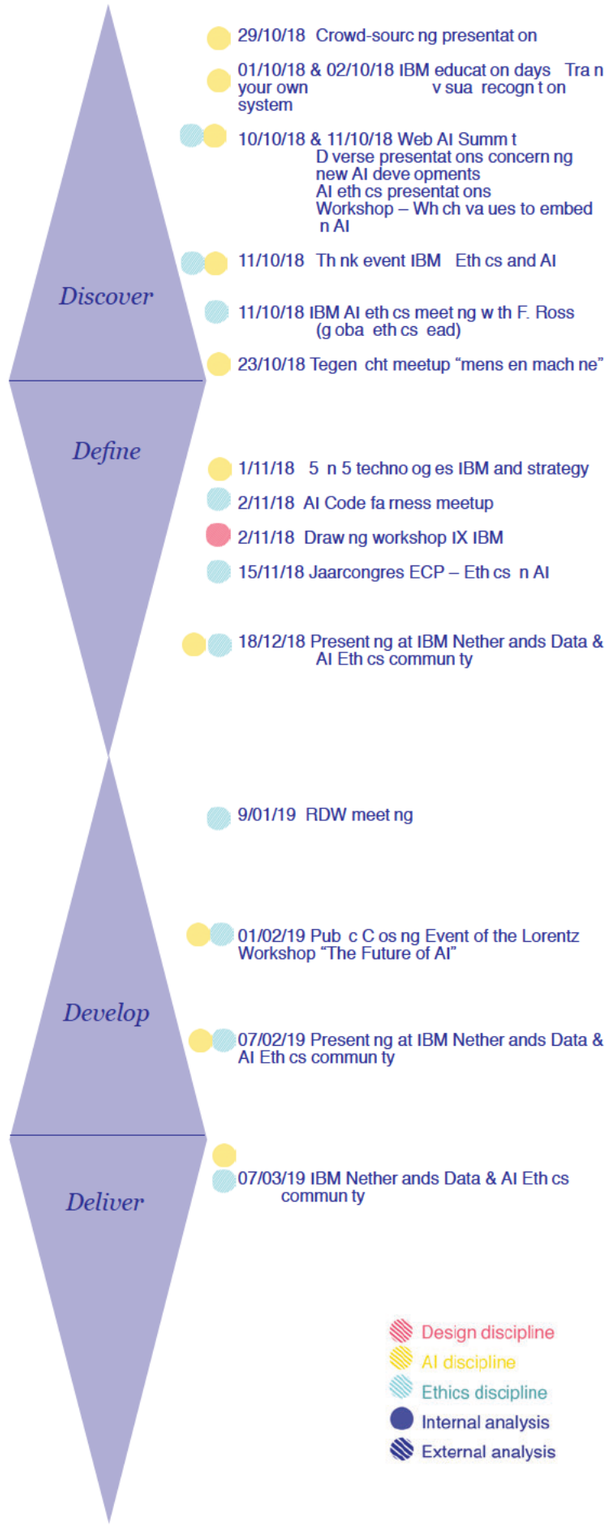
APPENDICES

Table of content

Appendix A I Overview events and expert interviews

Events



Appendix B I IBM values

IBM values, purpose & ambition

From the start IBM was initiated to save, process, tack down, analyze and pass on information. This is traceable in most of the products and services within IBM, from calculators to their AI service Watson. The overall purpose of IBM is **"to be essential to our clients and the world".**

It is divided in three values (dedication to client's success, innovation that matter for our company and the world, trust and personal responsibility in all relationship), which are described in the flowing paragraph and nine different practices

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.

The three main IBM Values:

01 Dedication to every clients' success. IBM aims to build long lasting client relationships and demonstrate personal dedication to every client.

02 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.

03 Trust and personal responsibility in all relationships. This focuses on building sustainable trusted relationships, by following words by actions.

The diverse AI services that IBM provides (Szlavik,2018)

Appendix C I Introduction to AI

Introduction to the AI field

The term artificial intelligence (AI) is brought up by John McCarthy and others in 1956. The first idea of the field arose shortly after the inventions of electronic digital computing. AI knew so called " AI winters" in which it lost interest of businesses. Burgess (2017) explains primarily caused by the disappointing results of high investments and expectations shown in the figure.

However, since 2009 the discussion of AI has increased sharply (Fast & Horvitz, 2017) leading to an increasing popularity. In 2016 AI is mentioned in twice as many articles, almost four times as many as in 2014 (Bughin et all. 2017). Now, the investments and research are rising, and it is even expected to have approximately $ 200 billion in cumulative spending from 2017 to 2021 in an array of sectors (Moses, Devan, Khan, 2018). The current popularity, after the two previous AI winters can be related to four aspects.

01 Data accessibility To train algorithms,there is a need for tremendous amounts of data, which nowadays is available. Sources differ but an approximate expectation is that by 2020 there will be 44 trillion gigabytes of data created annually.

02 Diminishing cost strorage The cost is the diminishing of storage of data. It becomes fast and the size of the machine to store it diminishes as well. This allows to actually store the data.

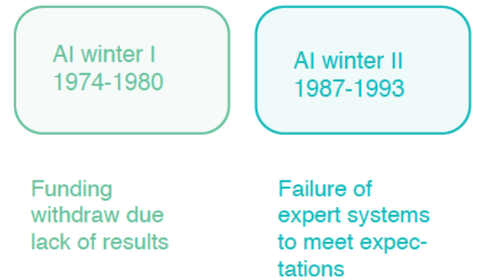
03 Faster processors. Not only we can store the data, but also faster process it. This increases the usefulness of AI systems, in development and use of its applications.

04 Ubiquitous connectivity Connectivity which is fast enough to not completely rely on the devices processor and therefore allows faster real-time processing and faster training of the AI.

These four advancements allowed for the new fast AI development and consequently initiated the current hype about AI (Burgess, 2017). However, AI keeps up being complex and difficult to study as experts have varied understandings of AI (Fast & Horvitz, 2017).

The World of AI

The world landscape of AI development shows a diversity of strategies towards AI. For an understanding of the world landscape of AI Ethics and EU's position, I have studied the worldwide strategies as input and support for determining IBM's AI ethics strategic direction.



AI winters

Currently the US has the most AI startups and investments in this field. Followed by the fast expanding China and Israel. China announced their ambition in the world of AI technology, application and research, by having one of the most complete plans for their national AI strategy. They aim to become the world's dominant player in the AI field by 2030. Chinese government has the ability to implement policies that are impossible in western cultures, due their different data policies and different notions form privacy. At the same time the US is increasing its AI know-how and investment are fast-growing.

The EU is challenged to take a different strategy to be able to compete with these two countries, as the EU does not have the US resources nor the controlling power of Chinas governments (Rossi, personal communication 10 October 2018). Francesca Rossi, the global ethical lead of IBM, proposes not copying the Chinas or American strategies but tackling it from a different perspective, in order to become the world leader in ethical responsible AI. This is expected to be a more sustainable strategy on the long run.

