a lack of actual creative solutions. At the same time, little to/no research has been performed into resolving value tension in AI. Even though it is of great importance to explicitly resolve these in the process to avoid ethical pitfalls. This part of the framework focuses on using creativity to solve the value tensions in a playful fashion. It aims to use ethics as a propulsion of innovation instead of a burden or restriction. The four steps to resolve value tension in AI are represented in the smaller circles and explained on the next page.

05


Ethical fulfillment in AI process

Implementation of actual solutions needs to be integrated in the actual AI development process. This needs to be done in a manner fitting with the AI projects, and not taking much time of the AI development team.

Figure 7.3 | Framework to design for fairness in AI

Ethical inventiveness for value tension in AI

This element of the framework serves as a basis for the design for resolving value tensions in **AI, inventively, co-creatively and context specifically**. The four main phases for resolving value tension are briefly explained. The framework in figure 7.4 intends to portray a flow steps for the AI team to resolve value tension related to fairness in their process. Beneath it shows the steps necessary for diverse types of values explained in detail at the end of this sub-section.

 The arrow shows the flow per value level.

01 Creatively Demystify Values

Founded in literature and expert interviews, explicitly addressing values leads to more ethical outcomes (Dignum, 2018; Flanagan et al. 2005; Miller et al. 2007 Van den Hoven, Shilton, 2018; Vermaas, & Van de Poel, 2015, p 838). It is crucial to analyze the ethical problem and explicitly resolve value tensions to construct fairer and value-aligned AI systems. This step in this framework aims to explicitly discuss the value tension in a fun relatable manner and not directly. So the AI team can easily relate to the values without being assaulted with the direct work they are performing. Infusing design techniques stimulating empathy and change of perspective. Creating a personal view and relation with the value is advised while changing perspectives to increase empathic thinking, then ethical viewpoints are more accessible for non-ethical experts (L. Kamphuis, personal communication 11th of December 2018)

02 Decompose Values

One strategy to resolve value tension which is repeatedly mentioned in literature is decomposing the values in a context specific fashion (Miller et al. 2007; Van den Hoven, Vermaas, & Van de Poel, 2015, p 838). Due to this thesis focuses on situational values

decomposing the values per case is a promising approach. This is the first step to proposing solutions (ethical process). Based on the ethical tool review is advised to decompose values step by step from abstract levels (values), to concrete (norms), to implementable features (procedures).

03 Situate AI system in user/societal context

Currently there is a lack of integration of the actual societal context and it's values in the AI development, while it has a vital impact on the (perceived) fairness of a system. It is proposed to explicitly discuss the interaction and consequences of the created AI system, early in the process (Despotou, 2005.) By the use of design, sparking creativity and placing the AI in the actual use context to catalyze designers both humanistic as well as technical imagination in a form of ethical reflection. With e.g. in designerly manners, stimulating synthetic thinking, empathy by creating scenario's, user stories, prototyping. Also, placing the AI system in the actual context use allows to see what might be a preferred feature/value in certain contexts. This is also the proposed stage to combine the two values of the value tension, allowing it to be resolved in a context specific fashion.

04 Decide for implementation

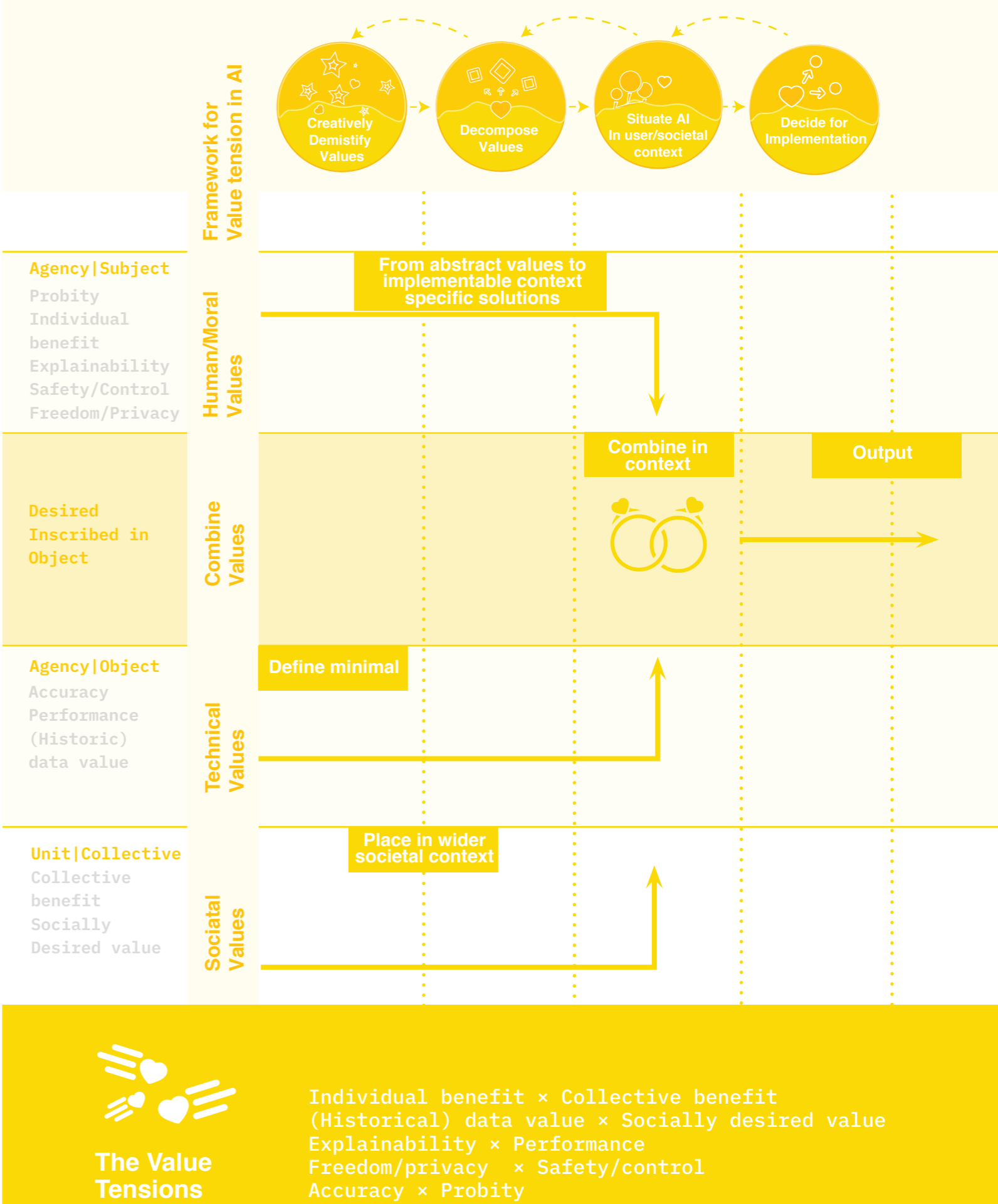
A critique on ethics is that serves discussions but little to no implementable solutions (Dorst & Royakkers, 2006). Thus by resolving value tensions by demystifying values, decomposing them, and placing in the user/societal context, decisions concerning the solutions need to be made for actual implementation with agreement of the entire team to stimulate practical uptake in the AI development process.

 05 Reflection

The reflective act in ethics is an essential one. Similarly in this framework reflection between the sequential phases is highly desired for ethical

Ethical inventiveness for value tension in AI

Figure 7.4 | Framework for ethical inventiveness



outcomes. This is in line with the prototypical kind of ethical thinking and iterations should be stimulated between the phases. The framework takes the two types of reflection into account. First order reflection, in which the team reflects on the outcomes. Also second order reflection is recommended, which requires a person to reflect on his or her background theories and value system (p. 41).

06 Value levels

Due to the different levels of values also diverse ways of decomposing and demystifying are proposed. Technical values such as accuracy can be defined by numbers and thresholds. These are values already (often) embedded in the AI system (agency: object) while moral and human values need to be explicitly discussed and decomposed in context specific fashion (agency: subject) which consciously and explicitly need to be inscribed in the AI system (move from subject to object dimension). Also distinction is made between the unit, individual or collective, in manners of decomposing on more use-case specific or if it is necessary to take societal benefit into account.

7.3 Design guidelines

Design guidelines per framework element are briefly explained. These are a result of the literature review, internal & external analyses, interviews and provotypes with the design vision in mind. The requirements are linked to the earlier mentioned insights.



I Ethical cognizance

- Create awareness and understanding of the technology and its implication, thereby the ethical importance of decisions, within the AI team (p. 39 & p.80)
- Align the AI team on ethical considerations (p.80)
- Stimulate moral responsibility & motivation (p. 40 & p. 80)
- Make consequences of the AI system and it's value tensions explicit (p. 67)
- An outsider perspective is strongly proposed (p.41 & p. 82)



I Empirical research

- Use triangulation of data
- Perform context specific research values (p. 76)
- Perform context specific interviews in combination with tools or techniques that reach latent levels of knowledge (provotypes and generative tools (p.76)



I Ethical inventiveness for value tension

- Value tensions related to fairness in AI are the ones discovered and selected (p. 88)
- Support the AI team in explicitly discussing value tension (p.72)
- Spark creativity with creative exercises (p.47)
- Create an understanding of the decision context (p.72)
- Do not use means as objectives (p.72)
- Aim to avoid exact fairness calculations at the beginning (p. 56)
- Actively stimulate discussion concerning the consequences of the AI system (p.72)
- Frame the decisions and make them explicit (p.72)
- Stimulate structuring of the problem (p.72)
- Simulate integration of stakeholders direct and indirect (p. 39)
- Stimulate asking questions about values to support resolving the value tension (p. 39)

- Dimistify value tension (p. 73)
- Decompose values (p. 73)
- Dimistify consequences (p. 73)
- Possibility to decentralize responsibility (p. 73)
- Create room for reflection (p. 41)
- Advance the values expressed in the AI system from accidental to purposeful (p.66)
- Advance the values expressed in the AI system from potential to performed (p.66)
- Strive for situational values, context specific fairness and perceived fairness (p.66)



I Ethical fulfillment in the AI process

- Integrated at current ethical decision moments in the development process, ideation (p.)
- Easily integrate ethical implementation made in earlier stages (p.82)
- Remind the ethical dimension of AI (p.80)
- Stimulate continuous reflection (p. 41)
- Stimulate well argued decision making (p.39)
- Stimulate discussion concerning the consequences (p. 39)
- Stimulate moral motivation further in the process (p. 40)

The guidelines for the overall design are presented at the next page

07 Preparing the ethical recipe

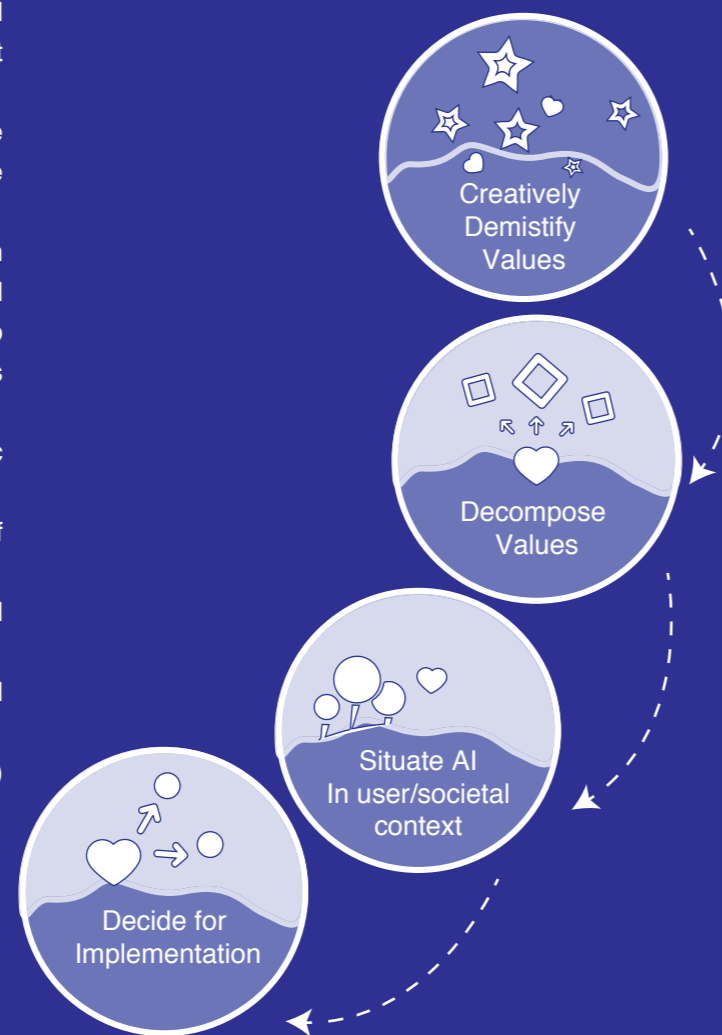
05 Guidelines for the overall design

- Early in the process (p. 39 & p.80)
- It should be facilitating and not prescribing (p. 86)
- A co-creative act with the entire team (p. 82)
- Ethical challenges do not always arise within AI development, thus there is a need for assessment beforehand (internal interviews)
- The AI team should be stimulated and provoked towards non technical thinking (p. 86)
- It should lighten the burden of the all the responsibility the data scientist has now (p. 82)
- It should stimulate reflection on the diverse design choices in iterative manners (p.41)
- It should explicitly mention responsibility and stimulate moral responsibility. Make clear when, who is responsible, or the company does not take responsibility (p.40)
- The design should stimulate intrinsic ethical motivation (p.40 & p.86)
- Reusable for a diverse range of projects (internal analyses)
- Fit in current processes (internal analyses)
- Easy understandable (internal analyses)
- Fun to use (not as a ethics burden) (p.37)

Framework ethical inventiveness for value tension in AI

- 01 Creatively demistify values
- 02 Decompose Values
- 03 Situate the AI system in user/societal context
- 04 Decide for implementation

Overall I continuous reflection & iteration is imperative
Distinction is made between different values human/moral values, technical values, societal ones.

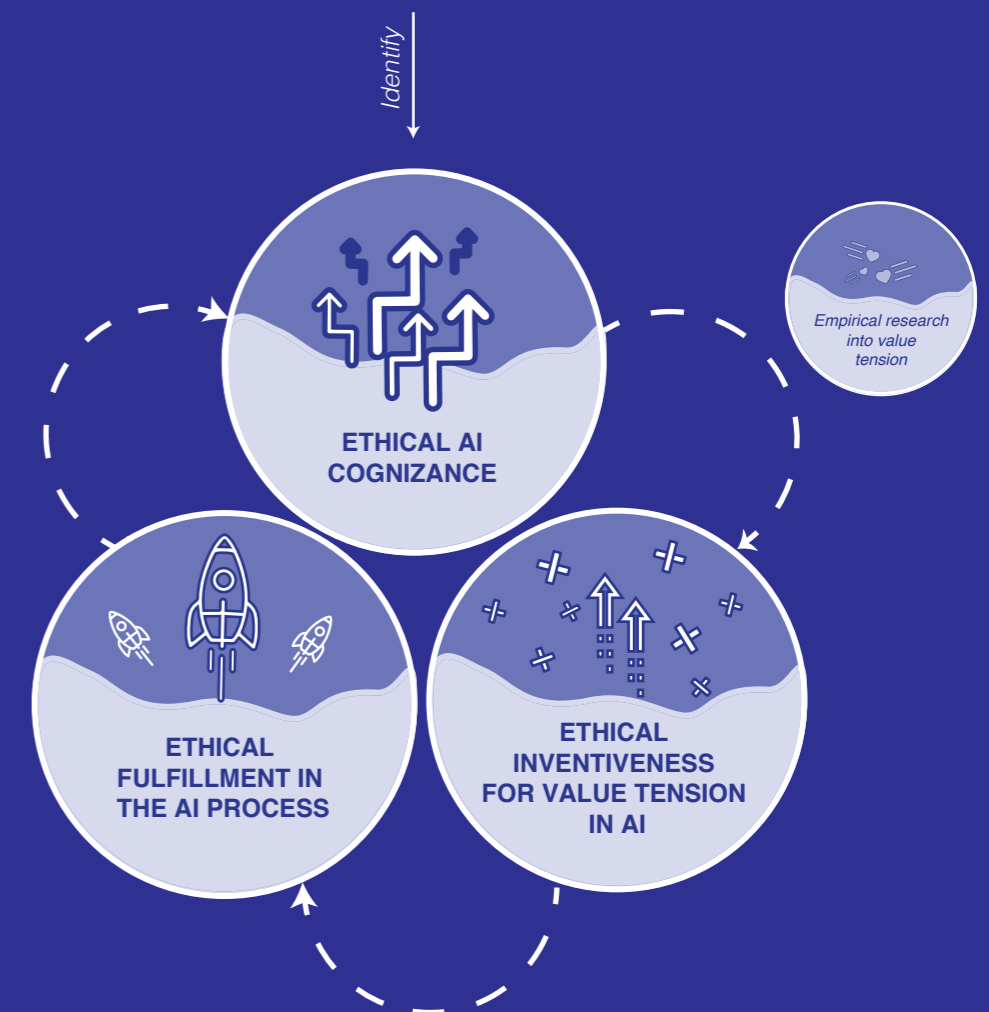


Vision for design consists of

- 01 Ethics as fuel of co-creative innovation
- 02 Prototypical kind of ethical thinking
- 03 Proactive stance in designing for fairness by reducing unfairness sources
- 04 Context matters
- 05 Increase moral motivation and reflexivity

The Framework I designing for fairness in AI

- 01 Identify
- 02 Ethical AI cognizance
- 03 Empirical research into value tension (optional)
- 04 Ethical inventiveness for value tension in AI
- 05 Ethical fulfillment in the AI process



Chapter 08 I

Designing for Fairness

This chapter elaborates on the design process of the organizational role with an accompanying modular toolkit. The framework explained in the previous chapter serves as a basis for the design. First, the overview of the ideation process is shared and the idea overview explained. Second, every step of the design is elaborated upon separately.

In this chapter

1. Iterative ideation
2. The Ethical AI coach
3. AI dish
4. Shape workshop

8.1 Iterative Ideation

This section elaborates on the ideation process which was performed iteratively. Two test session were conducted, once with designers and once with computer scientists (differing from master students to managers) during the course of ideation.

In figure 8.1 the iterative development process is visualized. A variety of ideation techniques was used to come up with suiting ideas and concepts with the use of the framework. Due to the complexity of the topic was chosen to not perform a creative session as ideation but rather use extra iterations and test moments with diverse groups of people. The first test has been performed with design students to get a designer perspective on the process and understanding of the different elements. Feedback was implemented to improve the toolkit and a new ideation round performed Then the concept was tested with two teams of (students) computer scientists. As computer scientist and data scientists have a very different way of thinking then designers, this session was performed to gather insights from the actual target group and similarly the feedback was implemented and ideated upon, leading to the ethical consultant starters pack.

8.1.1 Ethical Coach Starters pack

The concepts proposed to IBM Benelux are on

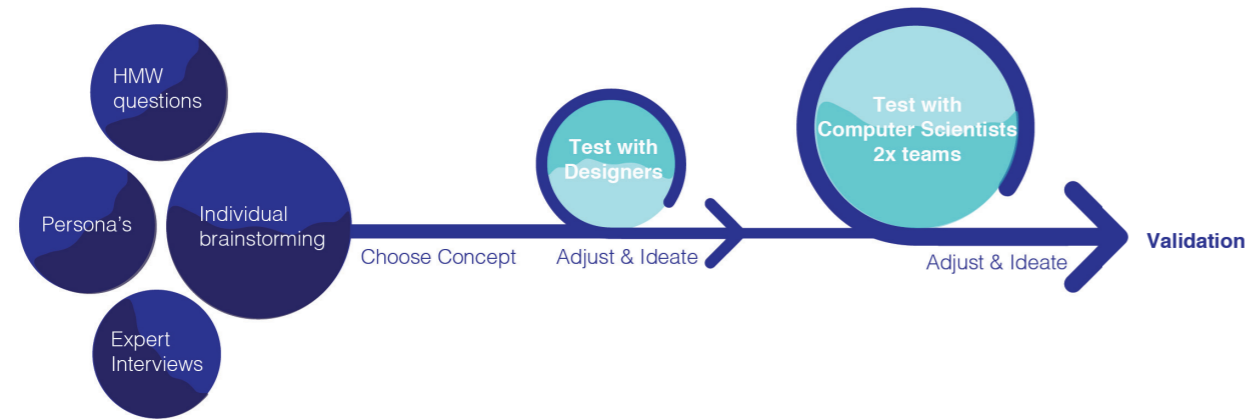


Figure 8.1 | The ideation and testing of the workshop design



Figure 8.2 | Overview of the idea set up with the goals (the ethical consultant starter pack)

8.2

The Ethical AI Coach

Goal I

Facilitating and co-creating a fairer and value-aligned AI system together with AI team.