The EU commission recently announced to concentrate on three main pillars (Dutton, 2018).

1. Boost the EUs technological and industrial capacity for AI uptake both in public and private sectors. (The investment in AI from €500 million in 2017 to €1.5 billion by the end of 2020)
2. Prepare EU citizens for the social economic changes it brings with
3. Establish an ethical as well as legal framework by the new EU AI Alliance that aims to establish AI ethics guidelines to focus on challenges for example transparency, safety and fairness. Therefore, I identify opportunities for IBM in supporting Europe's strategy, towards a more ethical approach to AI. As

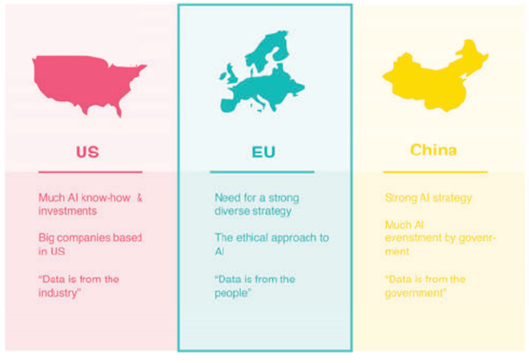
IBM has the resources and much knowledge in this field, they could gather a strong position in the EU by supporting the European Union in executing their more ethical and responsible strategy.

World of AI and the EU

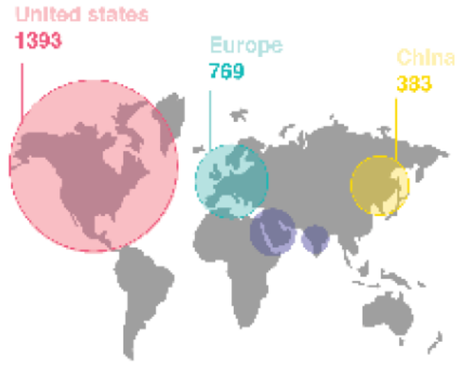
To compete with the US and China, the EU needs a more ethical strategy. I identify opportunities for IBM in supporting Europe's strategy, towards a more ethical approach for AI. As IBM has the resources and much knowledge in this field, they could gather a strong position in the EU by supporting the European Union in executing their more ethical and responsible strategy.

The EU Ethical Approach

An analysis of the worldwide strategies of AI is presented. Based on this I identify opportunities for IBM in supporting Europe's strategy, towards a more ethical approach to AI wich it aims to take. As IBM has the resources and much knowledge in this field, they could gather a strong position in the EU by supporting the European Union in executing their more ethical and responsible strategy.



World of AI Strategies visualized per country/ union



Global distribution of AI startups

World of AI startups in 2018, by Roland Berger

Thinking humanly

" The exciting new effort to make computers think.. machines with minds, in the full and literal sense."
(Haugeland, 1985)

" [The automation of] activities that we associate with human thinking, activities such as decision-making, problem solving, learning..."
(Bellman, 1978)

Thinking Rationally

" The study of mental facutlies through the use of computational models."
(Charniak and McDermott, 1985)

"The study of the computations that make it possible to perceive, reason and act."
(Winston, 1992)

Acting humanly

" The art of creating machines that perform functions that require intelligence with performed by people."
(Kurzweil, 1990)

" The study of how to make computers do things at which, at the moment, people are better."
(Rich and Knight, 1991)

Acting Rationally

" Computational Intelligence is the study of the design of intelligent agents."
(Poole et al , 1998)

" AI... is concerned with intiligent behavior in artifacts."
(Nilsson, 1998)

Various definitions of AI

Appendix D I Creative trend research

Synthesis of creative trend research

Trends

A creative trend research is conducted by me and shown in the figure. 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). The synthesis is divided into a trend hierarchy of macro trends, micro trends and actual strategic and design directions with relevance for this thesis. The topline identified macro trends are: morality recoded, anxiety rebellion, focus filter, global citizens and the new conscious. (The relevant topline micro trends at the company level are augmenting humans, democratizing AI, Niche AI, Design it all and Brand redemption.

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. This 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. Almost all severe competitors of IBM released this type of toolkit in 2018. These toolkits leave out the human aspect of AI, the human values and approach just a few sources of unfairness.

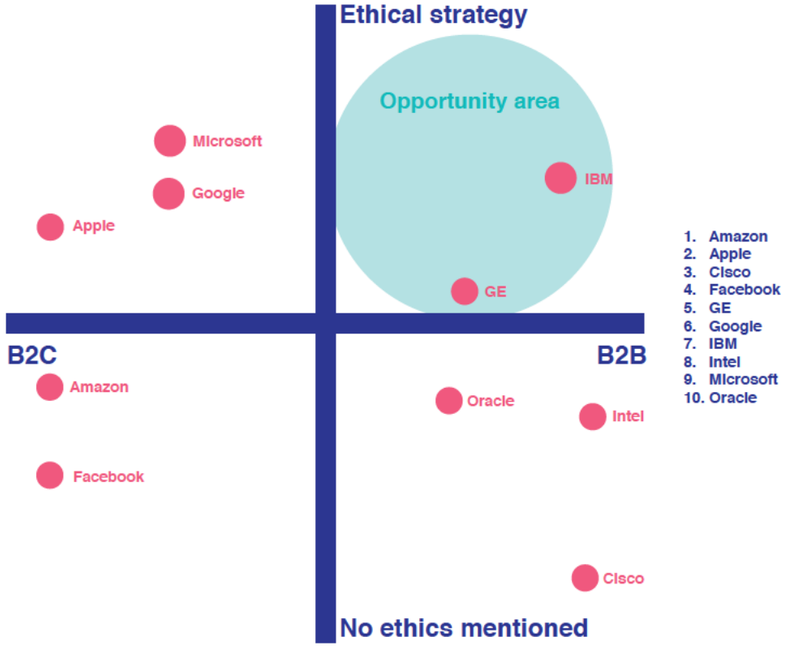
At the same time at AI events (such as the world AI summit Amsterdam 2018) and in AI strategies of companies such as Google and Microsoft, very inspiring principles toward more ethical AI development are described. Hence, the translation towards the day to day work of the AI team is lacking. Based on this trend driven innovation framework analyses, is extracted that a more human approach in the AI development process would benefit instead of a more technical one. A more bottom up approach from the people actually making the systems seems more suited. Also, aligning AI for the benefit of society and human needs is one perspective taken in this thesis

Appendix E I Competitor analyses

with the use of design methodologies and principles. **There is a need integrating societies perspective thinking about the people society and context not only from nice principles and talks but real bottom up help for the people who are making it.**
Overall, it provides strategy and design directions for my project that distinguishes IBM's approach from its competitors on the long term and aligns with human and societies values.

AI Ethical strategy

As IBM is a wide spread company with competitors in many branches, of interest for the scope is the ethical strategy of the competitors in the AI field (De Leon, 2018; Stoller,2018).