Organizational level design concept

8.2.1 Need

The need for an ethical coach role in the AI development team is briefly discussed. First, from the interviews and prototypes a lack in ethical knowledge, thinking and reflection is discovered in the AI development process. Second, is discovered that the data scientists role has plentiful responsibilities, deadlines and this person experiences much pressure. It is not feasible neither desirable to give the data scientist, extra ethical responsibility fully.

Third, the current teams miss a societal perspective with still a basic understanding of the technology AI.

Fourth, most people in the AI development team have an engineering background, thinking in features and processes straight away. A real change of mindset is necessary with the integration of ethics in AI. It is necessary to implement considerations about implications on (in)direct stakeholders.

As the design vision describes in a creative new solution space. The definition of ethics used in this thesis is not prescriptive. It is meant to be a reflective mindset. In the case of the AI team a co-creative process towards creating fairer and value aligned AI systems.

8.2.2 Theoretical background

The literature review points out that a team member, with responsibility of explicitly brining ethics and values to the table during technology development process, has benefits for ethical results (Fisher and Mahajan, 2010; Manders-Huits and Zimmer, 2012; van Wynsberghe and Robbins, 2014; Shilton and Anderson, 2017). A values advocate is a team member translating values for technical work (Shilton, 2018). The current identified benefits are: (1) it brings deep knowledge of interdisciplinary literature of ethics. (2) it provides an outsider perspective and break group biases, creative thinking. (3) Incomplete understanding of the technology can bring up new questions and make developers thing of the technology and problem in a different way (Mun et al., 2014). (4) Lastly, value consciousness, an explicit responsibility in the design team, aids to build values reflection into the scope of work and the success metrics of a team (Shilton,2018). However, there are also downfalls. (1) First, it might be difficult to fight for a presence in the design team and also to convince others why it is important (Manders-Huits and Zimmer, 2012), legitimacy makes their job difficult. (2) Second, responsibility on a single person may put a stronger emphasis on putting his/her values in the design process, therefore ethical pluralism



Figure 8.3 | The ethical coach role with requirements and responsibilities

Role description | Ethical AI Coach

Requirements

- Design Background/Experience
- Ethical knowledge
- A basic understanding of AI systems
- Experience with creative facilitation

Responsibilities & Activities

- Creates ethical project strategy
- Oversees the ethical project strategy
- Assists projects in the need of ethical advice in AI at the important decision moments
- Implements the ethical starters pack tools at the right moments in the process
- Facilitates a more ethical AI development process
- Simulates two orders of reflection
- Creatively facilitates the team towards the creation of fairer AI systems.

is advised (Borning and Muller, 2012). (3) Third, in real life commercial setting it is not always feasible to hire an extra person full-time.

8.2.3 Role Design

Literature, internal conversations within IBM concerning organizational roles and the earlier described need fueled the design space for this role. It is intended to fit IBM's current systems and processes

Design based on literature

The previous section explained the theory behind a value advocate alongside the benefits and disadvantages. The proposed role of the ethical coach is in line with the value advocate role but slightly different. The ethical coach role is meant as a facilitating role towards more fair AI development. It is not the responsibility of the of the ethical coach to make a more fair AI, this is a collective responsibility of the whole AI team. An intent of this design is to enrich the ethical solution space with a design perspective (chapter 3). In line, the ethical coach should have a design background or experience. In this manner the coach can integrate design principles, co-create and use creative facilitation methods and tools towards a new ethical solution space. In line with the theoretical background, this role should

not be a role for an AI expert, rather someone who can ask exploratory questions outside its field and make developers think and explain in varied fashions. Additionally it is the role of the ethical coach to discuss values, value tensions explicitly in the process to improve the ethical reflection, awareness and discussion.

Fit IBM

IBM as a multinational and one of the global leaders in AI development has the resources, knowhow as well as scale to proceed with a new organizational role. The quote of A. Wynsberghe indicates that also for the benefit of all AI ethics, companies such as IBM should invest in creating the knowledge base how to make fairer AI systems.

Internal informal interviews within IBM were held to create a fitting concept and support for this concept for later development. The main insights are shared. First, the name coach is proposed as a attractive name (i.s.o. consultant). In line this would fit the development of the agile coach role.

Second, IBM has many education and badging programs employees can earn and learn from. It is advised to create the role in similar fashion as the current roles are, such as the Agile Coach. This would support a fitting implementation in

the company.

Currently the badges for an agile coach are attained through a variety of online and offline trainings and courses. They start ranging from awareness towards in depth coaching roles. The role is designed in line to train in skills and knowledge as currently is done at IBM.

Third, for a coach soft skills are highly important. Not all employees even with the right knowledge are a suiting coach. Discussions have been held with the CIO at IBM Benelux (who interviews and selects the people to become an Agile coach) concerning the skills an Ethical AI Coach should have and incorporated in the final design.

“We need ethicists working in the companies that can afford them as part of the design team, where they can start to uncover the common issues other companies are running into”

- Aimee Van Wynsberghe, Founder Responsible Robotics Foundation (2019, Forbes interview)

The expert (ethics)

The creative facilitator

The outsider (fresh perspective, asking questions sparking reflective character)

IBM

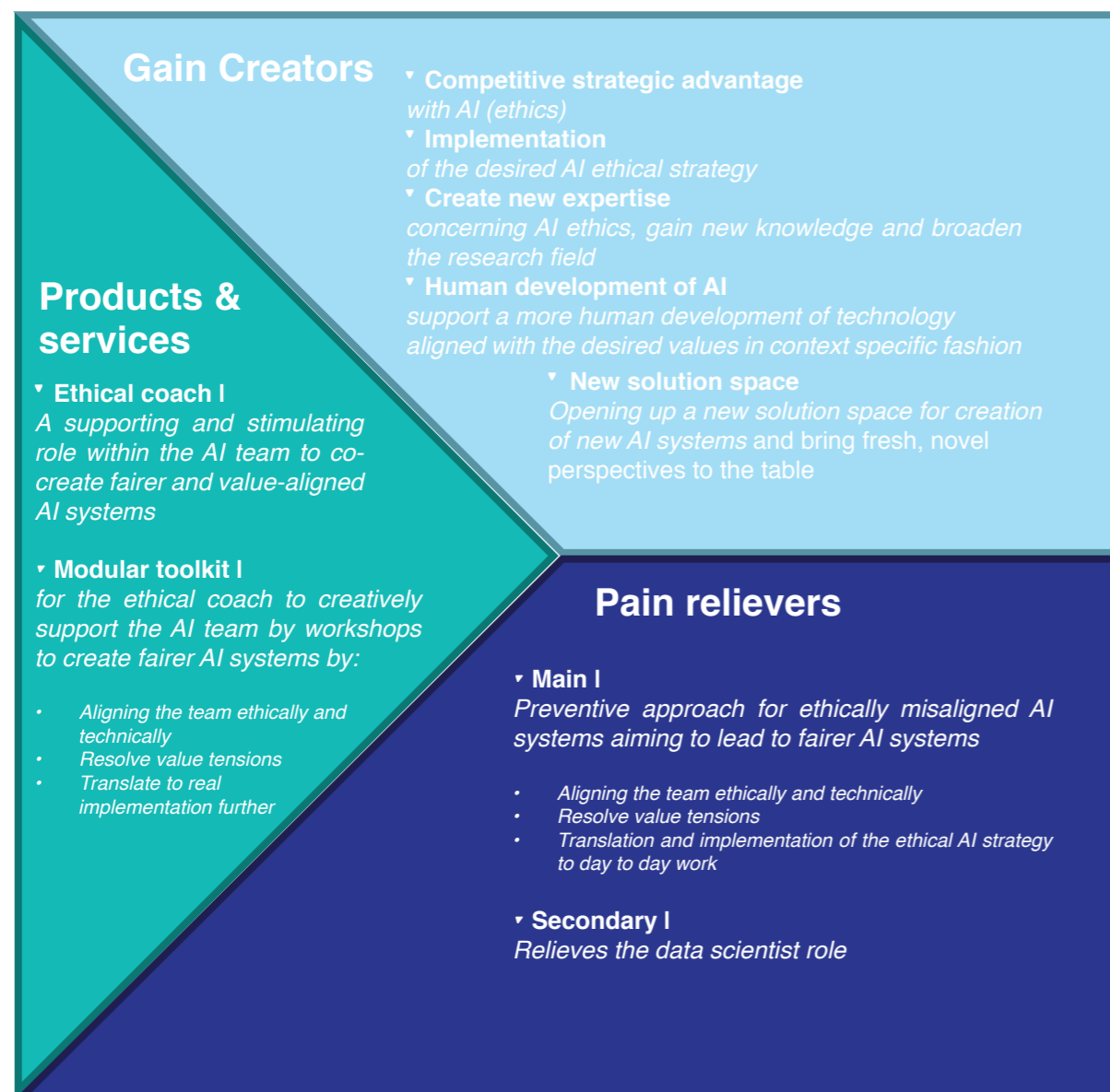


Figure 8.4 | Value proposition Ethical Coach with Starters pack for IBM

ORGANIZATION CLIENT

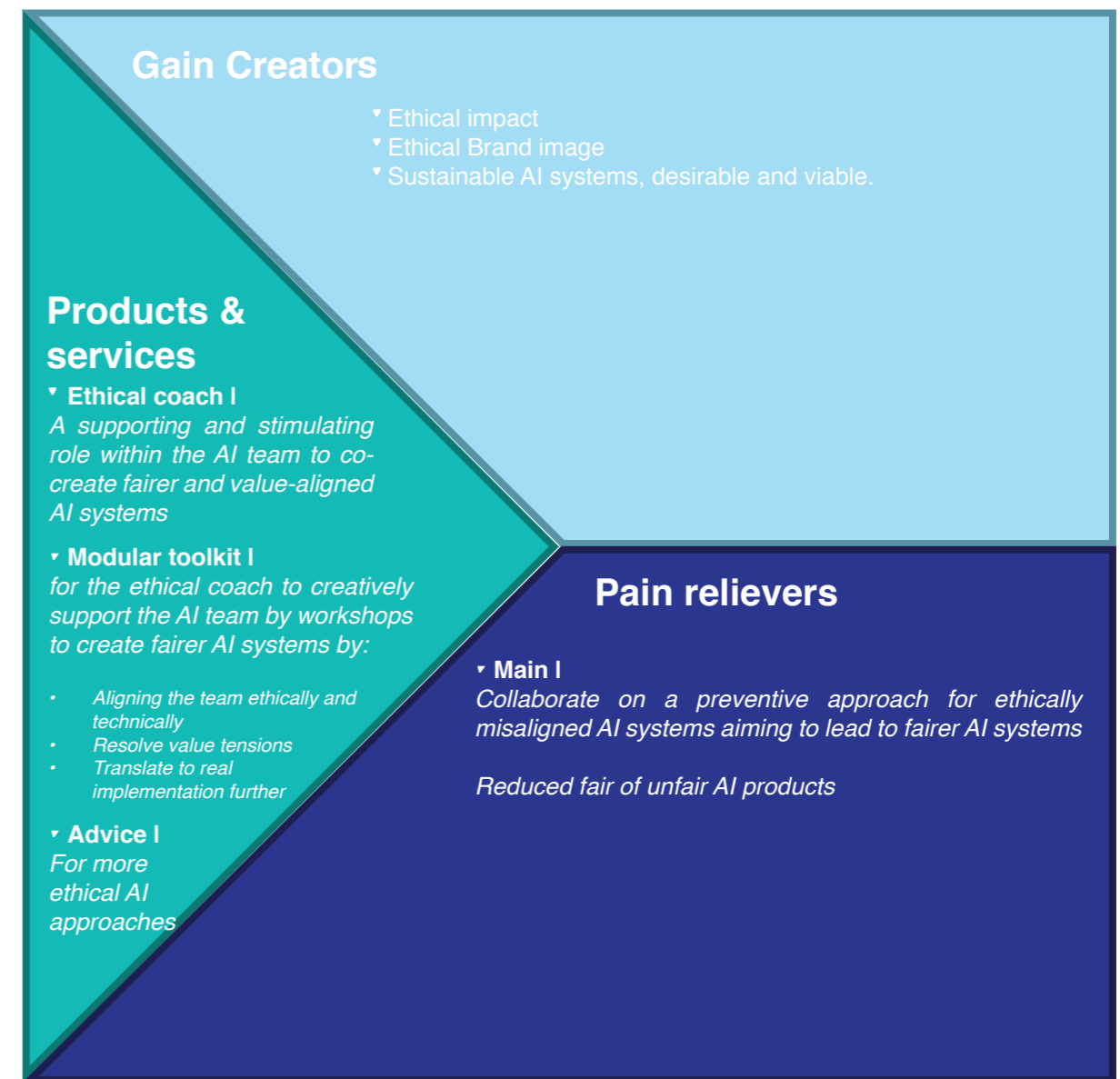


Figure 8.5 | Value proposition Ethical Coach with Starters pack for clients

Value proposition

The value proposition canvas is used to indicate the gains for both IBM and their clients shown in figure 8.4. and figure 8.5. This supported the development of the role and further recommendations.

“..people like new ways of working, make it attractive and people are more likely want it”

- Sophie Kuijt, Ethics Ambassador IBM Benelux

8.3

AI Dish

Goals I

Awareness, Alignment team, Basis for reflection

Implementation level design concept

8.3.1 Need

The interviews, prototypes and literature identified a lack in: (1) understanding of the whole team of the technology and why certain choices, in the AI development process have an impact on the ethical consequences. (2) awareness of the ethical implications decisions might have later in the process or in the implementation. (3) a clear basis as an overview of the system.

8.3.2 Theoretical background

01 Relatable metaphor

A metaphor I “.. to mean a linguistic, visual, or auditory construct in which one thing (the referrer or source) refers to another (the subject or target)” - (Saffer, 2005)

In other words, it is a mechanism to look and talk about something in terms of something else. Metaphors, when used properly, are an influential tool for designers. They can add benefit in multiple manners, in the develop process for ideation, and within the design itself. Metaphors support humans to understand complex and abstract topics, by referring to something more concrete (e.g. time is money) (Saffer, 2005). The AI development process is quite abstract and difficult to grasp for non-experts in the field. Thus, a relatable metaphor of the AI dish is chosen to easily communicate the process and impact of the consequences (in chapter 1

an explanation is given on the metaphor). For example, when one cooks, but the ingredients are of poor quality, the dish will never be of outstanding quality. Similarly, in AI development, when the data is of poor quality the model will not be of good quality. This metaphor is used in this first exercise to align the whole team on basic AI knowledge.

02 Describing choices

STIR (Socio-Technical integration research) is an ethical research tool/method. It intends to bring to light decisions about opportunities, technical considerations, alternatives and outcomes in engineering processes (Fisher, 2007). At heart of this approach is to ask designers to describe their decisions but not changing them. This increases reflexivity about what the team members decide. This is currently lacking in the ideation This currently lacks in the ideation phase of the AI development process. Decisions such as, which algorithm to choose (appliances) and which types of learning (recipe) are not communicated across the team with the advantages but also importantly the disadvantage. Most of the time is not thought in a reflexive manner about these technical decisions but from a technological standpoint (how fast, how accurate etc). Thus, in the AI dish, the team is supported in describing these decisions, without necessarily changing them, to increase their reflexivity.

04 Stakeholders

Ethical tools analyzed for this thesis, highlighted the importance of integration of stakeholders, both direct and indirect in the process (Friedman et al., 2013; Friedman & Hendry, 2012; Goodpaster, 1991; Guston and Sarewtiz, 2002; Mephram, 1994; Mephram & Kaiser et al., 2006; Miller et al. 2007). Concluding from the generative tool, there is a lack of thinking and integration about them in the current AI development. Thus, in the AI dish a discussion about the stakeholders both direct and indirect is facilitated.

05 Early in the process

Ethical tools analyzed for this thesis, highlighted the importance to use ethical tools early in the process for effective (Friedman et al., 2013; Miller et al. 2007; Van den Hoven, Vermaas, & Van de Poel, 2015). From the generative tool also appeared that many decisions relevant for the project later are made in the ideation phase. Thus this tool is designed for the ideation phase.

06 The interaction & implications

Thinking and discussing about possible implications of systems is seen as a strategy to resolve value tensions. In order to make the AI team think the implications early in the process a first step to stimulate this is integrated in the AI dish.

8.3.3 The AI Dish Design

The AI Dish design is derived from the framework part: Ethical AI Cognizance. This sheet aims to align the team on both technological knowledge, and stimulate to discussing the choices. The AI dish is a relatable and playful manner to discuss the new AI technology, think about implications, the choices, interaction in an understandable fashion for the whole AI team in a form of a canvas. The AI dish canvas is part of a bigger workshop in which the whole AI team participates or can be used separately to align the team. It is facilitated by the ethical coach. Once filled in, the AI dish provides an easy overview of all the components of the AI system. It provides

a basis for discussion and reflection both later in the workshop as later in the process. The first iteration is presented in appendix N.



Figure 8.6 session 2 AI dish

8.4 Shape workshop

Goal I

Resolve value-tension in AI development creatively with concrete implementation ideas

Implementation level concept

8.4.1 Need

Currently value tensions are not addressed and there is no practical support for AI teams to explicitly solve value tensions. Not addressing value tension in an explicit way can lead to a lack of appropriation by disadvantaged groups, system sabotage (Flanagan et al. 2005) or ethical misaligned AI systems. In this research value tensions in AI are identified (Chapter 6). Thus a workshop setting is designed, aiming to support the AI team in explicitly resolving these. The following value tensions are addressed of which accuracy vs probity workshop is tested.

- Individual benefit vs collective benefit
- (Historical) data value vs Socially desired value
- Explainability vs performance
- Freedom/privacy vs safety/control
- **Accuracy vs Probity**

The full overview of the workshop ideas is visualized in appendix P.

8.4.2 Theoretical background

In the literature review concerning value tension distilled towards six main strategies to resolve value tension (p. 72). These are: (1) untangle value (tensions) (Van den Hoven, Vermaas, & Van de Poel, 2015, p 838). (2) decompose values (Miller et al. 2007) (3) avoid problematic features for stakeholders (Miller et al. 2007) (4) Decentralize responsibility (Thacher, 2004) (5) Quantify values & consequences (Van den

Hoven et al., 2015); (6) Untangle consequences (Friedman and Hendry 2012). These together with the identified value tensions and strategies from the interviews and provotypes fuel the ideation around this workshop. Together with the vision of and the framework presented in chapter 7, the workshop setting is ideated upon and created in figure 8.7.

8.4.3 The shape workshop design

A workshop setting is chosen as the most fitting design matching with the earlier mentioned design requirements.

(1) First it is important that the AI team themselves realize the ethical implications and create new solutions. In this manner one increases the intrinsic moral motivation and commitment. (2) Second, the entire AI team needs to be present to have the different perspectives and backgrounds fueling the (ethical) ideation process (business owner, data scientists, IT and other (core) stakeholders). Also for input about certain restrictions and other types of knowledge of their own specialty the presence is valuable. During the design of the session, the target group was always in mind. More technology oriented people need more guidance in the sessions as well as in creative stimuli to change perspectives. This is taken into account in the choices made for the different phases of this workshop.

The ethical coach is facilitating the creativity of the participants of the session. It is crucial that

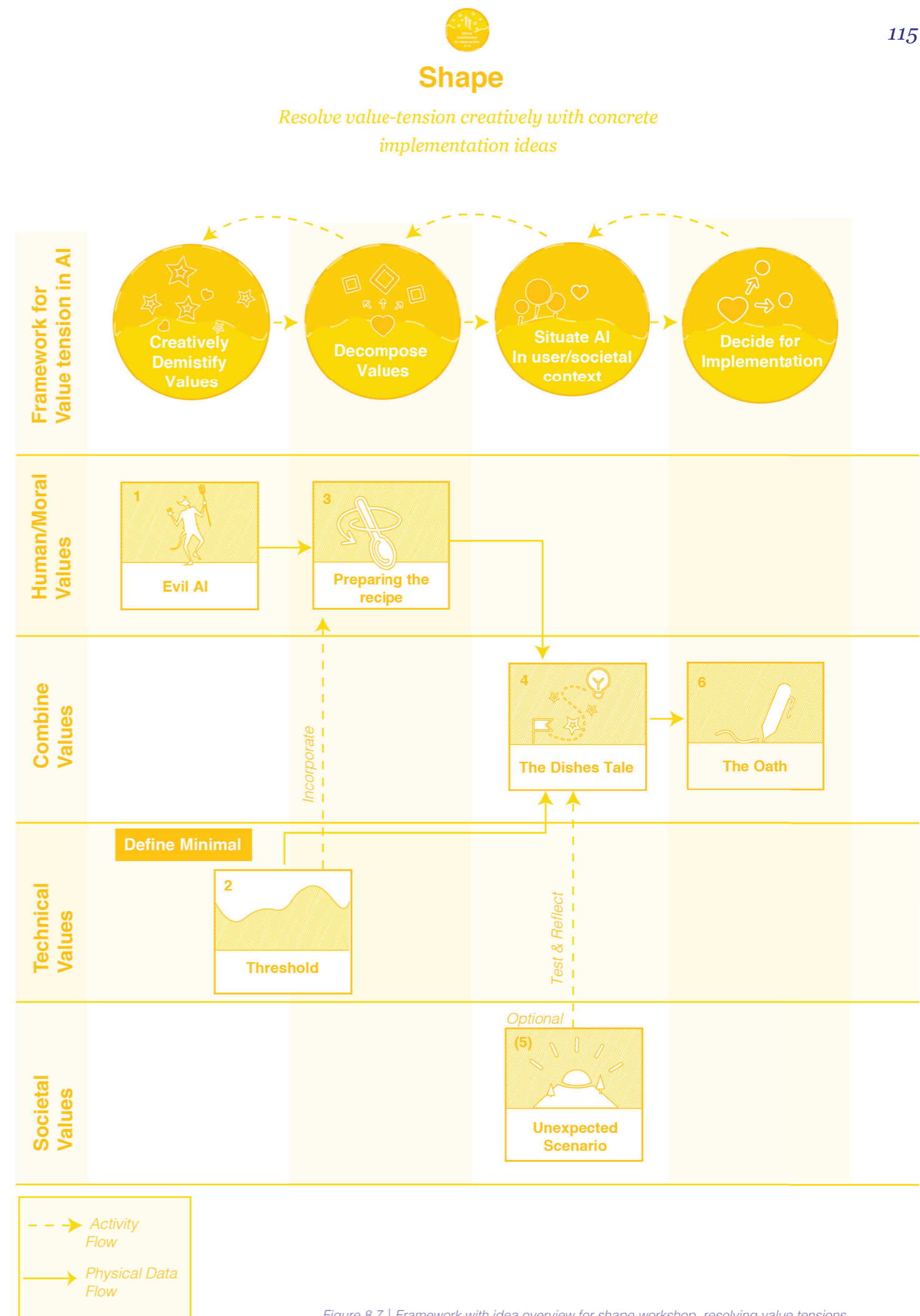


Figure 8.7 | Framework with idea overview for shape workshop, resolving value tensions

the ethical coach stimulates the target group with asking the right questions, energizers and ice-breakers to lead toward desired results.

01 Different levels of values

The chosen value tensions are not on the same level. For example probity and accuracy are two different type of values, whereas individual benefit and collective benefit are on the same level. Through the design process it was noticed (test session 1 with designers), when the two values are not on the same level they call for a different type of decomposition and resolving. For the current workshop the value tension accuracy vs probity was chosen as it is connected to multiple sources of unfairness. Also from the interviews and the provotypes appeared no attention is given to this one while it is a essential one. The value tensions are visualized in appendix P. As not all values of the value tensions are on the same level, the processes of the workshop change accordingly.

8.4.4 The framework for resolving value tension in AI

The insights gathered concerning value tension and resolving in AI are consolidated into small framework (by means of generative interviews, provotypes, expert interviews and literature) (chapter 7). The goal of the framework is to design tools to resolve the identified value tensions. The framework is open for reuse for the design for support for AI teams. Based on this framework the workshop design is created.

“It is sad truth that most evil is done by people who never make up their minds to be good or evil”

- Hannah Arend, *Philosopher political theory*



Figure 8.8 | Session 1 value tension

Although the workshop is modular in nature the steps of the full workshop spelled out. The steps with the proposed design are briefly discussed in the next paragraphs.

01 Creatively dimisitfy value

Evil AI & Evil stimuli

This step explores the ethical problem in extreme manners (linked to the ethical process). In order to spark the imagination in more playful and fun this extreme exercise is chosen. The participants are asked to think of the most unfair system in their use case. An important aspect of this exercise is to discuss why this idea is unfair. It distills the unfairness sources the team can think of. For example, in the case of probity, the participants are asked to think of the most unfair, immoral and prejudiced system and how to make the AI system like this. Additionally it stimulates thinking about the (in/direct) stakeholders in a playful manner. This

exercise inspires to think from the perspective of the user as the team starts to think of unfair systems for themselves. Currently this lacks in the AI development process and the evil canvas with the evil stimuli provide a starting point for that.

To stimulate the participants, evil cards are made to spark their evil side (especially with more technical oriented people this was necessary). Multiple iterations have been made (appendix N) of which the final one is described. Research has been performed into the what values are perceived as undesired in the European context. This led to the use of fables.

A fable is “a short story that tells a moral truth, often using animals as characters” (Cambridge dictionary).

These short stories are used to teach people (often kids) about right and wrong, what is morally desirable or is not (in the European context). The Fables de La Fontaine, even though written in 1679, continue to have impact due to the imaginary power and still current relevance of their content (figure 8.9). Thus, inspired by the exhibition of La Fontaine by Rob Scholten in the Hague, the negative emotions and human characteristics are extracted and translated into an evil card deck (appendix N). This is used to fuel the stimuli to think of inappropriate ideas for AI systems in the EU context.

The outcome of this exercise are evil AI system ideas with the sources of unfairness.

Threshold

When two values are not on the same level, also different manners are needed to explicitly discuss them. In the case of probity and accuracy is a technical value of the AI system and expressible in a number. Thus to discuss the acceptable accuracy level a different exercise is created in which the value of accuracy is demystified. The components that boost or hinder accuracy are discussed so the understanding of the entire team is bridged. This exercise explicitly discusses the threshold of accuracy that is acceptable, keeping in mind already the existing systems it



Figure 8.9 | Example fable the Fables de La Fontaine



Figure 8.10 | Session 2 value tension

needs to be integrated and translating this into requirements that can be implemented in the system (already partially decomposing the value of accuracy). Thus the outcome of this sheet is a list of requirements for the minimal acceptable accuracy.

02 Decompose values

Preparing the recipe *(propose solutions)*

This sheet turns the reasons of the evil ideas into manners how to prevent it. It guides the team into decomposing the values from abstract to concrete to implementable procedures and features inspired on the evil ideas the team had. This type of decomposition is partially inspired by the Design for Values ICT (Van den Hoven, Vermaas, & Van de Poel, 2015, p 838). The metaphor of the AI dish is used in all the steps of the process to keep the workshop understandable for the whole team and to give it a playful touch. The outcome of this sheet are implementable features or procedures to prevent unfairness of the system.

03 Situate AI in user & societal context

The Dishes Tale

This design focuses on the story line of the AI system in context it will perform in. It forces the team to discuss about the interactions with the system, which outputs it will need to have and the priorities. The decomposed values are implemented in the service proposition. In this manner consequences are seen in a tangible manner, which is currently lacking. Also, the interaction layer with the system gets a prominent place, early in the process. (Also, the human perspective and human to AI System perspective is injected at the software design level (Bitner, Ostrom, & Morgan, 2008).) User stories are chosen as inspiration to familiarize the IBM employees with the approach. It also prioritizes the different features that will need to be tested and implemented at first. It serves also as a foundation for reflection later on the process, providing the overview of the interactions happening. The output is the story line of the AI

system in context with the necessary features, hierarchically distributed.

Unexpected scenario

This step is performed for validation and reflection (related to the ethical process) on the proposed AI system. From the empirical research appeared that much surprises occur during the AI development process. This can lead to less ethical AI systems in the actual societal context. This exercise aims to prepare the team for some unexpected scenarios that occur more often and have implications on the fairness of the model. Inspired by the unfairness sources that are identified in the literature review, unexpected scenario cards are made. The participants can pick a card blindly. This card is put on the surprise canvas. Then questions are asked about the implications it would have on the proposed model as a form of reflection. Then is asked what can be changes to prevent negative consequences of the surprise or prevent the surprise. It supports the team in making the model more robust as well as leading to a model which is better resistant for sources of unfairness. The output of this sheet is an improved Dishes Tale and AI Dish as input for the AI development process.

04 Decide for implementation

The oath

This exercise aims to stimulate moral responsibility as well as form a start of the agreement of implementation towards a fair AI system with the entire team. It bridges both the procedures and requirements of both probity and accuracy, integrating it with the AI principles of IBM. This is a playful manner to close of the workshop while having made agreements concerning implementation instead of solely rich discussions. The output of this sheet is an morally binding agreement of the design of the AI system. It is not legally binding and aimed at stimulating argumentation for the changes of choices later in the actual process.

05 Reflection and Evaluation

Overall it is meant that the AI Dish and the Dishes Tale provide the projects overview and are reflected upon further in the process. For example during stand-up meetings that the AI team has. Already during the two test sessions this was really noticeable however the ethical coach need to stimulate the reflection with questions. Agreements about the features of the AI system that are made serve as the basis for the actual development process further towards a more fair AI system.



8.4.5 Role of the ethical coach

The ethical coach has an important role in this workshop. Both in stimulating the creativity of the team as well as asking reflective questions during the process. Also inspiring the discussion from a more societal and end user perspective is the task of the ethical consultant, supporting to think of the consequences of decisions and the (in)direct stakeholders. In the final workshop design a the guide for the workshop is visualized.

“I really like the surprise and the reflective act in it. It makes you identify the gaps and blind spots and make the system more robust. Also, that you went through the process and then need to go back in an iterative manner is really nice such as in real life. And if people do not want to go back to half an hour ago this will happen in real life but then with weeks or months. It might be nice to have everything on a wall and then you can make it an iterative process.”

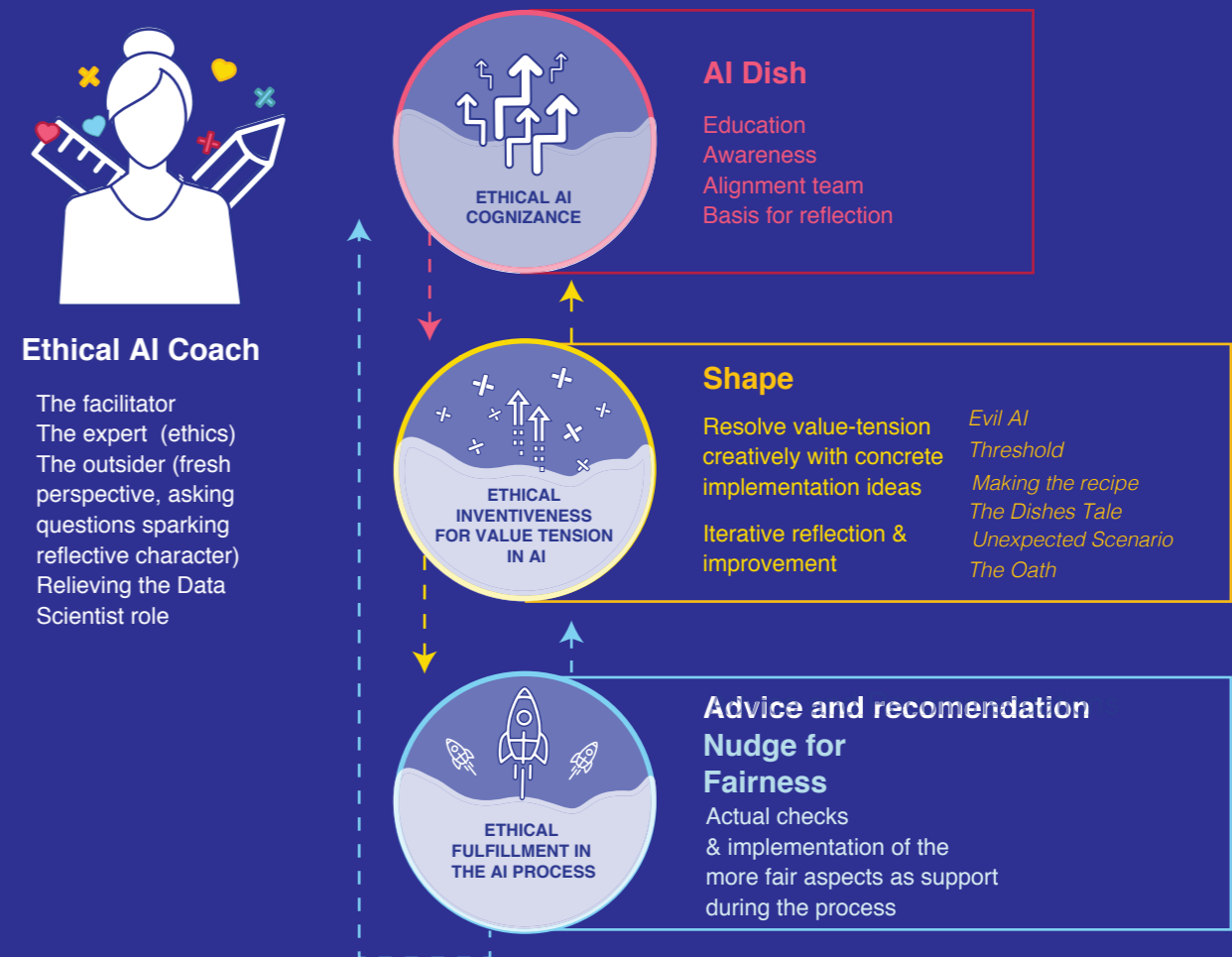
- second iteration (computer scientist) about the unexpected scenario

“The AI dish, works really well and really appeals, it works immediately”

- first iteration (designer) about the AI dish

08 Designing for Fairness

The Framework with design elements



Different levels of values

Due to the different types of values represented in the value tensions there is a need a different approach of decomposing and addressing these.

Chapter 09 I

The ethical coach starters pack

This chapter shares the final design outcome of this thesis using the ideas from the previous chapter as fuel. The final design is the ethical coach role with an accompanying modular toolkit. Additionally it describes the validations of it which lead to the recommendations in the next chapter.

In this chapter

- 9.1 The ethical coach with a modular workshop
- 9.2 Design validation

9.1 Ethical coach with a modular workshop

This section elaborates on the final design of the ethical AI coach and the tools designed for this role. This design is a consolidation of the ideation and validation phases of this thesis. The following parts describe the design gradually.

9.1.1 The Ethical AI Coach

The final concept is a proposition of a new organizational role, the ethical coach. For this role a set of tools to support the AI development team in explicitly resolving value tensions related to fairness is developed.

The ethical AI coach is a role described on the next page with the demonstrated skills, traits, knowledge, activities and guiding principles. The type of description is created based on other organizational roles such as the agile coach, within IBM, thereby it fits the current ways the organizational roles are described and implemented.

In figure 9.2 the new ethical AI team composition is visualized. The ethical AI coach is actively present at the ideation stage. Further in the process the coach supports the team at ethical decision moments such as decisive scrum meetings.

The next paragraphs briefly explain the activities step by step in the process.

Activities per phase



At the initiation of the project:

- explore the necessity of the ethical coach in the specific project
- ask the right questions concerning the project for assesment to propose a
- tailor the workshop and plan towards the teams needs

- In specific cases empirical research into context specific dimensions of values might be needed for the projects context (new industries)



At the ideation phase:

- Creatively facilitate “The AI dish” - supporting for alignment on ethical implications and technical specifications
- Creatively facilitate “The shape workshop” or parts of it - supporting in explicitly resolving value tensions
- Extract elements for implementation from the workshop session
- Stimulate reflection 1st and 2nd order



At the feature engineering phase, modeling phase and pilot phase:

- Challenge the process and project from an ethical perspective
- Apply the knowledge and decisions made in the ideation phase towards easily implementable features
- Guide ethical implementation
- Stimulate reflection 1st and 2nd order

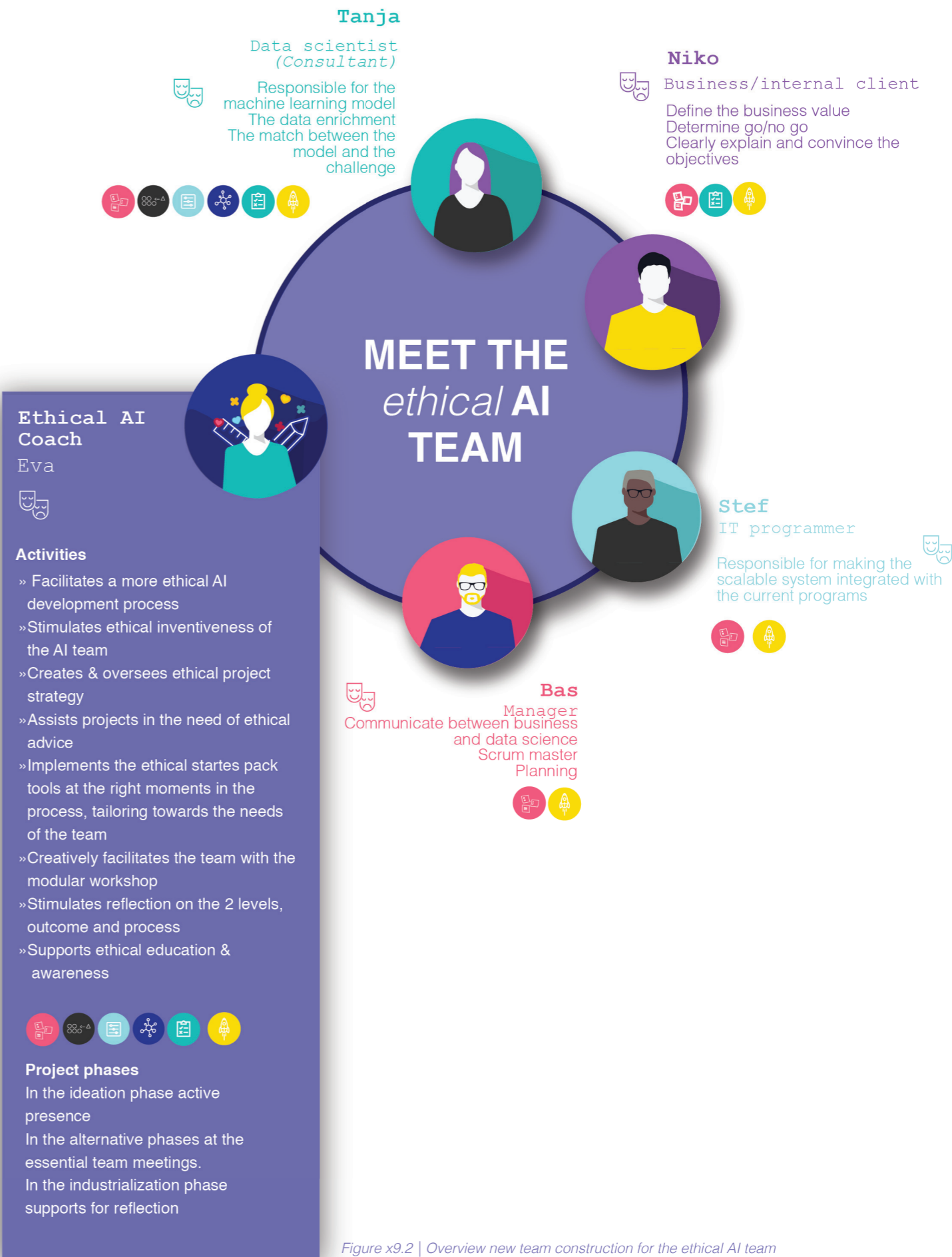
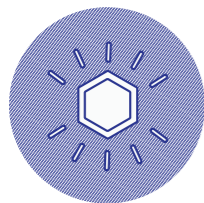


Figure x9.2 | Overview new team construction for the ethical AI team



Principles that guide the ethical coach



Ethics is a creative source of innovation



Reflection is key to learning



Strive for co-creation of fair AI systems

Role description

The ethical coach supports AI teams in the creation of more ethically aligned AI systems, according to IBM values and principles in a context specific fashion. Together they co-create fairer AI systems using their ethics and design background. For this they use the AI dish, shape workshop and tailored implementation of actions into the team processes. The coach challenges the team with the right questions towards fairer AI systems.