Competitor analysis

Appendix F I Fairness strategies

Example procedural fairness and outcome fairness
Specific groups of refugees got shelter in the Netherlands when their home country was unsafe. This could be for example over a course of 17 years. The children of these refugees are raised in the Netherlands. By the time they need to take their final high school exam, the country of origin is labeled as safe to return by the Dutch authority. This means due regulation; the family needs to go back to the home country, not giving the opportunity to the kids to finish high school. This leads to much debate, and is labeled as "unfair", although the rules and the process are executed "fairly" (Van den Berg, 2018).

In similar fashion, in legislation distinctions are made. For the translation of fairness into AI, it might mean one needs to look at manners to quantify it. For example, in legislation attempts are given to define what is not fair, for example anti-discrimination laws prohibit unfair treatment based on sensitive attributes (as race) (Equal Employment Opportunities, 1964).
These types of laws evaluate fairness of decisions processing two aspects: disparate treatment and disparate impact (Barocas & Selbst, 2016). There is a case of disparate treatment when decisions are based on an individual/group sensitive attribute (i.e. gender, race). It suffers from disparate impact when the outcomes hurt people with certain sensitive attributes. This is in line with the discussed process and outcome of fairness.
If there would be two very homogeneous groups and predictions would be made with the groups separately, then accuracy of the model might be in line with the fairness of it (Hardt, 2014). However, when they are in one group the classifiers from, for example, an Expectation-Maximization algorithm, it means that minorities are considered in unfair manners as deviations from the norm, when one aims to increase accuracy of classification it leads to under-appreciation of minorities. This is considered as an under-appreciated source of unfairness (Hardt, 2014).

3.5 Strategies for fairer AI

This section elaborates on the current approaches towards fairer AI (most of them released in 2018) The approaches are analyzed and the overlapping strategies extracted. This leads to the following four identified strategies in the field. Nevertheless, after the literature review, fairly little work is found on actual implementation on a day to day basis of fairness in AI development.

I Control

Due the ethically misaligned products on the market, people saw new business opportunities. Audit for AI is performed by some companies over the world (Hempel, 2018; Ghani, 2018). Hence, this acts as an afterthought, rather than at the beginning of the process. While in ethics literature prevented action early in the process is supported. Additionally it bears extra costs by changing the model at the end instead of the start.

II Code fix

- IBM fairness toolkit - code
- Google what if toolkit – code

Although multiple technical fairness" toolkits exist, these two are open source. Google's released it as a new feature of their TensorBoards web application. It gives its users the power to analyze a machine learning model, without coding. It creates an interactive UI which visually shows results of editing the model and diverse classification thresholds which account for numerical fairness criteria (Wexler, 2018). Fairness360, IBM's open-sources toolkit of metrics is also released this year to check for undesired biases in both data sets as well as the models themselves. Additionally, new algorithms are provided to ease the biases in one's model (Varnshney, 2018), These toolkits serve as starting point of working on implementable solutions for fairness in AI.
Albeit, 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.** The toolkits tackle the problem from a technology perspective and do not take context specific fairness and many of the identified unfairness sources into account.

III Reminders & checklists

Within the discipline of data science a few first support tools for fairer AI are created. (Mason & Loukides, 2018). These take the form of general checklists. These are generally applicable but loose richness in the extremely relevant and subtle context specific values and attributes.

IV Awareness & dialogue

Fairness toolkit by Probosics University of Oxford (Lane, 2018) is a physical toolkit analyzed for this thesis. It is created from a research design perspective to increase awareness and raise dialogue concerning bias, trust and fairness in algorithms. An interesting aspect is the closing of the gap between the understanding of the users and the actual algorithms. Thus, it is a very promising attempt to create awareness. Nevertheless, the day to day application of ethics in AI which is desired, is not tackled by this toolkit. Additionally, as far known this toolkit is limited tested.

Fairness360, IBM's open-sources toolkit of metrics (Varnshney, 2018)

Appendix G I Specific value tensions in AI development

I Accuracy vs Fairness

Accuracy I the degree to which the result the model conforms to the correct value or a standard.
Fairness: is 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.

Why is it a tension? As mentioned in the fairness section, AI systems can be unfair. Currently, the performance of algorithms is evaluated by comparing the algorithms outputs (the dish) and the expected outputs on a data set, representing this in a metric such as error. Albeit, these types of metrics are not taking the systems fairness into account (Chouldechova et al., 2017). Comparing two algorithms (appliances) using a general matrix for accuracy, even when output is very similar to each other, the fairness of the outputs can be extremely distinctive (Chouldechova et al., 2017).
"statistical patterns that apply to the major- ity may be invalid within a minority group." When making systems for the majority, these systems are " more accurate" (Hardt, 2014). In context where the use of sensitive attributes may be permitted, it is important to understand the implications that this choice has for fairness (Hardt, 2014; Chouldechova et al., 2017)

In some cases, accuracy of the model is very important and the fairness aspect less (ex. in medicine for example the discrimination could be made between man and women towards different treatments). In other cases it might be a more difficult trade-off (ex. a case of insurance companies predicting fraud one wants to be accurate bit simultaneously not discriminate between races). Thus there is a necessity for resolving this value-tension context specific, in order to create more fair AI systems.

II Explainability vs Performance

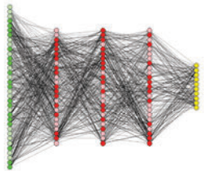
Explainability I The capability of the model to be understood, the model being interpretable and make the way it works and makes decisions understandable.
Performance I an action/process how well somebody/something carries out work or an activity. In this case the model, so for example how accurate, fast it preforms the tasks if that is demanded.
Why is it a tension? In AI systems these two are usually at odds with each other Many of the best-performing models (viz. deep neural networks) are black box in nature (Dhurandhar, 2018). When deep learning, " learns" it identifies patterns from the data and information it has access to. It uses for example neural networks and can quickly resemble a tangled mess of connections that are nearly impossible for analysts to disassemble and fully understand. In some cases, when decisions are made with real-world impact, an explanation is demanded for a fair perception and assessment if the model does not take into account sensitive attributes (such as: race, gender).

III Bias vs variance

Statistical bias is a feature in statistics, in which results (the predicted quantitative parameter) differ from the expected value. In other words, "The inability of machine learning techniques to capture the true
relationship is bias" (Desarda, 2018).
Variance, in statistics, is the expectation of the squared deviation of a random variable from its mean (Desarda, 2018). In other words, it measures how far a set of (random) numbers are spread out from their average value. Most of the time a data scientist strives for a low bias overall. But a model with high variance pays much attention to the training data. It has difficulties in generalizing based on new data. Thus, these models have a high error on test data while performing good on training data. This is a challenge as when increasing the bias decreases the variance and the other way around. It is difficult to find a balance between the two, to minimize the total error.

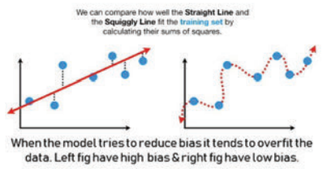
Examples of low-bias machine learning algorithms: Decision Trees, k-Nearest Neighbors and Support Vector Machines.