Responsibilities

- Make ethics a priority in AI projects
- Make the AI team aware of the necessity of ethical considerations and sources of unfairness in AI
- Creatively facilitate the AI team in the ethical reflection process
- Explicitly address value tensions and the consequences towards stakeholders in the process.
- Tailor the appropriate workshop formats as well as implementation

- Know the (IBM) AI ethics tools & developments
- Implement the ethical learning of AI projects into new ones and within IBM

Not responsible for

- The outcome of the project, this is a collaborative responsibility

Demonstrated skills

Team | Collaboration | Design thinking | Creative facilitation | Coaching

Traits

Passionate | Energizing | Creative | Empathic | Learning & Growth mindset | Empowering.

Knowledge

Ethics | Design | Coaching | AI (basics)

9.1.2 Starters Pack Design

The overall tool-set of the Ethical AI Coach is visualized in figure 9.1. The colors of the tools refer to the framework steps described in chapter 7. The tools are clustered per goal and per phases the tools should be used. This provides a concise overview for the coach. This toolkit is modular in its nature to fit the dynamic and constantly changing project backgrounds in the AI field. All canvases can be used together but as it is a modular toolkit some of them also can be used separately for different use cases. The intention is to make a workshop tailored for the specific projects, leading to suited support for the AI teams. The workshop canvases are designed to be facilitated by the ethical coach. The canvases are briefly addressed in the sequential order of a workshop comprehensively focused on value tension. **The entire workshop is presented in the accompanying file with this thesis.**

9.1.3 AI Dish Design



Accomplish preferably in the ideation stage
+- 1,5-2 hour
With the entire AI team

The AI dish canvas is a relatable and playful manner to discuss the new AI technology, think about implications, choices, interaction in an understandable fashion for the entire AI team. The metaphor of the AI system as a dish is used (p.32) The team will discuss the components of the new AI system and write the outcomes down on post-it's at the related sections, so the team can change the content later. As a facilitator questions asked concerning the arguments of the choices is essential. If difficulties arise guiding the group with techniques of creative facilitation is advised. Necessary is to explicitly mention the dishes ingredients are fluid, flexible for change later in the process. The output of the exercise is a filled in AI Dish canvas with initial ideas, propositions and discussions of the AI

system content. This prompts alignment of the team on understanding the decisions concerning the technology and stimulates argumentation for these decisions and the first realization of ethical consequences.

9.1.4 Shape workshop design

The following steps guide the reader in one profound combination of canvases advised to follow for a complete integration of value perspectives (see chapter 7). This workshop describes the steps of resolving the tension of probity and accuracy in AI system creation. Further necessities for the workshop are in detail described in the final workshop file (i.e. post its).



Accomplish preferably in the ideation stage
With the entire AI team

Evil AI & Evil stimuli - + 1 hour

This canvas explores the ethical situation in an extreme manner. It aims to spark imagination for more creative outcomes and trigger a change of perspective. The participants are asked to think of the worst, most evil ideas possible for the new AI system. The participants need to think of the most unfair, immoral and prejudiced systems and to write these down on post-it's. To stimulate participants, evil cards kindle their evil nature and imagination (inspired by negative values in the EU (p.132)(figure 9.3). The evil stimuli make use of pictures and HMW-questions to provoke ideas and emotions from the users side.

In this phase it is essential to guide the participants towards evil ideas and support them when difficulties are experienced. The second step is why these ideas are unfair. It is the role of the ethical coach to support the team in the categorization of these ideas in relation to data, context etc. The output of this sheet are evil ideas with a categorization on the evil motive on post-its. This canvas also can be used solely to stimulate the first ideas and reflection concerning unethical AI. This is already used by IBM during a presentation to stimulate discussion.

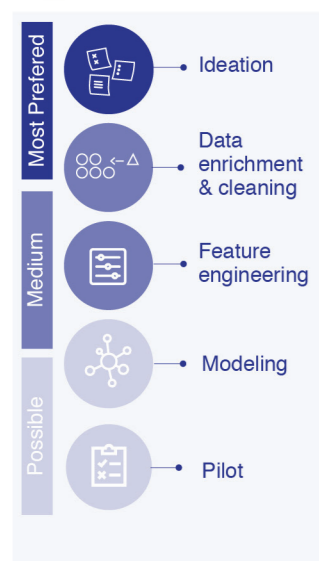


Ethical AI Coach Starters Pack

Modular tools for the ethical coach to support the AI team divided per goal and process stage

CANVASES | PHASE | GOAL

Legend



Accomplish preferably in the ideation stage
With the entire AI team

Threshold - 30 min - 1 hour

This exercise aims to create a minimal threshold for accuracy. Firstly, by writing down ideas what would be very unacceptable in terms of accuracy in this project and boiling it down to the causing rationale. Secondly, writing down which aspects would really increase accuracy of the system. Thirdly, a discussion concerning what would be the line of acceptability, should be stimulated by the facilitator. At the same time it is good to keep in mind the functional constraints and existing systems in which it might be implemented and discuss if these will be impacted by the accuracy levels. The outcome of this sheet are the accuracy requirements written on post-it's.



Accomplish preferably in the ideation stage
With the entire AI team

Making the recipe - 1 hour

In this exercise the earlier evil ideas are translated into manners to prevent this unfairness. Firstly, put the categorized post-it's from the evil sheet on this one. Then the team should be sparked by ideas how to prevent this evil. Secondly, categorization of the ideas is made ranging from more abstract ideas towards concrete implementation ones using the dish metaphor. It is essential generate concrete implementation ideas, supported by the entire team. The facilitator can support the team by guiding questions. The output of this exercise is a set of implementation ideas and principles for the project to prevent unfairness, written down on post-its.

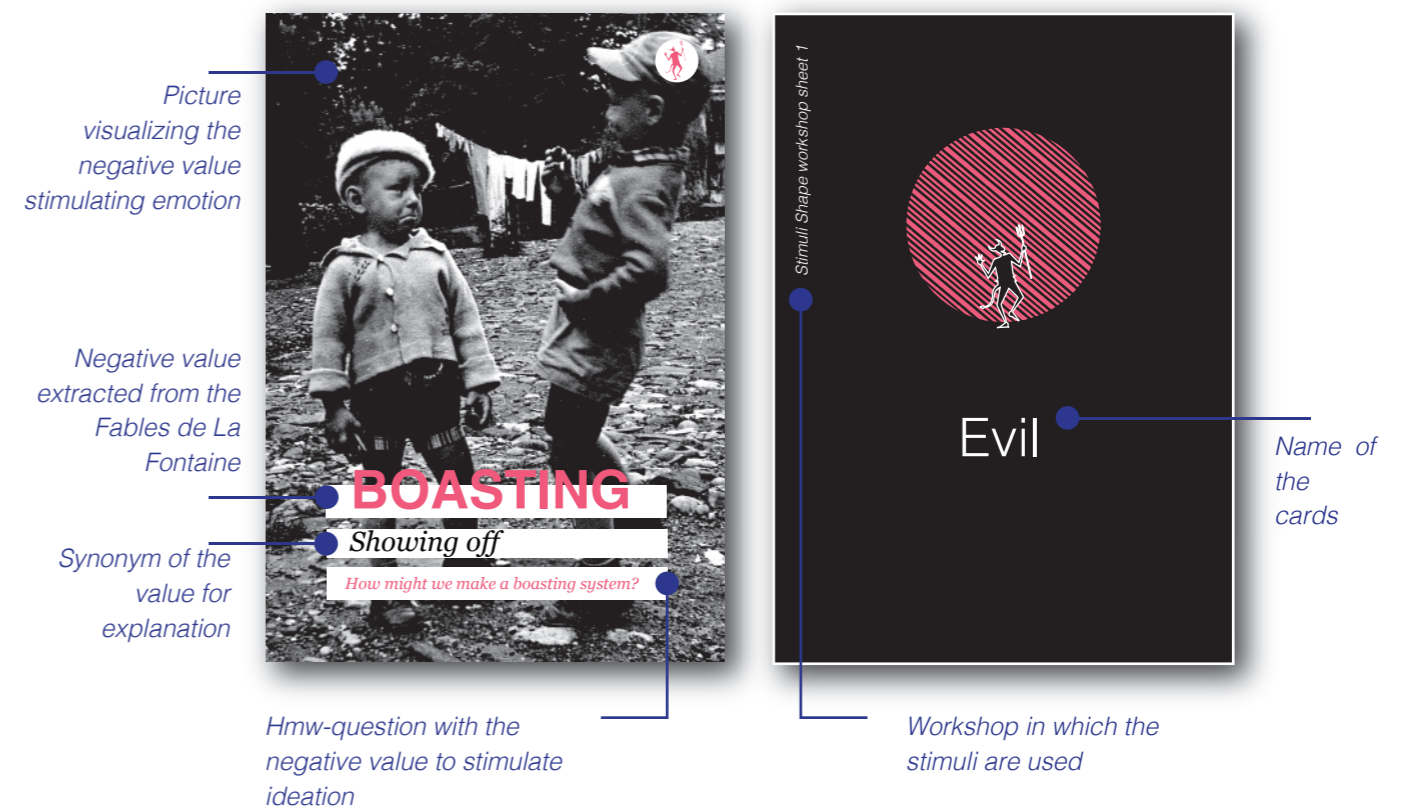


Figure 9.3 | Evil stimuli example and design explanation

Figure 9.1 | Overview modular workshop

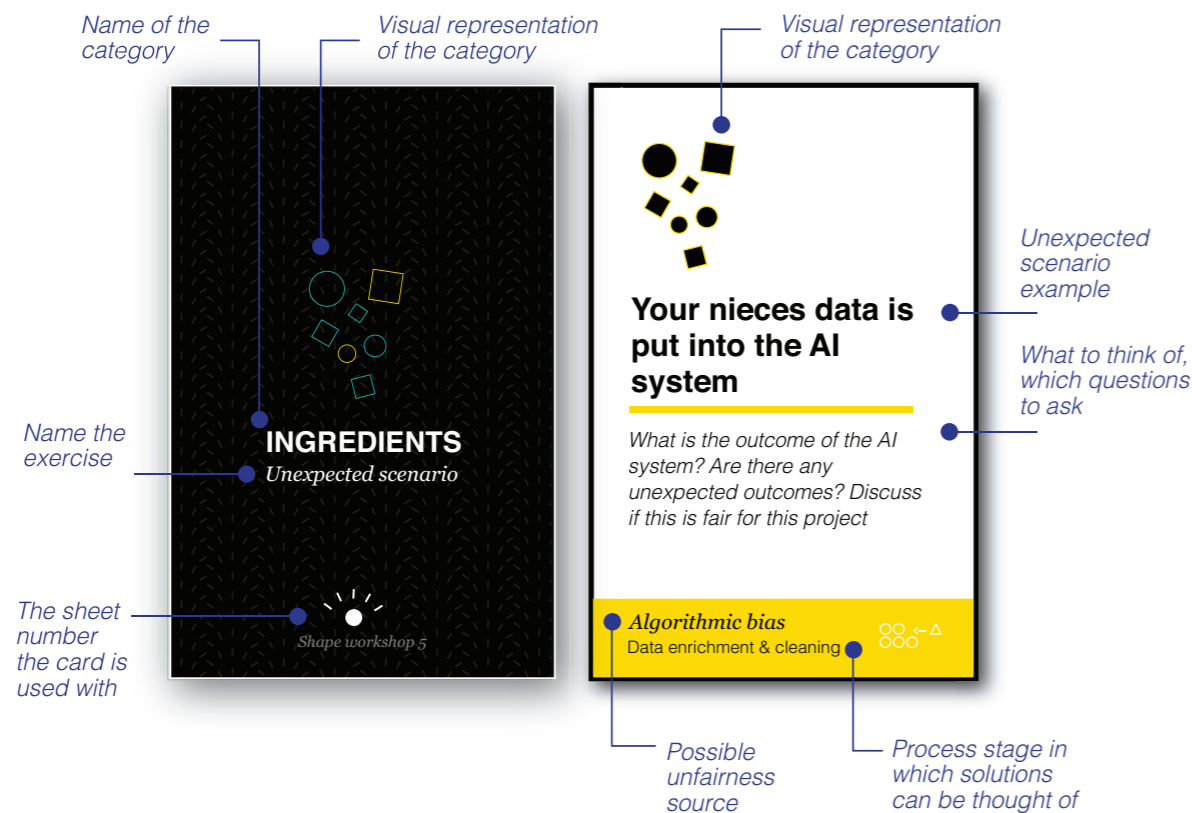


Figure 9.4 | Unexpected scenario card example with explanation



Accomplish preferably in the ideation stage
With the entire AI team

The Dishes Tale - 1/1,5 hour

This exercise aims to create the initial AI system story along the fundamental features for a fairer one. Firstly the decided upon post-its are placed from the “Preparing the Recipe” canvas and the “Threshold” one. In case the ideas are conflicting the facilitator stimulate the discussion with the team and ask questions to discover the rationale. Secondly, the story of the AI system will be constructed in the actual context it aims to be operating in. This is gradually consummated by starting with describing the context, then characters, touch points, outputs of the system, the AI system behavior, and the actual features/ characteristics. The facilitator may encourage the team to back up their decision with solid

argumentation. When needed the coach can remind to consider the indirect stakeholders and the consequences for society. The actual agreed upon features are hierarchically allocated based on priority. The output of this canvas is an initial story line of the use of the AI system with a prioritized list of implementable features, leading to a fairer AI system.



Accomplish at the Ideation stage or feature engineering stage (or as a separate workshop to test & reflect upon an AI system)
With the entire AI team

Unexpected scenario - 1,5 hour

This step is performed as reflection on a prospective an AI system. It has the format of a game. The participants can be divided in teams of two. They can pick an unexpected scenario card blindly at first, and later in the game also create cards themselves for the opponent team (figure

9.4). These cards are founded in real-life cases and sources of unfairness in AI. The scenario card is propound at the canvas. The other team may ask challenging questions concerning the implications this scenario would have on the proposed model in a form of reflection. The sheet follows several steps to be filled in concerning the implications for the diverse stakeholders and if something can/needs to be changed to prevent or be prepared for this. The ethical coach can aid the team in the reflection toward the AI Dish and Dishes tale. The aspiration of the game is to assist the team when created an (initial plan) for a new AI system to become aware of the unexpected scenario’s that often make these AI systems less fair. Next to awareness, actually altering the design/dish/ tale preventing these is advocated. The output is a more fair AI Dish/ Dishes Tale/ AI system design, with concrete implementation features.



Accomplish at the Ideation stage or feature engineering stage
With the entire AI team

The Oath - 30-40 min

This exercise aspires to stimulate moral responsibility. Correspondingly it aims to decide and agree upon the implementation steps for fairer AI systems. It bridges both the procedures and requirements of both probity and accuracy while integrating it with the AI principles of IBM. It is a playful manner to culminate and conclude the workshop, made real agreements and actionable statements rather than solely (rich) discussions. The output of this sheet is a morally binding agreement on the design of a fairer AI system. It is not legally binding. It intends to stimulate argumentation when changing certain discussed features/aspects for changes later in the development process.

9.1.5 Conclusion

This section describes the final design of the Ethical AI Coach (EAC) role and the workshop tools. Seven canvases with their intents and actions are presented with general introductions, processes, ethical coach actions and outputs per sheet. The concepts strive to assist in the creation of fairer AI systems and simulate ethical awareness and thinking throughout the AI development process in an inventive and playful fashion. The next section elaborates on the validation of the ideas IBM internally, externally and with client cases.

9.2 Design Validation

This section elaborates on the validation that is executed for the design, particularly is focused on the validation the desirability, feasibility and viability. Simultaneously is tested if the intended goals are met in a desired fashion.

9.2.1 Validation set-up

The ethical role, AI Dish and the Shape workshop are all validated from a variety of perspectives. Due to the diversity of disciplines inter-crossed in this thesis both from an AI background as from an ethical background validation is realized. IBM’ perspective is taken into account and therefore the design is tested within the ethical community as well as individual face-to-face meetings with employees to tailor the design toward IBM’s needs and deliver in an implementable formality. Figure 9.5 shows an overview of the validations performed.

» The central aim was to evaluate the modular toolkit based on perceived value, purpose and clarity. The validation of the ethical coach was aimed at the feasibility, the viability within IBM and the perceived value.

To evaluate the tools, AI teams were asked to use the tools in sessions. Instructions and facilitation of the workshop was provided with corresponding information. After the session

an organized evaluation discussion was held in which questions were asked related to the aim.

9.2.2 Ethical coach validation

The ethical coach role is validated in various ways. Firstly, this approach was discussed with the an ethical expert, Aimee van Wynsberge (Assistant Professor of Ethics and Robots TU Delft) to evaluate it from an ethics perspective. She believes in the role of a value advocate and is a proponent for a person with partial ethical motivation and responsibility of projects. Secondly, the concept of the ethical role is proposed towards the Data & AI ethics community within IBM. This consists of the ethics ambassador, data scientists, designers, marketing and communications employees, CTO Benelux, client executives, lead of CAS and many others. Next to this one on one meetings were conducted with Rob Nijman (Client Executive, Government Sector Business), Reggie van der Westelaken (CIO-Mobile IBM, Manager Europe),

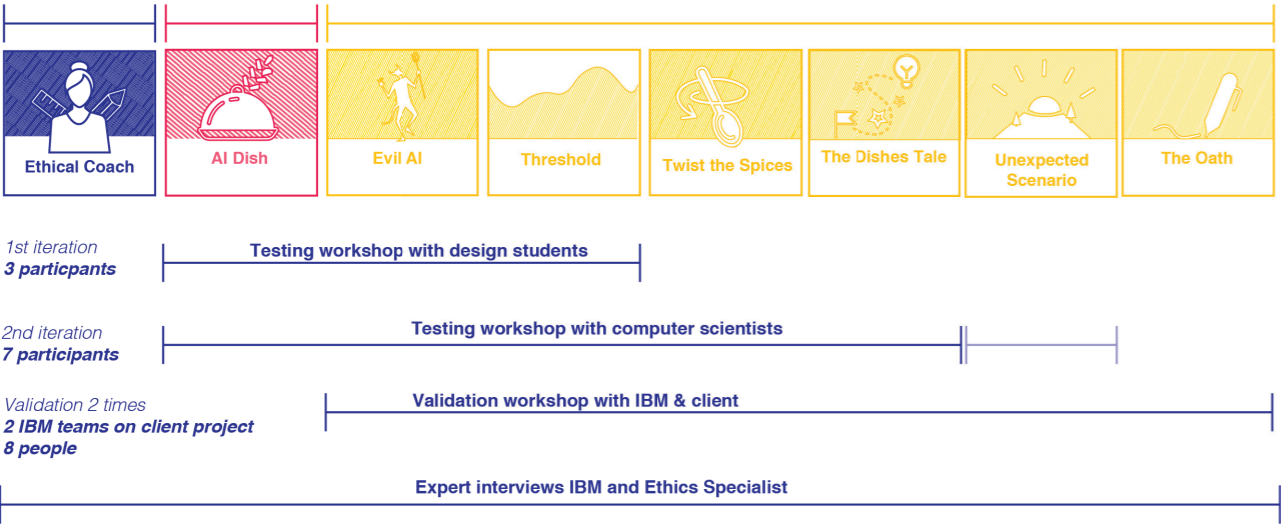


Figure 9.5 | Validation overview visual



”I think if you are actually at the beginning of a project, this could be a really good framework, otherwise you are sitting by yourself, while doing it in this way it formalizes everything, it would be very useful”

- Participant 1 Data Scientist
Validation | Client & IBM

Figure 9.6 | Visual validation session with IBM and client number 1

“ In general I really liked it, I think it is really nice to think try to think about first all the bad things that you can work back wards,I really liked it “

- Participant 1 Data Scientist | Validation 1 Client & IBM

Sophie Kuijt (Ethics ambassador Benelux) to envision the possibilities of creating this role as well as validating the actual possibilities and capabilities of implementing this role. Thirdly, during the course of this thesis contact is established with one of the internal global design teams. In line, they are working on similar topics and willing to collaborate to roll this role out open-source and global. The main insights of these validations are discussed.

Main insights ethical coach validation

Overall, the responses to the ethical coach role are affirmative and mentioned as a novel solution to support AI development teams. The next paragraph explain shortly the advices given during the validations.

Name

It was proposed to change the name from an ethical consultant to ethical coach. It is a new an attractive name, linked to agile coach. In this manner the role sounds more engaging for uptake.

Link to IBM tools

It was proposed to link the existing IBM (design thinking) to the design for straightforward uptake of the workshop and role. Also, it was proposed to make the link to the technical fairness tools and AI checklists of Francesca Rossi.

“ The content was interesting, we have never done it way, however it was a bit quick”

- Participant 2 Client | Validation 1 Client & IBM

Soft coaching skills

The necessary soft skills of a coach were named as very necessary to include in the role description. Similarities with the agile coach are proposed for the selection of suited people for these roles.

Test

Testing the role and workshop with clients and with a practical use case was supported by the CTO Benelux. When results are promising further implementation can be developed internally.

Double sided demand

Two sides of demand creation are mentioned. Both at the clients side demand needs to be created for fairer AI as well as at IBM side. These are both necessary for a successful implementation of the role with the tools.

Competitive advantage

It is proposed to add this to GBS (global business services). It is mentioned to provide competitive advantage for IBM and its services.

In line with the IBM strategy

It is really in line with IBM strategy for more ethical AI (Gerard Smit, 07/02/19, IBM Netherlands Data & AI Ethics community).

9.2.3 AI Dish Validation

Comparatively, the AI Dish is validated from the angles of the AI discipline, the design one (who would be an ethical coach) and IBM perspective. Firstly, the AI dish was tested for understanding with a computer scientist, one on one and iterated upon multiple times. Elements such as readability, wording and numbering were adapted. Secondly, the AI dish was tested with the design students which was emphasized with and advised to continue with the metaphor throughout the workshop. Thirdly, the workshop was tested with three different groups of computer scientists (students) at CAS IBM and another design student as facilitator as well (for

“ You just spend a few hours on this topic and you will benefit the months afterwards ”

*- Participant 2 Managing Consultant
Validation 2 | Client & IBM*

Figure 9.7 | Visual validation session with IBM number 2



detailed insights see appendix O).

The value of the AI Dish was clearly expressed in twofold: (1) as a manner of structuring and discussing information and (2) a tool for reflection. Th AI Dish canvas was adjusted for purpose of clarity and the visual appearance. Also example answers are created and provided.

Fourthly, the AI Dish is presented at the Data & AI ethics community within IBM and a global US team. Due to time limitations and reached level of confidence on the value of the sheet with earlier tests, it is not validated in the last two validation sessions.

9.2.4 Shape Workshop Validation

Similarly, the shape workshop is tested from the design perspective, AI team perspective, IBM's one but also from a clients one. Specific elements of the shape workshop were tested with design students. The second iteration of the workshop was with two teams of computer scientists. The identical version is discussed within IBM with the design department and the ethics ambassador within Benelux.

Lastly, elements of the workshop are validated twice with internal AI teams and clients working on

projects. Two client project of IBM Benelux were used in the workshop of both two sizable Dutch banks. The workshop took two hours due to time constraints. The following sections elaborate firstly on overall insights and advices retrieved, after which a closer look is given per sheet.

Main insights final validations

This is validated with two IBM data science teams and their clients, next to the tests performed during the ideation stage. The workshop setting was limited to two hours. These are founded in a feedback discussion with guided question after the workshop.

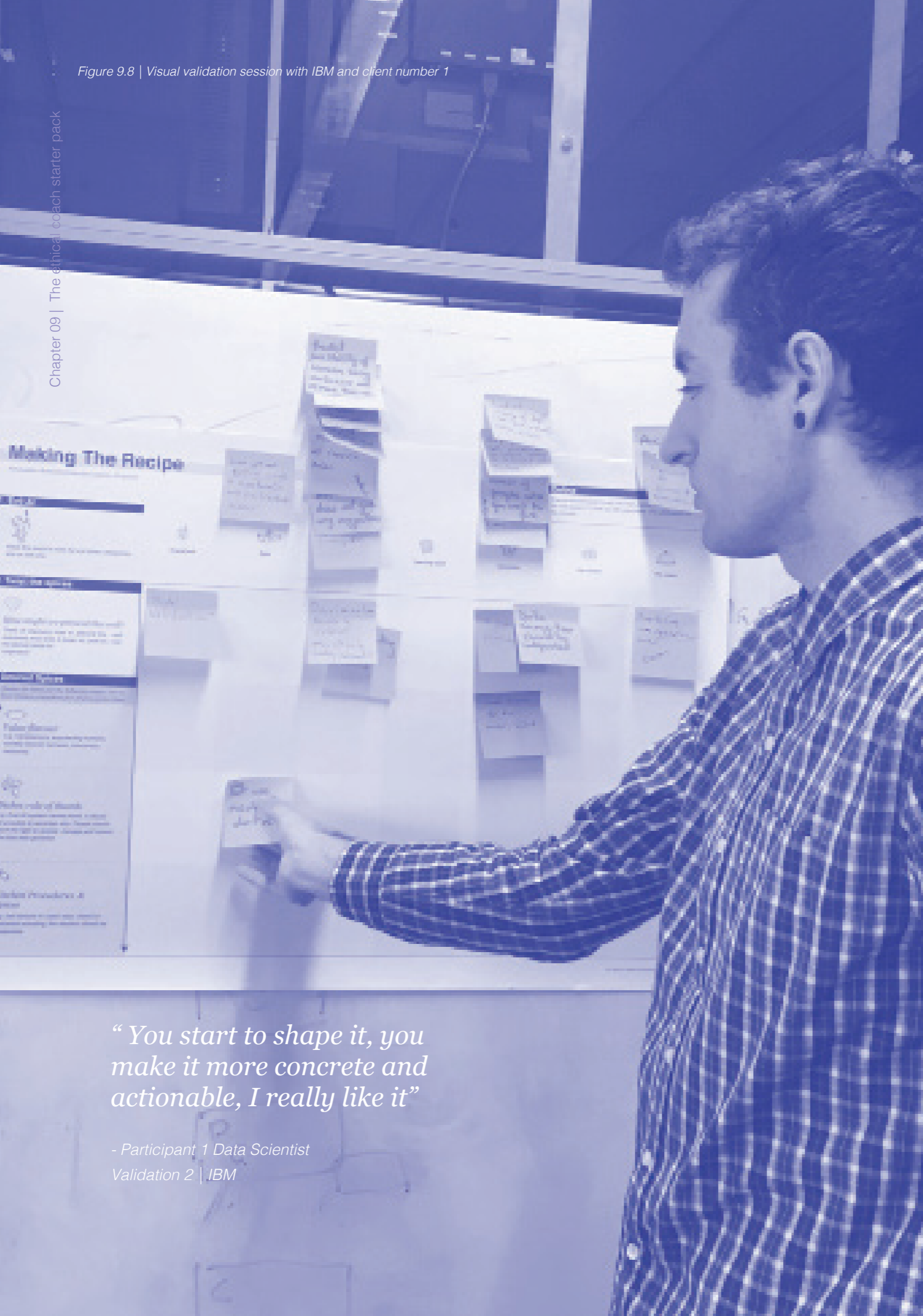
Overall

The workshop overall is seen as valuable manner to structure discussion, change perspective and a creative manner to incorporate ethics in the beginning of the project (quote on figure 9.6 and

“ What went well, when discussing this like seeming well obvious topic, I realized some of these I have not thought of, that are the new ideas I got from this workshop “

- Participant 3 Data Scientist Validation 2 | IBM

Figure 9.8 | Visual validation session with IBM and client number 1



“You start to shape it, you make it more concrete and actionable, I really like it”

- Participant 1 Data Scientist
Validation 2 | IBM

quote on figure 9.8). An advice to strengthen the visual structure is to use color coding with the post its in the different excises. In this way it is visually easier to grasp the content from a distance.

The time span of the workshop is repeatedly mentioned as a restricting factor in reaching the desired depth of the answers and discussions. The flow of the sheets was mentioned to be clear and refined. The explanations describing the analogy on the sheets were much appreciated. A few participants mentioned to prefer an example filled in at the presentation.

The analogy is affirmatively responded to, however more explanation is necessary then was given currently for a clarity.

Also it is mentioned by one of the participants that it would be valuable to do this workshop together with users and stakeholders.

“I think was really nice that you managed to connect everything”

- Participant 3 Data scientist | Validation 1 Client & IBM

Evil AI

The evil AI exercise stimulated the creativity and change of perspective of the team. In the feedback session all participants expressed their enthusiasm towards the evil sheet and the evil stimuli. It was noticed the way of reasoning altered and atmosphere changed effectively for the workshop. One comment was to stimulate the perspective of the user extra as a facilitator. Additionally it was proposed to strengthen the evil side, make people really get into the role by giving them for example hats.

“The cards really trigger the investigation. The ideas I really got from the cards”

“I agree this evil cards are really cool It definitely triggers the creativity”

- Participant 1 & 2 validation 2 | IBM

Accuracy threshold

This sheet is solely tested with the first validation

due to time constraints in the second session.

It is expressed that the word accuracy is a trigger word for scientists and thereby it is advised to change the naming into something more abstract. This was the main reason of unclarity in the beginning of the exercise.

Making the recipe

This sheet is seen as valuable manner to translate the discussion into actions. It is remarked to well relate to the different sheets in the workshop.

“I really like the preparing the ethical recipe, because it makes it actionable, it gives the recipe”

- Participant 3 validation 1 | IBM & Client

“You start to shape it, you make it more concrete and actionable, I really like it”

- Participant 1 validation 2 | IBM

Dishes tale

Diverse feedback is given to the dishes tale. For some participants it was a differing from the other exercises

“I lost the connection with evil, if you can strengthen that”

“Here I felt with the dishes tale, the value was not completely clear, however it could be due the time.”

- Participant 2 & 3 validation 1 | IBM & Client

Contrastingly by other participants in the same session it was mentioned to be the most difficult exercise, however the most valuable one as well. Thus more clarity of the goal of this exercise could support alignment of the team. Additionally, it was remarked the workshop could be tailored toward the phase the project is in. In very early stages of the project it might be valuable to construct the dishes tale around the development process. When the development process already started, it might be more valuable to look into the actual context of use more in detail. Thus as an ethical coach to sense the needs of the team and to

tweak the workshop toward them is proposed.

The Oath

“It is good when you start an AI project to have everybody on the same page ethically”

- Participant 2 validation 2 | Client & IBM

The oath is not filled in due to time constraints, however is discussed with the participants. All the participants expressed their enthusiasm concerning this sheet. Nevertheless it still needs to be tested while actually filling it out.

9.2.5 Conclusion

The conclusion relates to the aim of the validation: (1) the purpose, (2) perceived value and (3) the clarity of the tool.

Purpose & Value

Overall, participating in the workshop was appreciated and perceived as a creative novel way to bring the ethical dimension to the development process (). The approach (bottom-up) in specific was welcomed. The following quote resembles that:

“...no the ethical discussions are also important of course and the think-tanks, but we make it happen on the floor, and we if we decided to fully automate data, without any human intervention that is our decision and if that is risky or not”

- Participant 2 validation 2 | IBM

The workshop is perceived as an effective manner to bring the ethical dimension in the AI development process as well as a way to mitigate risk for ethically misaligned pitfalls.

“ It is good that we also look at different dimensions, time, budget dimensions, and I think it is good to consciously include the ethics dimension to this”

- Participant 3 validation 2 | IBM

“I think something like this would mitigate the risk, in the selling case for example, of not tragedy but big problems of delivering these use cases”

- Participant 1 validation 2 | IBM

Reflecting on the session is remarked that it would be more valuable to have also non-data scientist in the room. This would lead to a more diverse range of perspectives and answers.

Clarity

Most canvases effectively guided the participants, however the support of the facilitator is needed. It is proposed to spend more time on the explanation of the analogy although it is sympathized with. Providing an example of a filled in sheet could increase the clarity of the diverse steps. A few small remarks on wording were made to prevent confusion from a data scientist perspective.

Two main insights eliminated to increase clarity: (1) color coding of the post-its, due to the different levels from abstract to implementation and due to the flow of the post its from one to the other canvas, color coding the post-its would provide a clearer overview during the session but also would be easier to reflect upon later in the process (2) tailor the dishes tale towards the development process or the actual implementation of the AI, depending on the teams stand in the process. This would increase both the clarity of the dishes tale as well as the value it would bring.

09 Ethical AI Coach Starters Pack

Final design

An new organizational role is proposed, the ethical coach, with an accompanying modular workshop. The ethical coach starters pack consist of a supporting modular workshop, which can be tailored towards the AI development stage and the AI team need. The elements of the workshop are: (1) AI Dish, (2) Evil AI, (3) Threshold, (4) Making the recipe, (5) The dishes tale, (6) The Oath and (7) The unexpected scenario.

The central aim is to support the AI teams in the development of more fair AI systems. This is achieved by explicitly resolving value-tensions identified in the process by means of the ethical coach with an accompanying modular workshop, together named as: the ethical coach starters pack. Thereby, this design increases the capacity and infrastructure of IBM to facilitate and promote the development of fairer AI systems. The final design is presented in the accompanying file.

Validation

Validation is performed on all aspects of the design with the central aim to validate the purpose, perceived value and clarity of the workshop. This is realized by means of a variety of workshops from design, data science, IBM & client perspectives.

For the ethical coach validation is performed with the central aim for feasibility in IBM, viability and desirability by means of face-to-face validation sessions and presentations.

“ You just spend a few hours on this topic and you will benefit the months afterwards ”

- Participant 2 Managing Consultant
Validation 2 | IBM

Chapter 10 |

Recommendations & Discussion

This final chapter shares recommendations for IBM, implications on the research fields, discussions on limitations of this project. Finally it closes off with a personal reflection.

In this chapter

- 10.1 Recommendations & implementation requirements
- 10.2 Discussion & research implications
- 10.3 Contribution to practice
- 10.4 Limitations & future research
- 10.5 Personal Reflection

10.1

Recommendations & implementation requirements

This section elaborates on the recommendations concerning the various elements of the design. Also, it shares overall recommendations for IBM when proceeding with this ethical strategic direction founded in literature, expert interviewed, design research and validations.

From a strategic design perspective, the link between the organizational strategy and the design is made to create a fruitful fit. Strategic alignment of ethical values and actions is profitable for business (Shilton, 2018) and an ethical strategy gives a strong sustainable competitive advantage in the market on the longer term. In line the recommendations for IBM are to proceed with an ethical AI strategy. The following recommendations are arranged per design facet. After which general recommendations are elaborated upon.

10.1.1 Ethical Coach

IBM works with online and offline badging and training programs. Conversations are started with a US design team about developing a badge for an ethical AI coach. Throughout the validation of this concept, it appeared to be essential to focus on the soft coaching skills such as: empowering, energizing, learning and growth mindset. The following sections elaborate on four main recommendations for the ethical coach extracted from the validations.

Design Mindset

The recommendations based on this thesis for the ethical coach are to focus next to the obvious ethical knowledge, also on creative facilitation and design thinking skills and mindsets. During the validation sessions with the clients, it unraveled once more that one of the strengths of this workshop lie in the guidance of the

facilitator. Asking the right questions at the right moments is crucial to reach desired depths in the output of the workshop. Additionally, sparking imagination really strengthens the output of the sessions in terms of inventiveness and ethical considerations.

Implementation I Start Small Aim Big

For the implementation of the ethical coach role it is recommended to start with one coach, supporting a few AI projects in approximately 10% of their work week. This reduces the risk for IBM compared to making it a full-time role straight away. It also allows to experiment with the best suiting implementation for the role. Simultaneously, manners to measure the impact of the ethical coach on projects need to be created. In this style, the coach role can be adjusted and tailored towards an expanded role and badge program. Additionally, it can be proved in the organization if it prompts the desired benefits. If the results are beneficial, the coaching roles can be expanded to for example 20% and supplementary employees educated for this role. Per industry, value knowledge bases can be developed case by case. Concurrently the global expansion strategy can be created to further expand the role.

Shaping the clients' perspective

Next to internal education and communication, also external education is essential to lead to the viability of this role. By means of the value

proposition and internal interviews appeared that certain clients i.e. in the governmental sector, are realizing incorporating ethics in the AI development process in vital. However, with alternative industries awareness concerning AI ethics and the long term benefits need to be taught in order to create demand from the clients side for an ethical coach. It is encouraged to create this demand by building awareness, education and the creation of show case ethical AI projects

Perhaps this could even lead to a competitive advantage for IBM (discussed internally).

Win trust of data scientist

From the validation appeared it is crucial as a non AI expert (ethical coach) to gain credibility and trust of the data scientist in order to sincerely cooperate. This can be attained by shortly presenting new expertise concerning their own discipline they are not familiar with, such as a non famous ethical misaligned product or the unfairness sources.

10.1.2 The starters pack

For the ethical coach starters pack it is recommended to keep the tools open source due to the nature of the topic. Thereby expanding the knowledge base of building more ethical AI systems for society. To reach the full potential of the tool-set it is recommended to further build a website with the ethical AI tools easily accessible and explained in a practical way. Modularity is strongly emphasized with in this toolkit in order to allow tailoring per project and purpose for easier uptake in practice. Currently an US design team and I are in touch to further develop this together with the existing tools developed by them.

AI Dish

Specific for the AI dish it is strongly recommended to use it early in the AI development process. From all the test and validation sessions this observation is raised. Even tough particular questions might be difficult to answer early,

when used further in the process this exercises abolishes value as decisions are already made. It is the role of the ethical coach to support the team in making up their mind about these ingredients and make sure that is known the answers they provide are not final. A more detailed guide for an ethical coach could be developed to support this role.