Examples of high-bias machine learning algorithms: Linear Regression, Linear Discriminant Analysis and Logistic Regression
Examples of low-variance machine learning algorithms: Linear Regression, Linear Discriminant Analysis and Logistic Regression.
Examples of high-variance machine learning algorithms:



visual of a deep neural network connections

bias vs variance



Appendix H I Philosophy of technology

Philosophy of technology

Introduction

The roots of philosophy of science go back to the ancient Greeks and Romans as Aristotle and Plato. Philosophy is a training to perspectival flexibility. It challenges people to critically to look at apparently obvious ideas and actions. It does not give any answers however one can find a type of language of thoughts which might resonate. Essentially philosophy helps people in the capability understanding others, to question one's own views and be open for new ones (Kamphuis, 2018). Therefore, some philosophical views will briefly be described concerning ethics in philosophy and technology philosophy, for a greater understanding of the further reasoning in this thesis. It is not meant to give a complete overview of the existing literature but shed a light on the insights one can take with them into the AI development.

The last decades have seen a great technological development (Gonzalez, 2015; van den Hoven, 2017; Horviz, 2017). This impacted our thinking in diverse ways as well as our society and the way of living, as technology can be all-pervasive and ubiquitous (Van den Hoven, 2012). This is giving the feeling that humans are living in natural environments because of which humans tend to forget the fact that almost everything around us is artificially produced (Kool & Agrawal, 2016; van den Hoven, 2017). It is easy to dismiss form our mind that practically all products are used today are artificial and humans have difficulty to see how these artifacts shape our life's.

Some views of technology philosophy will be touched upon to broaden perspectives on technology in terms of this thesis. Specifically looking at the literature of philosophy of technology one of the main questions is the impact of technology upon the human race (Kool & Agrawal, 2016). This question became more prominent in the 20th century with the well-known technology philosophers: Martin Heidegger, Arnold Gehlen, Lewis Mumford, Jacques Ellul and Albert Borgmann, Don Ihde, Bernard Stiegler, and Bruno Latour. Some interesting and relevant viewpoints will be touched upon to broaden our view on how AI might impact humanity and vice versa. The lens of the research questions in mind is used.

Human technology relationships

Not only humans influence the technology when creating it, similarity, when the technology is released into the market, it influences our values and morals. Central to this thought, are the created relationships between the world and the human, by technology. When a technology is used, it functions as a medium between its user and the context (Verbeek, 20014). Don Idhe was one of the first to describe it in a systematic manner. In the figure the different relationships one can have according to the mediation theory of Ihde with elaboration by Verbeek are shown. This framework allows us to mediate between concrete technology with humans' actions inter-operations and experiences. In relationship to AI it is interesting to analyze which relationships it can influence or create, which might be relevant to take into account when designing new AI. It offers a framework to systematically account for the technology impacts in our lives. Firstly, people creating new technologies should be aware. Second, the developers can actively use this theory to make moral technologies in a moral beneficial way. On the other hand, it raises questions how far the creators of the technologies should go in moralizing technology and how to balance with social values and autonomy (Van den Hoven, Vermaas, & Van de Poel, 2015 p.236). Verbeek (2014) argues that these relationships also give humans overconfidence in technology and in other cases it makes humans scared of the new technology (see figure). These two streams are very clearly seen with the relationship with AI, some companies and people perceive it being scared of a future with AI, sketching dystopian visions. Contrastingly, other companies and people are overconfidently talking about AI technologies, creating the earlier mentioned hype with over expectations. Verbeek mentions that the moral should guide our technological development, walking the path hand in hand, for a beneficial, sustainable path.

Value laden technology

"We shape our dwellings and then our dwellings shape us."- Winston Churchill said in a speech in the House of Commons on October 28, 1944(3)

This statements resemble that not only humans create artifacts but also the artifacts, when in context and use, influence our behavior, our ways of thinking, norms morals and values. And with the technological advancement in the last decade we entered a new phase of digital shaping of society (van den Hoven,

2017). Correspondingly, it is appointed that technology is value-laden instead of value-free (Gonzalez, 2015). In technology ethics the designer choice, embeds (consciously/unconsciously) his/her values in the technology created. Then it transfers the values of humans (imperfectly) to the designed technologies (Fleischmann, 2013). From this perspectives values can be seen as properties of systems.

Thereupon technology is morally laden, due the people making it (Verbeek, 2011). In this respect one could say that technology is directly connected to ethical values and therefore ethics. This means also towards AI systems; this view can be applied. This leads to the perspective that AI systems are value-laden, starting to digitally shape our society. Unconscious embedding of values and morals in AI systems, might lead to undesired consequences for our societies. Thus, AI teams should become aware of this and aim to prevent the undesired outcomes. Contrastingly, researchers argue that people should not be moralized but technology should be. Latur believes that artifacts can help to shape human behavior. Artifact have so called “scripts”, prescriptions how to act, the same as one would do with acting in a movie (Latour 1992). These forms of scripts can be seen as a type of moralization and can reinforce moral decision making. Latour’s view shows us that not only people can answer the question of morality of how to act, but artifacts can too (Verbeek, 2005). This leads to questions if AI agents can make moral decisions for us? And can a moral AI system be made?

Levels of ethics analyses

Gonzalez discusses based on the work of Shrader-Frechette, three different levels of analysis of ethics in technology: general, specific and related to agents (Gonzalez., 2015). The general analysis type is relevant for any technology type. The specific analysis, takes a specific technology in a specific domain and the ethical problems that occur. The technological agent related level of analysis takes into account the ethical values used by them as criteria of what is worthy, as well as what ought to be done, taking analysis beyond the current morals to offer a future ethical proposal.

Additionally, there are two other distinctions of analyses of ethics of technology: endogenous ethics and exogenous ethics (Gonzalez., 2015). Endogenous ethics analyzes knowledge, human undertaking, artifact and product. This perspective is focused on aims, processes and results in technology. Albeit, exogenous ethics is focused on contextual aspects of the human activity in a social milieu, taking into account socially assumed or institutionally ethical values. The dimension of technology are persons/groups aiming to transform society/artifacts of social purposes. This might be acceptable in a specific milieu or not. From a more dynamic perspective, one takes the historical context into account. This means there are next to the ethical judgement itself also distinctive ways to analyze ethics of technology, leading to different outcomes. Currently for AI development I seems there is a lack of exogenous ethics both in training and development. The AI development process might benefit of both, nevertheless exogenous ethics seem missing.

Conclusion

To concluding from the philosophical theories, one can say that the designers and engineers (un) consciously design with their values and morals, thus the technology they are developing reflects that. 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). Additionally, different ways to analyze ethics in AI are elaborated upon and allows to see the bigger current gaps within the AI field, the exogenous perspective. In line, he mediation theory might support a more systematic manner to access the impact of technology on humans lives. Nevertheless, philosophy leaves us with new questions rather than answers. In the following section will ethics will be elaborated upon, how to make ethical decisions and what kind of capabilities does a company need.