Shape Workshop

Based on the validation of the shape workshop it is recommended to have diverse layers of depth the workshop can be given in of the phase the AI project is in. When in a specific project certain decisions are not made yet e.g. the dishes tale can be focused on the development part of the AI system. However, when used further in the project more specific sequences concerning the monitoring and use context can be focused on. A clear explanation of the analogy and steps of the workshop appeared desired from the validation i.e. the sheets filled in with an example case to increase clarity. It is most valuable for the output to have a diverse range of participants present to include a wide range of perspectives in the workshop.

10.1.3 Ethical fulfillment in AI

For the third element of the design framework, ethical fulfillment in AI, a concept is developed in this thesis and presented in appendix Q. This framework element is essential for a substantive implementation of ethics further in the AI development process after completing the shape workshop. The concept for the ethical fulfillment, the ethics fulfillment cheat sheet, is proposed in this thesis and is briefly explained in this section.

The analyses of the interviews and generative tool show that the data scientist, often works by him/herself in the feature engineering and modeling phases of the project. Critical ethical decisions are made in these phases often solely by one person. At the same time, the need for a nudge is the most pressing when choices have delayed effects, are infrequent, difficult, with poor

feedback and ones for which the relationship between choice and experience is ambiguous (Leonard et al., 2008). The consequences of created AI systems are often the long term (compared to pressing deadlines and financial KPI's) and have poor feedback. The effects are delayed (at the end of the development process) and the relationship between the choice made in the modeling phase and the actual systems output is not clear. Thus, this thesis puts forward that the translation of the shape workshop towards the feature engineering and modeling phase is a good candidate for nudging.

This concept proposes a task of the ethical consultant, to analyze the shape workshop and create the suiting nudge/behavioral override strategy for that specific AI team. It is supported to integrate reflection moments towards the “AI dish” and the “Dishes Tale” at scrum meetings/stand ups in the current AI development process. The proposed strategies in this thesis are based on the work of (van Lieren et al., 2018) and are: (1) Add small friction, (2) Increase decision moments in the process, (3) Highlight losses and therefore active choice, (4) Personal ranking, (5) Make commitment with an action plan (6) Checklists to easily remember information, (7) Real-time feedback of consequences, (8) Create personalized feedback and (9) Create reminders & alerts. These are consolidated into a “ethics fulfillment cheat sheet” (see appendix Q for more detail of this concept).

10.1.4 Employees education

From the literature review and design research it appeared the workshop on its own is not yet capable of reaching the desired impact. Ethical education is needed within IBM, not only for the ethical coaches, but also the employees working on the creation of AI systems. In this fashion more moral responsible engineers could be shaped. By education and awareness, they can become intrinsically motivated for making the right choice even when they are by themselves modeling the AI system or engineering features. Throughout the course of this thesis it has been

noticed a change of mindset is needed. In the interviews appeared that through sharing knowledge and discussion only certain levels of awareness are achieved. The prototypes, and the change of perspectives in the workshop showed promising results in terms of ethical reflection. It had a durable and substantial impact. Additionally, in this research it is discovered, that making the experience more personal (linking to the AI team or to people who are close to them), placing the AI system in actual context while making scenarios are examples of manners how to initiate this education next to the ethical knowledge taught. Therefore, it is recommended in further education to use practical hands-on examples, scenarios or provocations to modify the current perspectives towards the creation of more ethical outcomes.

10.1.5 Assessment & recognition

One facet necessary for ethical implementation in organizational AI processes, is that it becomes a priority in the daily agendas of managers, data scientists, IT etc. If people and projects are assessed on entirely different criteria, then when time pressure and deadlines come closer there is a risk ethical outcomes and considerations become secondary priorities. Two approaches to prevent this are discussed.

Recognition for ethical projects | Thus, to stimulate people and teams to contain ethics as a priority in projects diverse strategies can be taken. For example these can be stimulated by positive recognition and publicity of these ethical projects in order to motivate teams to work toward these. Or more fundamental changes can be advised, adding or changing ways in which projects and people are assessed (like KPI or competition matrices) and thereby change the motivations of the AI teams.

Intrinsic ethics | From this research appeared that the data scientists have a strong sense of freedom. Thus, it is recommended to pay attention to the way ethics is proposed. Hence, not obliged to the AI team to take it into consideration,

rather by aspiring ways in which more intrinsic ethical considerations are stimulated (such as education, change of perspective).

10.1.6 IBM

The three building blocks of an ethical organization capacity are unraveled in this thesis: ethical people (addressed by education), ethical processes (the workshop structure and implementation strategies) and an ethical organization (partially addressed by the coach). However, to create and be an ethical organization more is necessary than the ethical coach role with the accompanying modular workshop. Thus a few general recommendations are presented based on this body of work.

Establish an ethical culture | Diverse heterogeneous teams should be stimulated and feedback and openness in the company encouraged to lead to a more ethical culture.

Sustain the ethical community | The initiation of the ethical community during my graduation is a great start of aligning and sharing ethical initiatives and knowledge. Sustaining this community and actively sharing projects and knowledge in these meetings is strongly advised to create coherence in the external message and generate new ethical initiatives.

AI ethics communication strategy | It is advised to strengthen the external communication of IBM, spreading the message it is an ethical AI company.

Currently IBM is the only one of the prominent technology companies who keeps the clients the owners of their data. Additionally it strongly invests in research towards more ethical AI.

However the marketing is less active in this area. An ethical AI strategy really fits the foundation of IBM, it is in their veins. During the validation sessions it is mentioned that the introduction of the ethical coach, could be a competitive advantage for IBM. Founded in insights generated during the course of this thesis, it is recommended

to both internally and externally share these ethical projects and visions, distinguishing itself from competitors of whose business models are organized around selling data (which might be less ethically desired).

10.2

Discussion & research implications

This research sheds a critical light on the current AI and applied ethics fields through a design lens. This section elaborates on the answers to the research questions and design goal.

The destination of this thesis has been to discover and create practical implementation, through (strategic/critical) design, supporting AI teams in the creation of fairer and value aligned systems and thereby the organizational capacity of support in ethical AI development. This is achieved by consolidating insights of literature, internal and external analyses, expert interviews and design research into the day to day work of the AI team, into a conceptual framework (chapter 7). These insights laid the foundation for the design: a new organizational role, the ethical coach with a modular toolkit to co-create fairer AI systems. These are presented in chapter 9. This discussion of this thesis follows the structure of the research questions in the following sections divided into sub-topics (figure 10.1).

10.2.1 When ethical support is needed

In line with earlier research in other industries (Shilton, 2018; Spiekermann, 2015) this study detected, by means of semi-structured interviews and provotypes, a lack of the ethics integration in AI practice. In the current field, also a paucity of alignment of the team concerning the (technical) decisions and their societal consequences is identified.

Moreover, this thesis focuses on when this ethical support is needed most. Ethical decision moments are when ethical support is most urgent in the processes (Davis & Patterson, 2012). Herewith, the ethical decision moments are identified in the AI processes by means of the generative tools in this study. The ideation phase resulted as one of the significant ethical

decision moments. This is in line with the analysis of the ethical methods and tools from which appeared that the impact of incorporating ethics is most influential early in the process (Davis & Patterson, 2012; Goodpaster, 1991; Guston and Sarewitsch, 2002; Mephram & Kaiser et al., 2006; Ratto, 2011; Schot & Rip 1997). However, from the interviews also appeared the data scientists make essential decisions, often by themselves, in the modeling and feature engineering project phases. This causes these ethical decision moments to be critical and incline for ethical aid. Thus, next to supporting the entire team in the ideation phase also ethical fulfillment is necessary in the project stages later on.

10.2.2 How to create an organizational capacity and infrastructure to support ethical uptake in AI projects?

Three main elements are extracted from the research to answer this research question. These are elaborated upon in the following section.

01 Three strategic ethical building blocks

From the ethics literature review, three main ethical building blocks are aggregated for ethical organizational processes and outcomes: (1) ethical people, (2) ethical processes & tools and (3) ethical company (p. 40). These building blocks describe the necessities for more ethical outcomes in organizations. Founded upon these building blocks is argued that to deliver more

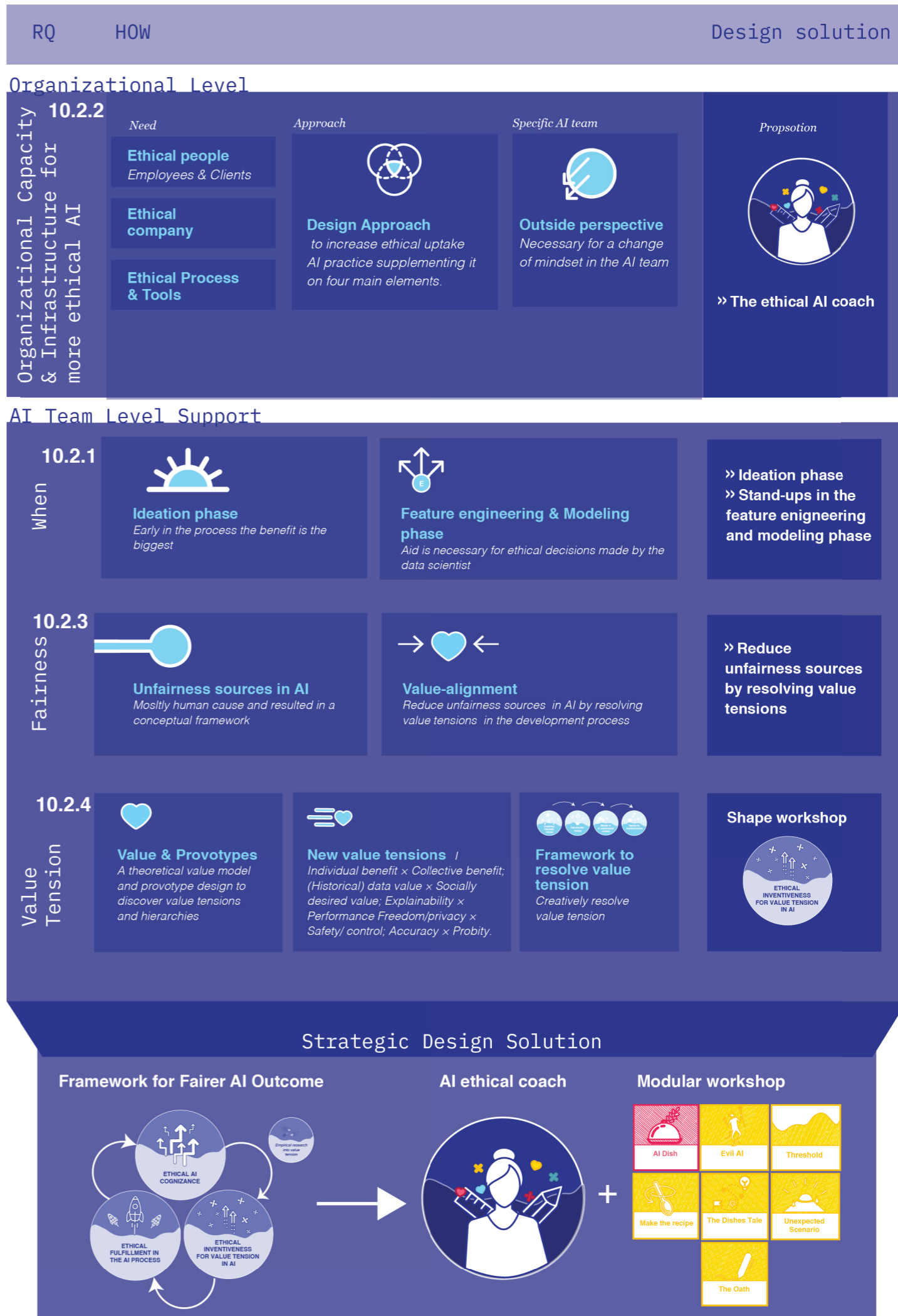


Figure 10.1 Discussion visual

ethical AI systems, not only the ethical processes are important but also the characteristics of the employees and the companies' culture/structure. Hence, in the AI teams also the ethical motivation, morality and knowledge should be triggered and taught. By means of the semi-structured interviews and generative tools a shortage of ethical awareness and motivation in the contemporary AI field is presented.

02 Approach I Design enhancing ethics integration

Despite lots of interest in ethics in research, there is a lack of incorporation of ethics in decisions in commercial development, even though it has competitive and strategic benefits for companies (Shilton, 2018). In line scholars mention a lack in translation of incorporations of "abstract" ethics into the AI development process and coding (van den Hoven 2013; Shilton, 2018). Thus, new approaches to incorporate ethics in AI processes are necessary for ethical uptake in practice.

This thesis looked into how applied ethics can benefit from a design perspective. Based on the literature review, this thesis identified four ways the design field can complement the applied ethics field: (1) new methods and tools for complex and uncertain problems (Whitbeck, 1998; d'Anjou, 2011); (2) use of imagination to stimulate creative solutions by means of synthesis (Lloyd, 2009); (3) deal with conflicting demands by use of empathic and creativity simulating tools (Dorst, & Royakkers, 2006); (4) opening up the opportunity space by exploring the problem iteratively (Van de Poel et al., 2007). In line with literature is argued that design can support the applied ethics field with synthetic reasoning and creativity to come up with practical and novel solutions, rather than solely discussions, using design as a prototypical kind of ethical thinking (Lloyd 2009).

Thus, this thesis proposes (strategic) design as the approach to close the gap between "abstract" ethical AI principles/discussions and the AI practice. Thereby, design methods and tools

are used to spark imagination and the solution space, creating the desired implementable ethical solutions that currently lack.

However, this research shows that design additionally can support the applied ethics field not only in the solution and synthesis spaces but also by means of inquiry. Generative tools and critical design are used to research, provoke and discover novel value-tensions, subsequently fueling the design space, while critiquing and questioning the state of the art of the contemporary AI field. This led to new propositions of value tensions to the applied ethics field and thereby contributed by means of inquiry.

03 Outside perspective I Ethical AI Coach

The design research in this thesis shows it is very troublesome, if even possible, to achieve an ethical change of mind-set by the AI team themselves. In line with literature the AI team is often a homogeneous group of people (Flasiński, 2016) mostly not educated in ethics but are taught more technical perspectives. A quote to illustrate:

"I choose Advanced analytics because it is currently the wild west, there is practically no regulation so we can make models the way we want" - Interviewee (data scientist)

This thesis argues that a change of mindset is needed and that the AI team cannot achieve that independently. The literature review shows that giving a team member explicit ethics responsibility and values during the technology development process, has benefits for ethical results (Fisher and Mahajan, 2010; Manders-Huits and Zimmer, 2012; van Wynsberghe and Robbins, 2014; Shilton and Anderson, 2017). Also, value consciousness and explicit responsibility of the design helps to build values reflection into the scope of work and the success metrics of a team (Shilton, 2018).

An ethical AI coach role is therefore proposed, based on characteristics of a design creative facilitator, value advocate and IBM agile coach. In line with literature of a value advocate, an ethical coach has the role to pro-actively stimulate ethical discussions in the team, asking questions concerning the development. Also in line with Shilton (2018) who proposes introducing value levers (entry points for ethical discussion). Contrastingly with value advocate literature, it is not the responsibility of the ethical coach to create more ethical AI. It is intended this role stimulates the co-creative act of imagining and developing new AI systems with an ethical fuel of innovation, incorporating not solely technical feat but social and human one. The creative facilitation capabilities is strongly empathized with to create alignment of the team on ethics.

" We need ethicists working in the companies that can afford them as part of the design team, where they can start to uncover the common issues other companies are running into"

- Aimee Van Wynsberghe, Founder Responsible foundation.

Hence, is argued that next to the benefits for ethical capacity of the AI team, the ethical coach also increases the knowledge concerning AI ethics as a whole, creating experts in AI ethics in society.

» Concluding, for an organizational capacity leading to more ethical AI development, next to ethical tools and process, also people need to be educated. Their moral motivation needs to be stimulated. This thesis does this by means of design principles and tools. An ethical AI coach role is put forward as a designed solution based on this study. In particular is focused on resolving value-tensions in relation to fairness which is discussed in the next sections.

10.2.3 How to support AI teams for fairness in AI projects

Current translation of AI ethical principles towards the day to day work of AI teams is lacking. Thus is researched how to create practical support for the AI development teams. Two main elements to answer this research question are discussed.

01 Identified unfairness sources I Human

The present research field occupied with fairness, sheds a light on attempts of finding agreed up definitions or defining fairness (Gajane & Pechenizkiy, 2017; Taylor, 2017; Zhong, 2018). The argument to put forward this research is defining what fairness is, and thereby what a fair AI system is very context depended. A universal definition of fair AI can therefore lead to ethically misaligned AI systems, due to misfit in i.e. context. A novel perspective taken in this thesis is examining fairness of AI systems not by defining fairness, but rather examine the sources of unfairness in AI (p. 54). Ten sources of unfairness in AI are identified by means of research into ethically misaligned AI systems (p. 56): (1) algorithmic bias & incomplete training data; (2) subjective measurement of data; (3) choosing target variables; (4) oversimplification; (5) redundant encoding; (6) reinforcement in feedback loops; (7) self-fulfilling predictions; (8) inconclusive evidence; (9) untransparency; (10) reinforcement of prediction. Surprisingly the sources of unfairness in AI systems appeared to be mostly from human origin (p. 50). Concluding, reducing these identified sources of unfairness is not solely a technical feat. There is also necessity in changing the human processes of developing AI to reduce these. From the semi-structured interviews appeared these sources of unfairness are generally not addressed in the contemporary AI development.

In this thesis, context specific fairness is taken into account, in line with peripheral values

literature (Borning & Muller 2012; Van den Hoven, Vermaas, & Van de Poel, 2015) albeit contradictory with AI fairness literature (Zhong, 2018). The way to incorporate more context specific fairness is approached from a design perspective.

02 Value-alignment and thereby value tensions

Currently there is the lack of integration of “desired” values into (AI) systems development processes and their outcomes which can result in unfair AI systems. Nine main challenges are identified in value-alignment in AI (p. 65). One of them in particular related to fairness is value-tension. Not addressing value tension in an explicit way can lead to a lack of appropriation by disadvantaged groups, unfair AI systems or even more drastic consequences such as system sabotage (Flanagan et al. 2005). Additionally, little research has been performed into value tensions in AI. Thus, this thesis took up this under-researched topic.

By means of provotypes is discovered that the present AI teams do not explicitly resolve value tensions related to fairness. Hence support is needed to resolve these.

»Concluding, to support AI teams in creation of fairer and value-aligned AI systems, sources of unfairness in AI need to be reduced by resolving value tensions. However, first research to unravel value tensions in AI is required.

10.2.4 How to support resolving value-tension for AI teams in AI projects

Three main elements are extracted from this research to answer this research question. These are elaborated upon in the following section.

01 Conceptual framework values

Before addressing value tensions, a deep understanding of value is required. A literature

review is conducted into value sources and manners to describe them, as the diverse fields addressed in this thesis have distinct definitions. This is boiled down in a conceptual framework presented at p.62. This led to the identified need of context specific values as well as strategically fitting IBM values. Furthermore the preferable dimensions of values are extracted: (1) Performed; (2) Purposeful; (3) In sync; (4) Both peripheral as central values (p.62). These are taken into account in the development of the practical support for the AI team.

02 New value tensions

To discover the value tensions specific for AI development processes, provotypes are designed (p. 84). Provotypes provide an opening to conflicts in processes, these are artifacts/pictures that embody tensions in a certain context in order to explore new design opportunities (Boer & Donovan, 2012). Results from the literature review, generative tool and provotypes derived five new value-tensions in AI development related to fairness. These are the following: (1) Individual benefit × Collective benefit; (2) (Historical) data value × Socially desired value; (3) Explainability × Performance (4) Freedom/privacy × Safety/control; (5) Accuracy × Probity.