3.2.2Artifacts & Scripts

Achterhuis, teaches us that people should not be moralized but technology should be, when he further argues upon the idea of Bruno Latour. Latur believes that artifacts can help to shape human behavior. Artifact have so called “scripts”, prescriptions how to act, the same as one would do with acting in a movie (Latour 1992). These forms of scripts can be seen as a type of moralization and can reinforce moral decision making. Latour’s view shows us that not only people can answer the question of morality of how to act, but artifacts can too (Verbeek, 2005). This leads to questions if AI agents can make moral decisions for us? And can a moral AI system be made?

Levels of analyses of ethics philosophy of technology

Gonzalez discusses based on the work of Shrader-Frechette, three different levels of analysis of ethics in technology: general, specific and related to agents (Gonzalez., 2015). The general analysis is relevant for any technology type. The specific analysis, takes a specific technology in a specific domain and the ethical problems that occur. The technological agent related level of analysis takes into account the ethical values used by them as criteria of what is worthy, as well as what ought to be done, taking analysis beyond the current morals to offer a future ethical proposal.

Additionally, there are two other distinctions of analyses of ethics of technology: endogenous ethics and exogenous ethics (Gonzalez., 2015). Endogenous ethics analyzes knowledge, human undertaking, artifact and product. This perspective is focused on aims, processes and results in technology. Albeit, exogenous ethics is focused on contextual aspects of the human activity in a social milieu, taking into account socially assumed or institutionally ethical values. The dimension of technology are persons/groups aiming to transform society/artifacts of social purposes. This might be acceptable in a specific milieu or not. From a more dynamic perspective, one takes the historical context into account. This means there are next to the ethical judgement itself also distinctive ways to analyze ethics of technology, leading to different outcomes. Currently for AI development I seems there is a lack of exogenous ethics both in training and development. The AI development process might benefit of both, nevertheless exogenous ethics seem missing.

Agentive amplifiers

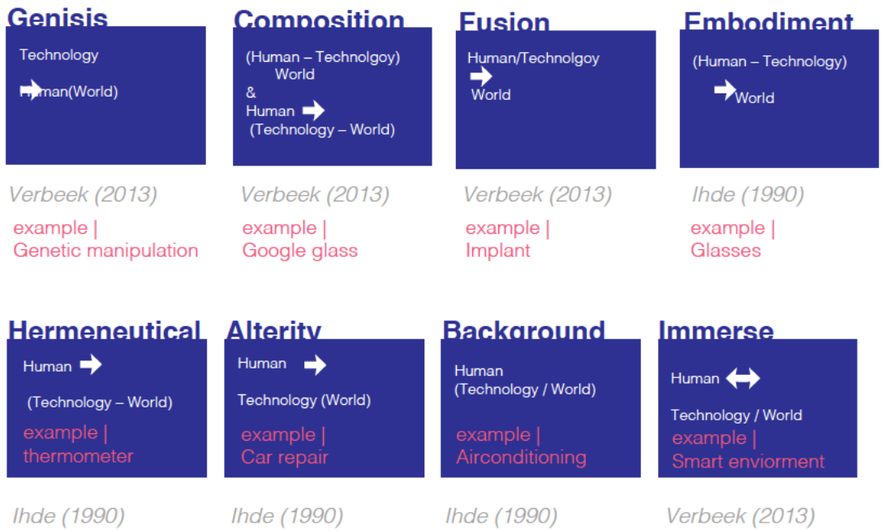
The well-known Spanish philosopher Ortega, argues that technical artifacts can be conceptualized as “agentive amplifiers”, creating opportunities that would have been impossible without them (Oosterlaken & Hoven, 2012). He argues technology is “contributing to people’s capabilities to lead flourishing human lives.” It is crucial to point out that humans can do without technology, however humans would be cold and hungry etc. Every new artifact a homo sapiens made was introduced with the goal to make the world a better place to live for him/her (less cold etc.). Otega argues: “the terminus ad quem of technology is there the good life”. According to Basalla, this suggests that the array of technologies is representing the distinctive visions of a good live (Basalla, 1989). A difference pointed out by Oosterlaken (2009), between technical artifacts and other varieties is that it is integrated in a use plan. This includes the require actions performed by the user to reach a certain goal. Also, Jeroen van den Hoven argues that technical artifacts and devices he sees as agentive amplifiers. This influences the technology assessment, namely to evaluate in the quality of contribution to flourish human lives. Thus, looking from this perspective, the assessment of AI systems could be in very distinctive manners. One with the view of Otrega in mind would lead to the assessment if it is contributing to the flourishment of human lives?

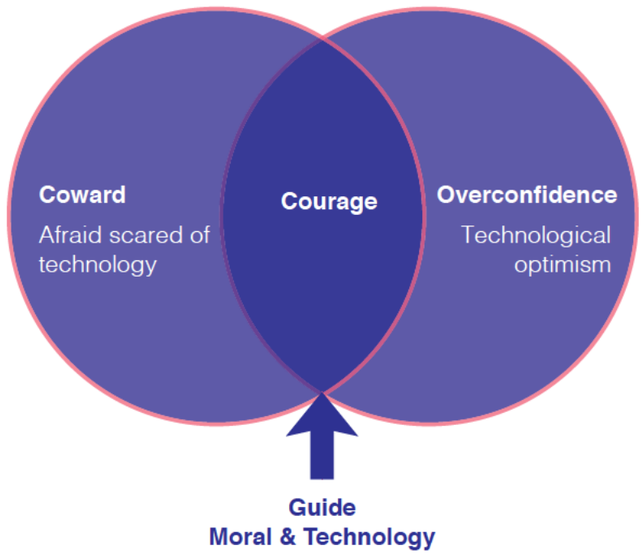
Mediation theory

P.P. Verbeek (2014), points out another relevant aspect at the intersection of humans and technology. Not only humans influence the technology when making it. When the technology is released into the market it influences simultaneously our values and morals. It is based on the mediation theory of Don Ihde. Central to this thought, are the created relationships between the world and the human, by technology. When a technology is used, it functions as a medium between its user and the context (Verbeek, 20014). The different relationships one can have according to the mediation theory of Ihde with elaboration by Verbeek are shown. This framework allows us to mediate between concrete technology with humans’ actions interoperations and experiences. In relationship to AI it is interesting to analyze which relationships it can influence or create, which might be relevant to take into account when designing new AI. It offers a framework to systematically account for the technology

impacts in our lives. Firstly, people creating new technologies should be aware. Second, the developers can actively use this theory to make moral technologies in a moral beneficial way. On the other hand, it raises questions how far the creators of the technologies should go in moralizing technology and how to balance with social values and autonomy (Van den Hoven, Vermaas, & Van de Poel, 2015 p.236).