03 Framework to resolve value tension

By means of literature reviews from a diversity of disciplines seven strategies for resolving value tensions are identified: (1) untangle value tension; (2) decompose values; (3) avoid problematic features for stakeholders; (4) avoid problematic features for stakeholders; (5) decentralize responsibility; (6) quantify values & consequences; (7) untangle consequences. In addition to these design methods and tools were explored to consolidate in a framework to design for resolving value tensions. A novel approach of using design methods and tools with ethical ones led to a conceptual framework to resolve value tension in AI development in a context specific fashion (p.98). The framework

consists of the following steps: (1) creatively demystify values; (2) decompose values; (3) situate AI in user/societal context; (4) decide for implementation; (5) continuous reflection. It accounts for diverse levels of values such as moral ones and technological ones.

This framework is used to design a fitting support for AI teams to resolve value tension in the development process.

»Concluding, values sources and their tensions have been unraveled in the AI development process. This body of work presents five novel ones. A framework to resolve these AI value tensions inventively, co-creatively and context specifically is designed. Furthermore a workshop is designed with this framework disclosed in the next section.

10.2.5 Design solution bridging the three questions.

To answer the research questions in a coherent fashion, a framework is created to support organizations to aid the AI teams in the actual AI development for fairer AI (p. 96). In particular it focuses on resolving value tension in relation to fairness.

To create an ethical AI organizational capacity, a new role is created: the ethical AI coach.

And the way this role incorporates and implements support is by a modular workshop created on the foundation of this body of work. The final design for support is a starters pack with an accompanying modular workshop. This is validated from a diverse range of perspectives of both research, clients and practice.

10.3

Contribution practice

This research is performed at the intersection of three disciplines, AI, applied ethics and (strategic) design. This section shares the contribution towards the three fields.

This research contributes to closing the gap between AI ethics and the AI industry. Overall, it is intended and aspired that the ethical coach and the toolkit which are introduced in this thesis, will contribute towards the adoption of ethical considerations in AI practice by reducing the sources of unfairness and resolving value-tension, within IBM and externally. It is encouraged to design diverse kind of ethical support for the AI teams with the use of the proposed frameworks.

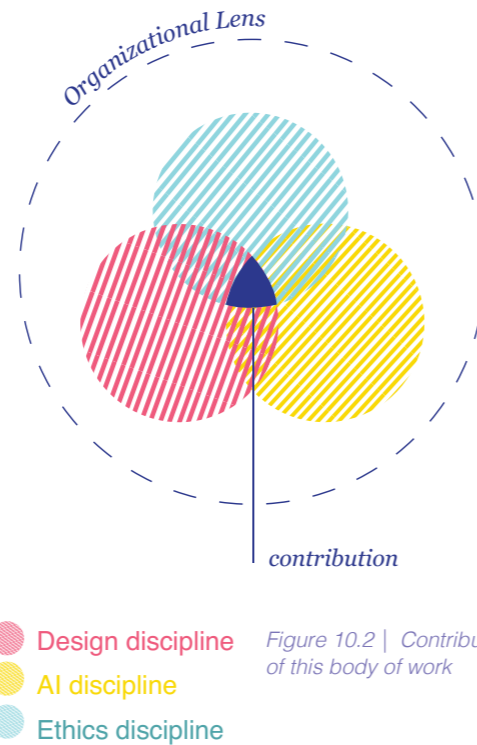
10.3.1 Design practice

The ethical coach starters pack aspired to contribute to design practice by providing tools and handles to understand and get involved in the AI development process by supporting it in becoming more ethical.

The ethical coach can spread awareness of the importance of ethics in AI and the designerly approach to it. The coach can stimulate discussion concerning the AI ethics field and in that be a proponent in proposing ethics as a design quality and also as quality to facilitate ethical development approaches by design. Thereby, new design roles in organizations might be flourished.

It is aspired to support other designers in the creation of new tools and methods based on the frameworks presented, to increase the uptake of ethics in AI which is desperately needed.

This thesis combines critical design (provotypes) with a strategic design approach which supplemented each other by critiquing the type of systems that are made by the use of provotypes, discovering new value tensions. This supplemented the strategic design



approach in the ideation and solution space s. It showed the benefit of the combination of both techniques. Thereby, it proposes interesting ways to combine these two approaches in fruitful fashion.

10.3.2 Applied ethics practice

This thesis proposes a new manner of integration of ethics into practice by the use of the conceptual framework and the design tools and mindset. Additionally, it aims to provide a basis for the design of organizational ethical capacities and infrastructures for AI development.

Also, it proposes a new practical proposition the ethical AI coach, which combines design skills and mindset with an ethical one.

Furthermore, it proposes a new way to resolve value tension with a variety of value levels. Simultaneously, it provides a practical manner to deal with both situational and central values in context specific fashion presented in a framework open for reuse.

Finally, it consolidates value literature in a graspable manner for practice to align value discussions and conversations based the value sources and ways to describe it.

10.3.3 AI Practice

This thesis aspires to provide new approaches to create fairer and value aligned AI systems in AI practice. It does so, by creating a supporting role at the intersection of AI ethics and design with an accompanying modular toolkit.

It presents a new approach to address fairness in AI by reducing unfairness sources in context specific fashion.

Furthermore, new value tensions in the AI development process are derived. New approaches how to resolve these are presented. Finally it shed a light on industries which appeared to be more ethically challenging. These are the insurance industry and the financial one.

10.4 Limitations & future research

Finally, a few influential limitations need to be considered. The following sections elaborate on the limitations per process stage. Furthermore, it proposes a few future research directions.

The intent of this thesis is to support the AI teams with in the creation of more fair and value-aligned AI systems and the organizational capacities to support that ethical uptake. To arrive at that stage, this body of work researched the current state of the AI field, the value-tensions appearing in the AI field, at the intersection of three disciplines AI, applied ethics and design. It is not meant to discard the research of these industries rather to shed a fresh perspective in the field in search of novel implementable solutions and boost the awareness concerning these topics. Following the limitations of the research phases are shared.

10.4.1 Limitations of literature research

In the first phase a literature study into AI, Ethics, Fairness and combination of these was executed for the creation of a in depth understanding of these fields, transpire gaps, relationships and how these can be supplemented by a design perspective. The field of AI and ethics, fairness are fields professionals study their (entire) careers. The amount of literature in these fields is astounding. Due to the breadth and depth of the fields, insights could have been missed.

10.4.2 Limitations of design research

In the design research phase the choice was made to focus on machine learning insurance teams within the Benelux. The value-tensions identified in this research phase are more likely to occur in other industries AI operates. However, further research needs to be conducted to

confirm/disapprove these within the insurance industry and outside. Also, industry and country specific value tensions need to be researched to draw substantial concisions for other countries, cultures, industries and eventually sport designers in developing support for these.

10.4.3 Limitations of design

The framework for design is founded in literature, the generative interviews and provotypes. The final design is based on this framework. However, due to the time frame of this research the framework has not been tested further for the development of new tools.

Additionally, in this project also trade-offs are made concerning the depth of the diverse tools in the toolkit. For example the “ethics fulfillment cheat sheet” is solely a first idea on how to implement the ethical decisions in the subsequent stages but further has not been tested, neither developed in detail.

10.4.4 Limitations of validation

The ethical consultant starter-pack is tested with a variety of disciplines as well as with two clients of IBM. Nevertheless the impact further in the AI project from these workshops has not been researched due to the time-frame of this graduation. Researching the actual impact of the workshop settings with the ethical coach role would be the next step. Also, due to the time-frame of the research the “full” workshop could not be tested from start to end at once, rather the workshop is tested modular to spread the load on the different teams contributing to this research.

10.4.5 Future research

Based on the results of this thesis propositions can be made for future research. Due to the newness of the topic it is aspired to spark curiosity within others to take up further research steps.

Framework for design for fairness in AI I

In future investigations it might be valuable to further validate and explore the frameworks to design for fairness in AI.

Measuring impact I It would be promising to research manners how to measure or access the impact on the ethical outcomes of projects.

Research value-tensions I It is recommended to continue research into value tensions in AI. Firstly confirm/oppose the value tensions identified in this thesis in other industries. Additionally, it is proposed to discover new value tensions in AI in other industries to develop an extensive knowledge base.

Research unfairness sources I More research is desired into the sources of unfairness in AI. A quantitative analyses of cases of ethically misaligned AI systems is proposed to detail and validate the identified sources of unfairness.

Ethical fulfillment in AI I In the designed framework the element of ethical fulfillment in AI is solely touched upon in a form of a first concept. It is advised to research this element further and discover practical ways of implementation of the ethical nudges/behavioral override strategies for more ethical (AI) development processes.

10.5

Personal Reflection

As a final note, thoughts on the personal development goals and the reflection of the project are shared.

10.5.1 Overall reflection

*‘Transforming a current state into a preferred state’
- Simon (1996)*

This quote describes a broad definition of design. My personal motivation for design is strongly connected to it, to advance the world we live in. I do not believe in utopia or dystopia, but contributing to a protopia, a state that is better than today than yesterday, although it might be only a little better (Kelly, 2017).

However, in the course of this graduation I discovered a new valuable corner of design. In which it is not concerned with design towards a preferred future state, but rather design used by means of inquiry, a form of criticism, a form of redefining the solution spaces and researching interactions. This led to rich and valuable insights which fueled the foundation of the strategic design solution.

I have truly enjoyed this project with my supervisors and experts intersected during the course of this graduation, exploring uncharted affairs. I relished diving into the intersection of the relevant subject of AI ethics and thereby joining many inspiring events and discussions. Overall, I am honored that next to research relevance of this thesis, also a design came out that currently is being developed further in the US globally.

One of the reasons I choose this graduation topic was to get more knowledgeable in AI. Thus I dared to dive into perplexing disciplines AI and ethics, I had little knowledge about. I necessitate to say it was a challenge to grasp these topics and their concepts such as fairness, deep neural

networks and value. At a certain moment during the project I dreamed away into the philosophical discussions and dialogue, trying to define words of which no agreed upon definitions exist. Taking a step back and zooming out helped me at these moments and led to re-framing of the concepts such as fairness in AI in a novel manner. However, I believe it could be prevented by scoping the topic more at the beginning. It showed me the significance of a sharp scope.

My bachelor degree is earned in at Eindhoven University of Technology. The study of industrial design there, focuses more on the creation and application of new technology (systems) in a societal context (mission statement). My master degree, strategic product design focuses on “mastering designs impact on business and markets”. Which leads to the following definition of strategic design: *“The use of design principles and practices to guide strategy formulation and implementation towards innovative outcomes that benefit people and organizations alike.”* (Calabretta, et al., 2016).

Looking back, I truly believe I complemented both approaches in this final master thesis. It focused on the creation of new technology systems (AI) in a societal context, by looking at the organizational transformation capabilities and resources at team levels. Furthermore this led to the balancing act between the use of more strict methodologies/approaches (Delft approach) and more free design approaches (Eindhoven approach), simultaneously balancing between the two what my final contribution would be.

Concluding, I think this project sincerely resembled the variety of skills and knowledge I gathered through the course of my entire studies, of which

I learned the value of both and their limitations of them. I believe these approaches therefore fruitfully supplement each other.

I experienced the value of strategic and critical design approaches in an environment which is greatly technology driven. I believe that it are these environments in which ways of working are disparate, the design approaches can bring serious benefit.

In line, I acquired new approaches how design can support ethics and thereby is a relevant approach for incorporating ethics in AI development. I am very curious to the future and discover further manners in which design can bring benefit and explore the opportunities after my studies!

10.5.2 Personal development goals

01 Corporate environment

In the course of my graduation I experienced the corporate environment both within and outside the Netherlands. I realized the advantages of a company which has a tremendous knowledge base and a surplus of highly skilled people. Next to the fact it the name IBM opened doors for interviews, web summits, meetings and events, also the people working there are willing and exited to support each other, more then expected. Obviously, also the disadvantages were experienced such as bureaucratic process and rules which in some cases slowed down the project. Nevertheless, the overall experience was a flourishing one. Due to the size of the company, also the impact of the projects can be substantial. Currently a collaboration is initiated with a design team in the United States to make this project a part of a bigger one, expand this modular workshop and ethical coach role globally. At a smaller company this would not have been possible.

02 Simplify complexity

One of the crucial challenges in the development of new applications of AI is the lack of understanding the technology of all involved (Burgess, 2017). Therefore, I aimed to give an

attempt at simplifying a technology such as AI, to communicate and understand its ethical pitfalls. I need to say it took me quite some time too grasp the topics in the field of AI well enough to simplify it in a desirable fashion. In the end I think I managed to come up with a metaphor which simply provides the understanding to the basics in a playful way. Furthermore, consolidating the literature and design research into “simple” frameworks was experienced challenging and iterated upon. However I think I managed to simplify elements of topics into relatively simple visuals. Hence, it was a thin line of communicating the depth while simultaneously do this in a simple and graspable manner. This balancing act I further aim to practice and in which I can develop myself further.

03 Translation

One of the strengths of a (strategic) designer is the ability to translate between industries, departments and people (Calabretta, 2016). Overall, I believe I found a complementary manner to bridge three main disciplines, applied ethics, AI and design.

04 Focus

I have a broad interest and see the connections between research areas and topics leading to new opportunities for interesting projects and directions. This led sometimes to side directions in the project; sometimes valuable and sometimes not. It is an characteristic I aim to develop further and as mentioned before, it showed the relevance of a sharp scope beforehand. This could have led to a more concise report as well.

Finally, I enjoyed the process together with the supervisors and IBM, which led to a fruitful collaboration. I truly hope to have sparked the curiosity and desire within others to further explore the field of AI ethics implemented by design.

Bibliography

The references used in the thesis and in the appendices are listed.

A

Ala-Pietilä, P. (2018) Europe’s AI ethics chief: No rules yet, please. Retrieved from: <https://www.politico.eu/article/pekka-ala-pietila-artificial-intelligence-europe-shouldnt-rush-to-regulate-ai-says-top-ethics-adviser/>

Agre, P., & Agre, P. E. (1997). Computation and human experience. Cambridge University Press.

Allen, C., Smit, I., & Wallach, W. (2005). Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and information technology*, 7(3), 149-155.

Arnold, T., Kasenberg, D., & Scheutz, M. (2017). Value alignment or misalignment—what will keep systems accountable. In 3rd International Workshop on AI, Ethics, and Society.

B

Barocas, S., Hardt, M., & Narayanan, A. (2018). Fairness and Machine Learning. NIPS Tutorial.

Basalla, G. (1989). The evolution of technology. In G. Basalla & O. Hannaway (Eds.), *Cambridge studies in the history of science*. Cambridge: Cambridge University Press.

Bashirieh, S., Mesbah, S., Redi, J., Bozzon, A., Szilávik, Z., & Sips, R. J. (2017, July). Nudge your Workforce: A Study on the Effectiveness of Task Notification Strategies in Enterprise Mobile Crowdsourcing. In *Proceedings of the 25th Conference on User Modeling, Adaptation and*

Baxter, K. (2018) How to Build Ethics into AI — Part I Research-based recommendations to keep humanity in AI. Retrieved from: <https://medium.com/salesforce-ux/how-to-build-ethics-into-ai-part-i-bf35494cce9>

BBC. (2018). BBC. Retrieved from BBC News: <https://www.bbc.co.uk/news/topics/c81zyn0888lt/face-book-cambridge-analytica-data-scandalc>

Balayn, A. (2018) On the fairness of crowd-sourced training data and Machine Learning models for the prediction of subjective properties. The case of sentence toxicity. To be or not to be #\$\$%*! toxic? To be or not to be fair? Unpublished master’s thesis

Banavar, G. (2016). What It Will Take for Us to Trust AI. *Harvard Business Review*.

Bennett, R. 2003. Factors underlying the inclination to donate to particular types of charity. *International Journal of Nonprofit and Voluntary Sector Marketing* 8(1): 12–29. doi:10.1002/nvsm.198

Berg, R. K. (1964). Equal Employment Opportunity Under the Civil Rights Act of 1964. *Brook. L. Rev.*, 31, 62.

Berleur, J. J., & Brunnstein, K. (Eds.). (1996). *Ethics of computing: codes, spaces for discussion and law*. Springer Science & Business Media.

Binns, R. (2017). Fairness in Machine Learning: Lessons from Political Philosophy. *arXiv preprint arXiv:1712.03586*.

Bitner, M. J., Ostrom, A. L., & Morgan, F. N. (2008). Service blueprinting: a practical technique for service innovation. *California management review*, 50(3), 66-94.

Boddington, P. (2017). *Towards a code of ethics for artificial intelligence*. Springer International Publishing.

Boer, L., & Donovan, J. (2012, June). Provotypes for participatory innovation. In *Proceedings of the designing interactive systems conference* (pp. 388-397). ACM.

Björgvinsson, E., Ehn, P., & Hillgren, P. A. (2012). Design things and design thinking: Contemporary participatory design challenges. *Design Issues*, 28(3), 101-116.

Borning, A., & Muller, M. (2012, May). Next steps for value sensitive design. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1125-1134). ACM.

Bostrom, N. (2003). Ethical issues in advanced artificial intelligence. *Science Fiction and Philosophy: From Time Travel to Superintelligence*, 277-284.

Bostrom, N. (2012). The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds and Machines*, 22(2), 71-85.

- Bughin, J., Hazan, E., Ramaswamy, S., Chui, M., Allas, T., Dahlstrom, P., Henke, N., Trench, M. (2017) Artificial Intelligence: The Next Digital Frontier? McKinsey Global Institute
- Buolamwini, J. (2018) When the Robot Doesn't See Dark Skin. Retrieved from <https://www.nytimes.com/2018/06/21/opinion/facial-analysis-technology-bias.html>
- Burns, C., Cottam, H., Vanstone, C., & Winhall, J. (2006). RED paper 02: Transformation design.
- Burgess, A. (2017). The Executive Guide to Artificial Intelligence: How to identify and implement applications for AI in your organization. Springer.

C

- Calabretta, G., Gemser, G., & Karpen, I. (2016). Strategic Design: Eight essential practices every strategic designer must master. BIS Publishers.
- Cambridge dictionary
- Carr, C. L. (2017). On fairness. Routledge.
- CBS (2008) Naar een half miljoen alleenstaande ouders in Nederland. Retrieved from: <https://www.cbs.nl/nl-nl/nieuws/2008/36/naar-een-half-miljoen-alleenstaande-ouders-in-nederland>
- Chouldechova, A., G'Sell, M. (2017) Fairer and more accurate, but for whom? arXiv preprint arXiv:1707.00046, 2017.
- Clickatell (2018) Trends in artificial intelligence technology. Retrieved from <https://www.clickatell.com/articles/technology/trends-artificial-intelligence-technology/>
- Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017, August). Algorithmic decision making and the cost of fairness. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 797-806). ACM.

D

- d'Anjou, P. (2011). An alternative model for ethical decision-making in design: A Sartrean approach. *Design Studies*, 32(1), 45-59.
- Davenport, T. H., & Ronanki, R. (2018). Artificial Intelligence for the Real World. *Harvard Business Review*, 96(1), 108-116.
- Davis, K., Patterson, D. (2012). Ethics of Big Data: Balancing risk and innovation. "O'Reilly Media, Inc.".
- Deloitte (2018) Global Human Capital Trends. Retrieved from <https://www2.deloitte.com/insights/us/en/focus/human-capital-trends.html>
- Desarda, A. (2018) Bias-Variance & Precision-Recall Trade-offs: How to aim for the sweet spot. Retrieved from: <https://towardsdatascience.com/tradeoffs-how-to-aim-for-the-sweet-spot-c20b40d5e6b6>
- Despotou, G., & Kelly, T. (2005). Using Scenarios to Identify and Trade-off Dependability Objectives in Design. In Proceedings of the 23rd International System Safety Conference (ISSC), CA USA, System Safety Society August.
- Dignum, V. (2018). Ethics in artificial intelligence: introduction to the special issue.
- Dobrin, A.D.S.W. (2012) It's Not Fair! But What Is Fairness? Retrieved from <https://www.psychologytoday.com/us/blog/am-i-right/201205/its-not-fair-what-is-fairness>
- Dorst, K., & Royakkers, L. (2006). The design analogy: a model for moral problem solving. *Design Studies*, 27(6), 633-656.
- Dourish, P., & Bell, G. (2014). Resistance is futile: reading science fiction alongside ubiquitous computing. *Personal and Ubiquitous Computing*, 18(4), 769-778.
- Durón, R. C., Simonse, L., & Kleinsmann, M. (2019). Strategic Design Abilities for Integrated Care Innovation. In *Service Design and Service Thinking in Healthcare and Hospital Management* (pp. 211-232). Springer, Cham.
- Dutton, T (2018) An overview of national AI strategies. Retrieved from <https://medium.com/politics-ai/an-overview-of-national-ai-strategies-2a70ec6edfd>

E

- Estrada, D. (2018). Value Alignment, Fair Play, and the Rights of Service Robots. arXiv preprint arXiv:1803.02852.
- Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., & Venkatasubramanian, S. (2017). Runaway feedback loops in predictive policing. arXiv preprint arXiv:1706.09847.
- Erdlyi, J. G. (2018). Regulating Artificial Intelligence Proposal for a Global Solution. AAAI/ACM Conference on Artificial Intelligence, Ethics and Society.