Relationships human - technology



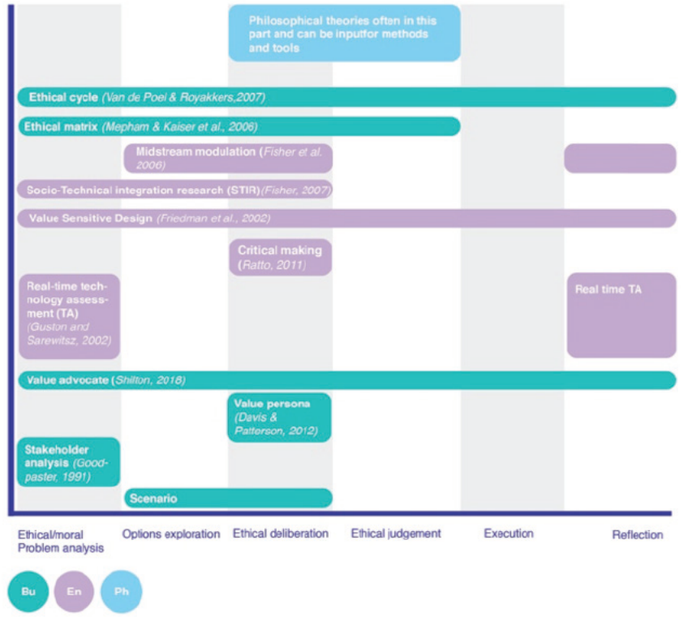


Verbeek, P. P. (2014).

Appendix I | Ethics tools & method review

In the figure the different analyzed theories are mapped against a standard process to see when they are most valuable and for which goal of the ethical decision-making process they are focused. When designing ethical support for the AI team this analysis will fuel the link to the process as in some phases it appears there is a stronger support needed for a structured argumentation while in other parts new perspectives and new critiques are supported.

Figure shows the analyzed tools mapped to the ethical decision making process.



Ethics tools in detail

1 Ethics tools/methods general

Stakeholder analysis

One of the most well-known approaches for ethical decision making is the stakeholder analysis. Business Management ethics suggests including all relevant stakeholders as well as include them in the decision-making process (Goodpaster, 1991). The foundation of this tool is to identify the relevant stakeholders, empathize with them and include their opinions and consideration into the process. Also, stakeholders input can be used for understanding of the moral issues in a field (Frost,1995). This tool is often a part of further discussed methods and processes for more ethical decision making and is advised in early stages of the process. Also, templates to guide this mapping exist to assist the process, focusing on visualizing and distinguishing the different types of influences, relationships etc. Thus, the incorporation of stakeholders (opinions) in the development of AI might shed a new, ethical, light on its consequences. The current state of this incorporation within IBM will be researched due empirical study.

Ethical cycle

The aim of the ethical cycle is to assist a structured way to address moral problems, iteratively. It is based on the opinion that moral challenges are complex and fuzzy, therefore cannot be described beforehand (Whitbeck, 1998). In other words, the outcomes are mostly provisional. It takes the standpoint that applied ethical theory is also relevant to identify and formulate the moral challenges however also judging them, using them as a heuristic tool. Van der Poel mentions that 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 needs to be articulated. In the problem analysis phase, it is important to map the stakeholders and their interest, the moral values and the relevant facts. In the third step, creativity is of great importance to broaden the solution space. In the fourth step of ethical judgement the formal or informal manners described in the earlier section are chosen and applied. The goal of last step, reflection is to come to a well-argued choice, it is a process of getting to a mutual balanced decision. Criticism is supported by van de Poel (2007) on two levels, on the ethical framework used as well as the concrete situation and action. This method shows the importance of clearly describing and communicating the challenges of a moral problem (for example in AI) to be able to solve it. In line with previous tool it incorporates stakeholders' interests. The quality of the outcome is highly depended on the creativity used during the process. Therefore, in this thesis is argued that a designer perspective will beneficially support a more ethical AI development. Additionally, reflection on the decision and the teamwork are of great importance for the quality of the outcome.

Ethical matrix

The founder of the ethical matrix Mepham in 1994, proposed it as a methodological way to the development of principles, fueled by common morality. The aspiration of the tool is to assist users in the identification of ethical issues with the rise of new technologies, arriving at intellectually defensible decisions (Mepham & Kaiser et al., 2006). As most all of the mentioned tools, it does not lead to one particular answer after using it. It starts with ethical deliberation " i.e., a process which entails the careful consideration and discussion of the ethical implications of an issue" (Mepham et al., 2006). Incorporated are different stakeholders with their particular perspectives as well as the different concerns the technology has, will be analysed (i.e. ethical principles). The principles are chosen with the different perspectives of the stakeholders. The approach considers the principles with hierarchy, some should be decided upon with more importance than others, based on evidence (i.e. scientific/economic data, assessments of consequences assessments of intrinsic values, tacit, folk or practical knowledge). Then the assessments of the impacts are put into the ethical matrix, leading to roadmap of ethical judgements made (qualitative or quantitative). In the third evaluation phase, consists of the current situation and the future desired one.

2 Ethics in engineering

Specifically, for engineering several ethical tools, approaches, methods are developed. With the lens of the research question in mind some of this will be shortly described.

- Critical Capability Approach of Technology (CCAT)
- Design for values approach
- Value sensitive design
- Constructive/real-time technology assessment (CTA)
- Ethical system development life cycle
- Socio-Technical integration research (STIR)
- Critical technical practices/reflective design/critical making
- Contextual value methodologies
- Value advocate
- Value levers

Constructive/real-time technology assessment (CTA)

Real-time technology assessment builds on constructive technology assessment but performs the assessment cooperatively with design teams during technology development (Guston and Sarewitsz, 2002). The basis of real-time TA compared to TA is that it meets ethical problem during the process instead of just assessing the impact after the technology is already in use (Rip et al. 1955). Real-time TA demands embedded social scientist or policy experts for four tasks: research historical case studies on analogous technologies, identify stakeholders, empirically document the attributes and perceptions of stakeholders and analyze and assess technical decisions in light of stakeholder needs and values (Guston and Sarewtiz,2002). It aims to put the project in social historical context and it aims to include more aspects and more actors in an early stage which Schot and RIP argue to realize better technology in a better society (Schot & Rip 1997). Which is unique in alterations of this approach later on, is that also technology developers are addressed instead of the government. This allows to make the technology assessment proactive and anticipatory (Van den hoven, 2015). In AI development this might increase the assessment quality as well as democratize the process more. Currently few people understand the AI development process. Bridging both ethicists into and AI into ethics might support

Socio-Technical integration research (STIR)

STIR uses a structured decision protocol to help humanists embedded in technology design teams to conduct collaborative inquiry (Fisher et al., 2013). Ethical reflections, sustainability and democratic governance are at its foundation. STIR researchers guide design through semi-structured interview protocol intended to bring to light decisions about opportunities, technical considerations, alternatives and outcomes (Fisher, 2007). At heart of this approach is to ask designers to describe their decisions, not changing them. This increases reflexivity about what they decide.

Midstream modulation

The framework midstream modulation is based on STIR. It is a framework for intervention-oriented activities to improve and make clear the "responsive capacity" of laboratories concerning the bigger societal dimensions (Fisher et al. 2006). It aims to support research participance to critically reflect on their work with the broader socio-ethical context. (Shuurbiers,2011). To reach this first order reflective learning ("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). Engagement tools for feedback, discussion and exploration of the decisions are used to reach this.

Critical technical practices/reflective design/critical making

In contrast with VSD, critical technical practice, reflective design and critical making, critically question the whole enterprise of a technological trajectory instead of empathizing design for specific values (Shilton, 2018). The founder of critical technical practise is Agre (1997), who points out that space for critical reflection is beneficial for the following reasons. It supports technical fields to evaluate their research, allows space for moral and ethical discussions, and encourage integration of knowledge form other fields. Specifically, to AI to push the boundaries of what counts as learning or knowledge. Critical technical practice requires questioning the metaphors, forms of representation and discourse of an entire field (Agre, 1997).

Reflective design

Reflective design asks what theoretical and methodological commitments values and assumptions underlie in HCI as a field (Dourish et al., 2014) or are appropriated during the process of designing (Sengers et al., 2006). The aim is to, by both designers and users, to identify and subvert limitations and to center values or assumptions previously left at the margins of design. Techniques as interpretive flexibility and technology as a probe Sengers et al. explore " un-designed" for spaces and values, such as social experiences at

an art museum.

Critical making

Instead of critique making a part of design, critical making uses design to conduct critique. It construes the material work of design itself as a practice that can help us question fields, disciplines and technological trajectories. Ratto (2011) argues we should experience technology itself instead of only describe it which might lead to mischaracterization of technology. It uses the generativity of engaging with material production to improve technology critiques. This can take form in a workshop.

Value advocate

Much literature points out that giving a team member the responsibility explicitly of ethics and values during the technology development process, has benefits for a 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 currently identified benefits are: can bring deep knowledge of interdisciplinary literature of ethics. Second, it can provide an outsider perspective and break group biases, creative thinking. 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). Lastly, value consciousness an explicit responsibility of the design, in helps to build values reflection into the scope of work and the success metrics of a team (Shilton,2018). However, there are also downfalls. 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. Second, responsibility on a single person may put a stronger emphasis on putting his/her values in the design process, therefore ethical pluralism is advised (Borning and Muller, 2012). Third, in real life commercial setting it is not always feasible to hire an extra person full-time.

A Sartrean model

This ethical decision model is developed for design following Sartrean line of thought. It is founded on that ethical decisions are not found through the use of formal ethical judgment theories, but it puts its focus on the designer's responsibility and freedom, as well as the practical limitations of the situation. In Sartrean ethics, freedom is an important ethical value (d'Anjou, 2011). The model consists of five phases of which the first one is about accepting one's complete freedom and its responsibility. Reflection concerning the prior design choices, after which is reflected upon the external demands. Fourthly, it reflects on the practical limitations, and lastly acting upon one's choice reflecting the conscious freedom and responsibility (d'Anjou, 2011). Thus, the main focus is awareness and reflection.

	Manner to deal with value tension	Translation values	Accounted for which values	Process phase	Other Insights	Drawbacks
Design for values (Van den Hoven, Vermaas, & Van de Poel, 2015, p 838).	No clear support for value tension	The abstract level (highly abstract statutes of a system, not yet contextual), the concrete level (specific model components in terms of concrete functionality) and the implementation level (system components as the basis for implementation)	Allows different prespectives for design. For example a more participatory design one for end user values or more designer driven design as with the VIP method. However it does not give the tools or handles to choose/account for wich values	Early in the process	Three main activities namely: election of values, development between business and modeling views (domain specific) and the execution one	Method procedures for the design are lacking and relation between views is lacking.
Value sensitive design (Friedman et al., 2013).	In the empirical investigation questions are researched as. How do they prioritize competing values in design trade-offs'?'	The different methods in conceptual, empirical and technical phases aim to also (partially) bridge the translation	Originally for a set of moral values and how these are affected (also towards direct and indirect stakeholders) Human welfare, ownership and property, privacy, freedom from bias, universal usability, trust, autonomy, informed consent, accountability, identity, calmness, and environmental sustainability (Friedman and Kahn 2003)	-	the three complementary viewpoints might be of value in value-alignment in AI. The view of direct and indirect stakeholders is relevant to prevent unwanted consequences and address desired values. Also, to research the deeper meaning of values in philosophy and that it is beneficial to research value trade-offs empirically can be extracted from this method.	Much critique on universel values. Researchers overclaim knowledge and authority over the informants and a lack of attention to suttle differences of designer. Most work is focussed on already built technologies instead of building new ones.
Value dams & Flows (Miller et al. 2007)	Features that are experienced as problematic are avoided and discussed and systematic manner to address the value-oriented design tradeoffs.	Trough context analyses value decomposition with a user and societal view of value. Then every value dam leads to a practical solution	All stakeholders	Early in the process but after value elicition	This tool shows us a way to translate values into functorialities as well make sure that conflicts are discussed. It makes stakeholder value conflicts explicit which is quite unique	It does not necessarily account for the designers' own values. Additionally, it is little tested, and the case study mentions a need for involvement of more indirect stakeholders as well.
Values at play (Flanagan et al. 2005)	Identifying values-based conflicts and during implementation and prototyping close attention is paid to the value conflicts generated per functional component	Through context analyses value decomposition with a user and societal view of value. Then every value dam leads to a practical solution	In values discovery phase relevant values from diverse sources are included. Namely: project goals and making hypotheses, earlier work, designer values, user values and other stakeholder values. It is remarkable from this method, that they explicitly use the designer's values, as often that is lacking in methods.	Covers process of game design	This process teaches us the different levels of values that can be integrated as well as the perspective of looking at value-based conflict in an iterative, functional and context specific perspective, making this trade off explicit.	This method is only used in research context and was tested with game design
Envisioning card (Friedman, & Hendry, 2012)	It does not give concrete handles how to deal with discussion or the value-conflicts however it provokes discussion	-	Stakeholders of which one would not think of in the first place as well as own values	Early in the process	The cards have four so called "envisioning-criteria" namely stakeholders, time, values and pervasiveness,	Mostly used for educational purpose. And is more a tool for discussion then one that helps on a daily work basis.

lit is clear support for more value-aligned AI development is needed. Next to new regulations for more ethical AI, also tools, methods or any other forms of support are thought about and researched to provide a well-founded basis of the current research and practice field and identify where it needs more support.
In AI creation we need to take into account also (social) values moral consideration, with the priorities of values by the different stakeholders in diverse multicultural context while still explaining reasoning and guarantee transparency (Dignum, 2018).