F

- Fast, E., & Horvitz, E. (2017). Long-Term Trends in the Public Perception of Artificial Intelligence. In AAAI (pp. 963-969).
- Fisher, E., & Schuurbiers, D. (2013). Socio-technical integration research: Collaborative inquiry at the midstream of research and development. In *Early engagement and new technologies: Opening up the laboratory* (pp. 97-110). Springer, Dordrecht.
- Fjord (2018) Computers have eyes. Retrieved from: <https://trends.fjordnet.com/computers-have-eyes/>
- Flanagan, M., Howe, D. C., & Nissenbaum, H. (2005, April). Values at play: Design tradeoffs in socially-oriented game design. In Proceedings of the SIGCHI conference on human factors in computing systems (pp. 751-760). ACM.
- Flasiński, M. (2016). Introduction to artificial intelligence. Springer.
- Friedman, B., & Hendry, D. (2012, May). The envisioning cards: a toolkit for catalyzing humanistic and technical imaginations. In Proceedings of the SIGCHI conference on human factors in computing systems (pp. 1145-1148). ACM.
- Friedman, B., Hendry, D. G., Hultgren, A., Jonker, C., van den Hoven, J., & van Wynsberghe, A. (2015). Charting the next decade for value sensitive design. Aarhus series on human centered computing, 1(1).
- Friedman, B., Kahn, P. H., Borning, A., & Hultgren, A. (2013). Value sensitive design and information systems. In *Early engagement and new technologies: Opening up the laboratory*(pp. 55-95). Springer, Dordrecht.
- Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems (TOIS)*, 14(3), 330-347.
- Frost, F. A. (1995). The use of stakeholder analysis to understand ethical and moral issues in the primary resource sector. *Journal of Business Ethics*, 14(8), 653-661.
- Fung., K. (2015) The ethics conversation we are not having. *Harvard Business Review*

G

- Gajane, P. and Pechenizkiy, M., 2017. On formalizing fairness in prediction with machine learning. arXiv preprint arXiv:1710.03184.
- Giaccardi, E., Nicenboim, I.P. (2018) Resourceful ageing: Empowering older people to age resourcefully with the Internet of Things
- Gispen, J. (2017) Ethics for designers. Incorporating ethics in the design process (unpublished master thesis)
- Gonzalez, W. J. (Ed.). (2015). *New Perspectives on Technology, Values, and Ethics: Theoretical and Practical* (Vol. 315). Springer.
- Goodpaster, K. E. (1991). Business ethics and stakeholder analysis. *Business ethics quarterly*, 53-73.
- Guston, D. H., & Sarewitz, D. (2002). Real-time technology assessment. *Technology in society*, 24(1-2), 93-109.

H

- Hamilton, A.H. (2018) Amazon built an AI tool to hire people but had to shut it down because it was discriminating against women. Retrieved from: <https://www.businessinsider.nl/amazon-built-ai-to-hire-people-discriminated-against-women-2018-10/?international=true&r=US>
- Hardt, M. (2014) How big data is unfair. Retrieved from: <https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de>

- Harris, J, Charles, E, Pritchard, M S and Rabins, M J (2000) Engineering ethics: concepts and cases Wadsworth, Belmont
- Hempel, J (2018) Want to prove your business is fair? Audit your algorithm. Retrieved from : <https://www.wired.com/story/want-to-prove-your-business-is-fair-audit-your-algorithm/>
- Horviz, E. (2017). AI, people and society. Science.

I

- IEEE (2018) Embedding Values into Autonomous Intelligent Systems. Retrieved from: https://standards.ieee.org/content/dam/ieeestandards/standards/web/documents/other/ead_embedding_values_v2.pdf
- Illari, P., & Russo, F. (2014). Causality: Philosophical theory meets scientific practice. OUP Oxford.

J

- JafariNaimi, N., Nathan, L., & Hargraves, I. (2015). Values as hypotheses: design, inquiry, and the service of values. Design issues, 31(4), 91-104.
- Johnson, M (1993) Moral imagination: implications of cognitive science for ethics University of Chicago Press

K

- Kamphuis, L. (2018) Filosofie voor een weergaloos leven. De bezige bij. Amsterdam 1ste druk
- Kirshna, A. (2019) IBM Marks More Than a Quarter Century of Patent Leadership with Record Year. Retrieved from: <https://www.ibm.com/blogs/think/2019/01/ibm-marks-more-than-a-quarter-century-of-patent-leadership-with-record-year/>
- Kluckhohn, C. (1951). Values and value-orientations in the theory of action: An exploration in definition and classification.
- Kool, V. K., & Agrawal, R. (2016). Psychology of technology. Springer International Publishing.
- Kouchaki, M., Smith, I. H., & Netchaeva, E. (2015). Not All Fairness Is Created Equal: Fairness Perceptions of Group vs. Individual Decision Makers. Organization Science, 26(5), 1301-1315.

L

- Lane, G (2018)Fairness toolkit. Retrieved from: <https://unbias.wp.horizon.ac.uk/fairness-toolkit/>
- Latour, B. 1992. Where are the missing masses? The sociology of a few mundane artifacts. In Shaping technology/building society, ed. W. E. Bijker and J. Law, pp. 225–58. Cambridge, MA: MIT Press.
- 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). ACM
- Lee, M. K., & Baykal, S. (2017, February). Algorithmic Mediation in Group Decisions: Fairness Perceptions of Algorithmically Mediated vs. Discussion-Based Social Division. In CSCW (pp. 1035-1048).
- Leonard, T. C. (2008). Richard H. Thaler, Cass R. Sunstein, Nudge: Improving decisions about health, wealth, and happiness.
- Leventhal, 1967 “The distribution of rewards and resources in groups and organizations.” Advances in experimental social psychology 9 (1976): 91-131
- van Lieren, A., Calabretta, G., & Schoormans, J. (2018) Rational Overrides: Influence Behaviour Beyond Nudging. Design research society. doi: 10.21606/dma.2018.699
- Lohr, S. (2018, Feb 9). Facial Recognition Is Accurate, if You’re a White Guy. Retrieved from New York Times: <https://www.nytimes.com/2018/02/09/technology/facial-recognition-race-artificial-intelligence.html>
- Lloyd, P. (2009). Ethical imagination and design. Design Studies, 30(2), 154-168.
- LSN Global (2018) Macro trends. Retrieved from: <https://www.lsnglobal.com/macro-trends>

M

- Macnish, K. (2012). Unblinking eyes: the ethics of automating surveillance. Ethics and information technology, 14(2), 151-167.
- Malpass, M. (2017). Critical design in context: History, theory, and practices. Bloomsbury Publishing.
- Manders-Huits, N. & Zimmer, M. (2009). Values and pragmatic action: The challenges of introducing \

- ethical intelligence in technical design communities. International Review of Information Ethics, 10.
- Mason, H., Loukides. M. (2018) Ethics and Data science. Published by O’reilly Media Inc
- Mason, H., Mattin, D., Dumitrescu, D., & Luthy, M. (2015). Trend-Driven Innovation: Beat Accelerating Customer Expectations. John Wiley & Sons.
- McCarthy, J., Minsky, M., Rochester, N., and Shannon, C. (1955). A Proposal for the Dartmouth Summer Research Project on Artificial intelligence. Technical report.
- Mckinsy report (2018) Notes from the ai frontier insights from hundreds of use cases
- Mephram, B., Kaiser, M., Thorstensen, E., Tomkins, S., & Millar, K. (2006). Ethical matrix manual. LEI, onderdeel van Wageningen UR.
- Miller, J. K., Friedman, B., Jancke, G., & Gill, B. (2007, November). Value tensions in design: the value sensitive design, development, and appropriation of a corporation’s groupware system. In Proceedings of the 2007 international ACM conference on Supporting group work (pp. 281-290). ACM.
- Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society, 3(2), 2053951716679679.
- Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). Foundations of machine learning. MIT press.
- Moses, R., Devan, P., Khan, A. (2018) Cognitive technologies: A technical primer. Deloitte insight
- Moutafi, J., Furnham, A., & Paltiel, L. (2004). Why is conscientiousness negatively correlated with intelligence?. Personality and Individual Differences, 37(5), 1013-1022.
- Mun, M., S. Reddy, K. Shilton, N. Yau, P. Boda, J. Burke, D. Estrin, et al. 2009. PEIR, the personal environmental impact report, as a platform for participatory sensing systems research. In Proceedings of the international conference on mobile systems, applications, and services, 55–68. Presented at the international conference on mobile systems, applications, and services. Krakow: ACM.
- Muller, M., & Liao, Q. V. (2017). Exploring AI Ethics and Values through Participatory Design Fictions. Human Computer Interaction Consortium.

N

- Narayanan, A. (2018) Tutorial: 21 fairness definitions and their politics. Retrieved from <https://www.youtube.com/watch?v=jlXluYdnyyk>
- NG. A. (2017) the state of artificial intelligence. Retrieved from: https://www.youtube.com/watch?time_continue=60&v=NKpuX_yzdYs

O

- O’Neil, C (2018) The truth about algorithms. Retrieved from: <https://www.youtube.com/watch?v=heQzqX35c9A&feature=youtu.be>
- O’Neil, O (2016) Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. New York Times Bestseller. New York
- Oosterlaken, I. (2009b). The capability approach, technology and neutrality towards the good life .Paper presented at 2009 annual conference of the Human Development and Capability Association, September 10–12, 2009, Lima, Peru.
- Oosterlaken, I., & Hoven, J. v. (2012). The capability approach, technology and design. Springer.
- O’Regan, G. (2008). A brief history of computing. Springer Science & Business Media.

P

- Pariser. E (2011) The filter bubble: How the new personalized web is changing what we read and how we think. Penguin.
- Parfit, D (1997) Equality and priority. Ratio, 10(3): 202–221
- Patton, M. Q. (2002). Two decades of developments in qualitative inquiry: A personal, experiential perspective. Qualitative social work, 1(3), 261-283.
- Pereira, D. (2018) Andrew NG’s “The State of Artificial Intelligence” reviewed. Retrieved from: <https://medium.com/@dpereirapaz/andrew-ngs-the-state-of-artificial-intelligence-reviewed-7007d95a72a1>
- Pick, R. (2016) Watch Google Research’s Robots Learn Hand-Eye Coordination. Retrieved from https://motherboard.vice.com/en_us/article/wnxdvy/watch-google-researchs-robots-learn-hand-eye-coordination

R

- Purcell, B (2018) The Ethics Of AI: How to Avoid Harmful Bias and Discrimination. Forrester Research
- Rahwan, I & Cebrian ,M. (2018) Machine Behavior Needs to Be an Academic Discipline. Retrieved from: <http://nautil.us/issue/58/self/machine-behavior-needs-to-be-an-academic-discipline>
- Rahwan, I. (2018). Society-in-the-loop: programming the algorithmic social contract. *Ethics and Information Technology*, 20(1), 5-14.
- Raconteur (23.05. 2018) Artificial intelligence for business. Raconteur N. 0521
- Robach, C. (2005). Critical Design: Forgotten History or Paradigm Shift. L. Dencik, Shift: Design as Usual-Or a New Rising, 30-41.
- Roeser, S. (2012). Emotional engineers: Toward morally responsible design. *Science and Engineering Ethics*, 18(1), 103-115.
- Rossi, F. (2018). AI ethics. IBM AI Academy
- Roos, T. (2018) Elements of AI. Retrieved from <https://www.elementsofai.com/faq/who-created-this-course>. Computer Science of the University of Helsinki
- Russell, S., Dewey, D., & Tegmark, M. (2015). Research priorities for robust and beneficial artificial intelligence. *Ai Magazine*, 36(4), 105-114.
- Russell, S. J., & Norvig, P. (2016). Artificial intelligence: a modern approach. Malaysia; Pearson Education Limited,.
- Ryan, A. (2006). Fairness and Philosophy. *Social Research*, 73(2), 597-606. Retrieved from <http://www.jstor.org/stable/40971838>

S

- Saffer, D. (2005). The role of metaphor in interaction design. *Information Architecture Summit*, 6.
- Sánchez-Fernández, R., & Iniesta-Bonillo, M. Á. (2007). The concept of perceived value: a systematic review of the research. *Marketing theory*, 7(4), 427-451.
- Saxena, N., Huang, K., DeFilippis, E., Radanovic, G., Parkes, D., & Liu, Y. (2018). How Do Fairness Definitions Fare? Examining Public Attitudes Towards Algorithmic Definitions of Fairness. *arXiv preprint arXiv:1811.03654*.
- Schatsky, D., Schwartz, J. (2015) Redesigning work in an era of cognitive technologies. Deloitte review
- Schmid, D. (2018) Values Build Trust: The Universal Declaration of Human Rights Secures the Ethical Use Of AI, Retrieved from <https://www.digitalistmag.com/improving-lives/2018/11/07/universal-declaration-of-human-rights-secures-ethical-use-of-ai-06192695>
- Schot, J., & Rip, A. (1997). The past and future of constructive technology assessment. *Technological forecasting and social change*, 54(2-3), 251-268.
- Schon, D. A. 1988. "Designing: Rules, Types and Worlds." *Design Studies* 9 (3): 181-90.
- Shilton, K. (2018). Values and Ethics in Human-Computer Interaction. *Foundations and Trends® Human-Computer Interaction*, 12(2), 107-171.
- Shilton, K., & Anderson, S. (2017). Blended, not bossy: Ethics roles, responsibilities and expertise in design. *Interacting with Computers*, 29(1), 71-79.
- Shilton, K., Koepfler, J. A., & Fleischmann, K. R. (2013). Charting sociotechnical dimensions of values for design research. *The Information Society*, 29(5), 259-271.
- Shrader-Frechette, K. (1997). Technology and ethical issues. *Technology and values*.
- Simonse, W. L., & Badke-Schaub, P. G. (2015). Business model design through a designer's lens: Translating, transferring and transforming cognitive configurations into action. 31st EGOS: European Group for Organisation Studies Colloquium-SGW 65, Athens, Greece, 2-4 July 2015.
- Snoek, C. "Tegenlicht meet up Mens en Machine" Interview, Tegenlicht, Amsterdam, 23rd of October, 2018.
- Soares, N., & Fallenstein, B. (2014). Aligning superintelligence with human interests: A technical research agenda. Machine Intelligence Research Institute (MIRI) technical report, 8.
- Spiekermann, S. 2015. Ethical IT Innovation: A Value-Based System Design Approach. Boca Raton: Auerbach Publications.
- Sylvester, J., & Raff, E. (2018). What About Applied Fairness?. *arXiv preprint arXiv:1806.05250*.
- Szlavik, Z. (2018) Chatbots to learn about with and from. Retrieved from <https://www.nextlearning.nl/wp-content/uploads/sites/11/2018/04/Presentatie-Zoltan-Szlavik.pdf>

T

- Taylor R. (2017) The Philosophy of Fairness. Retrieved from: <https://owlcation.com/humanities/The-Philosophy-of-Fairness>
- Teich, P. (2018) Artificial Intelligence Can Reinforce Bias, Cloud Giants Announce Tools for AI Fairness. Retrieved from <https://www.forbes.com/sites/paulteich/2018/09/24/artificial-intelligence-can-reinforce-bias-cloud-giants-announce-tools-for-ai-fairness/#11916d289d21>
- The Fairness Measure (2018) Fairness Definitions in Machine Learning .Retrieved from: <http://fairness-measures.org/Pages/Definitions>
- The future of life institute (2017) Asilomar AI principles. Retrieved from <https://futureoflife.org/ai-principles/>
- Trendwatching (2018) Five trends for 2019. Retrieved from: <https://trendwatching.com/quarterly/2018-11/5-trends-2019/>

V

- Van den Berg, J. (2018) Taida, Sahar, Mauro: hoe verging het andere kinderen zonder verblijfsvergunning? Retrieved from: <https://www.volkskrant.nl/nieuws-achtergrond/taida-sahar-mauro-hoe-verging-het-andere-kinderen-zonder-verblijfsvergunning--b691e3b6/>
- Van den Hoven, J. (2012). Human capabilities and technology. In *The capability approach, technology and design* (pp. 27-36). Springer, Dordrecht.
- Van den Hoven, J. (2017). Ethics for the Digital Age: Where Are the Moral Specs?. In *Informatics in the Future* (pp. 65-76). Springer, Cham.
- van den Hoven, J. V. (2015). *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains*. Springer.
- Van Otterlo, M. (2013). A machine learning view on profiling. In *Privacy, Due Process and the Computational Turn* (pp. 55-78). Routledge.
- Van de Poel, I., & Royakkers, L. (2007). The ethical cycle. *Journal of Business Ethics*, 71(1), 1-13.
- Varnshney, K. (2018) Introducing AI Fairness 360. Retrieved from: <https://www.ibm.com/blogs/research/2018/09/ai-fairness-360/>
- Verbeek, P. P. (2011). *Moralizing technology: Understanding and designing the morality of things*. University of Chicago Press.
- Verbeek, P. P. (2014). *Op de vleugels van Icarus: hoe techniek en moraal met elkaar meebewegen*. Lemniscaat.
- Verbeek, P. P. (2006). Materializing morality: Design ethics and technological mediation. *Science, Technology, & Human Values*, 31(3), 361-380.
- Verbeek, P. P., & Slob, A. (2006). *User behavior and technology development*. Springer.
- Vincent, J (2018) Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech. Retrieved from: <https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai>

W

- Wexler, J. (2018) The What-If Tool: Code-Free Probing of Machine Learning Models. Retrieved from: <https://ai.googleblog.com/2018/09/the-what-if-tool-code-free-probing-of.html>
- Whitbeck, C (1998) *Ethics in engineering practice and research* Cambridge University Press, Cambridge
- Van Wynsberghe, A., & Robbins, S. (2014). Ethicist as Designer: a pragmatic approach to ethics in the lab. *Science and engineering ethics*, 20(4), 947-961.

Y

- Yudkowsky, E. (2008). Artificial intelligence as a positive and negative factor in global risk. *Global catastrophic risks*, 1(303), 184.
- Yudkowsky, E (2016) AI Alignment: Why It's Hard, and Where to Start. Retrieved from <https://intelligence.org/2016/12/28/ai-alignment-why-its-hard-and-where-to-start/>

Z

- Zarsky, T. (2016). The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. *Science, Technology, & Human*

Values, 41(1), 118-132.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. (2013) Learning fair representations. In International Conference on Machine Learning, pages 325–333

Zimmerman, Michael J. (2015) “Intrinsic vs. Extrinsic Value”, The Stanford Encyclopedia of Philosophy), Edward N. Zalta (ed.). Retrieved from: <https://plato.stanford.edu/archives/spr2015/entries/value-intrinsic-extrinsic/>.

Zhong, Z. (2018) A tutorial on fairness in machine learning. Retrieved from <https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb>

Zijlema, P. “ 5 in 5: Five Innovations That Will Help Change Our Lives Within Five Years” on 1st of November 2018, IBM.

Zou, J., & Schiebinger, L. (2018). AI can be sexist and racist—it’s time to make it fair.