The following tools and methods are found in literature:
•Value elicitation (Van den Hoven, Vermaas, & Van de Poel, 2015).
•Value sketches (Woelfer et al. 2011)
•Value dams and flows (Miller et al. 2007).
•Value scenario (Nathan et al. 2007)
•Value levers (Shilton, 2018)
•Envisioning cards (Friedman & Hedry 2012)
•Value-Sensitive Action-Reflection Model (Yoo et al. 2013).
•Value sensitive design (Friedman,et al., 2008 till Friendman and Hendry, 2012)
•Values at play (Flanagan , Howe, Nissenbaum 2005)
•Design for values (Van den Hoven, Vermaas, & Van de Poel, 2015).
•Value Personas (Davis, 2012).
•Society in the loop (Radhwan, 2017)

This list contains diverse approaches towards the problem of value-aligning from more generative to engineering based ones. However, the current tools, methods and approaches little to no evaluation of the toolkits beyond academic setting (Miller et al. 2007; Shilton, 2018) or bear still much critique. Additionally, is discussed the field is still at the beginning of systematically thinking about design and values (Flanagan et al. 2005). As well as few practical methods address value tensions among diverse values (Miller et al. 2007).

For the scope of this thesis will be looked at five methods that have a stronger link toward resolving value tension, tradeoffs or focus on the translation phase of these values which are on a fuzzy abstract level towards practical day to day work of the AI development team. Therefore, a light is shed on the following approaches/ tools/methods as well as they will be described through the above described lens.

Design for values in ICT

(Van den Hoven, Vermaas, & Van de Poel, 2015); (The Value-Sensitive Software Development Framework)

Design for values offers a perspective to create technology in line with the moral values of the users and society and it is more an over-coupling term of several approaches.. In this thesis is specifically looked into design for values in ICT.
The method is based on three main claims. Namely, values are embedded in technology, through the embedding of these values in technology values space action of users to be. As well that explicit thinking concerning values which are built into the system is morally significant. Lastly, that value consideration needs to be early in the process where it will have the biggest impact.
One of the first steps of this method is the translation 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) and the implementation level (system components as the basis for implementation) (Van den Hoven, Vermaas, & Van de Poel, 2015, p 838)
Design for values consists of three main activities namely: election of values, development between business and modeling views (domain specific) and the execution one which is the result of the modeling. Based on the book an abstract visualization is made of this method as well as a filled in example of how this method should be used in real life (Van den Hoven, Vermaas & Van de Poel, 2015).
This method shows us different abstraction levels of values to implementation, which in value alignment is experienced a severe challenge. This distinction of levels might help value alignment in AI development. The explicit use of values in software development has benefits for traceability of effects and allows for shorter development cycles (Van den Hoven, Vermaas & Van de Poel, 2015).
Additionally, making the different views explicit: of a value view, modeling view and business view can assist bringing the multidisciplinary of the AI field. Also, the different approaches toward design for values, for example more designer driven or user driven help to clarify the processes of AI development.
Nevertheless, in this method procedures for the design are lacking and even though vertical translation in the is deeply researched, the relations between the views is lacking.

Value sensitive design (Friedman et al., 2013).

Value sensitive design (VSD) is one of the most widely used described methods in this thesis. The basis of the approach is a tripartite methodology which combines conceptual, empirical investigations and technological ones (Friedman et al., 2002), and is a based approach to the design of technology that incorporates human values principled and comprehensibly during the design process (Friedman et al., 2013). It is fueled by the belief that product that humans engage with, influence the experiences as well as the ability to meet our aspirations.
Shorty the three parts of the VSD framework will be discussed. First, the conceptual investigation. In this phase, direct and indirect stakeholders are identified, as well as who's and which values are affected. Additionally, how value trade-offs should be addressed. For example (autonomy vs security). The meaning of specific values is researched in philosophical literature (for example the meaning of trust). Which later on will give a basis of comparison for the team. As conceptual investigation cannot go further there is a need for empirical investigation of the human context in which the technology will be used (Friedman et al., 2013). Almost all types of quantitative or qualitative research methods can be applied in this phase to gather insights. Example question given by Friedman et al is " How do stakeholders apprehend individual values in the interactive context? How do they prioritize competing values in design trade-offs?". Last the technical investigation comes which has two forms. One focuses on how existing properties of technology support or block human values. The other one, focuses involve proactive design of systems that were found in the conceptual phase. The distinction between the second and the third phase is the technical analysis really focuses on the technology whereas the empirical one focuses on the humans affected by technology.
Originally VSD has a list of "core" values with origin in moral philosophy (Friedman and Khan, 2003, p.1187). However, this got much critique as values play differently in diverse cultures and universality of values is extremely problematic (Borning & Muller, 2012). Additionally, is argued that researchers overclaim knowledge and authority in this method over the informants and a lack of attention is given into the subtle differences of designer's own values and the stakeholders ones (Borning & Muller, 2012). Lastly is argued that most VSD work is focusses on already built technologies and systems instead of building new systems (Flanagan et al. 2005). Nevertheless, concluding from this widely spread method it can be said that the three complementary viewpoints might be of value in value-alignment in AI. The view of direct and indirect stakeholders is relevant to prevent unwanted consequences and address desired values. Also, to research the deeper meaning of values in philosophy and that it is beneficial to research value trade-offs empirically can be extracted from this method. Additionally, the combination of a proactive stance (designing for values) as well as an interactional perspective (values in design and its co-constative quality) is an interesting perspective to take into account in this thesis (Shilton, 2018). Due the critiques and many research in the VSD field the decision is made to look at a combination of "core" values and situational ones.

Value dams and flows (Miller et al. 2007)

This tool can be used to understand stakeholders value tensions after the values already have been elicited. Basically, the method is based on three aspects. First, features that are experiences as problematic are avoided. Second, design in for desired stakeholder's values. Third, in a systematic manner address the value-oriented design tradeoffs.
Value dams are "technical features or organizational policies that are strongly opposed by even a small set of stakeholders" (Miller et al. 2007). This contains a strong ethical aspect, to recognize the desires and harms of the minority. Value flows are "technical features or organizational policies that, for value reasons, a large percentage of stakeholders would like to see included in the overall system, even if the features or policies are not absolutely necessary for successful appropriation " (Miller et al. 2007). This explicit use of value conflicts and desires results in solving conflicts earlier in the process. As well as due being aware and being explicit about values and the conflicts, during the case studies new crucially important values arose and were discussed. This tool shows us a way to translate values into functorialities as well make sure that conflicts are discussed. It makes stakeholder value conflicts explicit which is quite unique. However, it does not necessarily account for the designers' own values. Additionally, it is little tested, and the case study mentions a need for involvement of more indirect stakeholders as well.

Values at play (Flanagan et al. 2005)

This is a hybrid methodology that aims to discover rele-

vant values for a particular project and resolve the value-trade off, in this case specifically for game design. Due the explicit description of trade-offs this method will be shorty described. Flanagan et al. describe four stages of this process. Firstly, the values discovery in which relevant values from diverse sources are included. Namely: project goals and making hypotheses, earlier work, designer values, user values and other stakeholder values. It is remarkable from this method, that they explicitly use the designer's values, as often that is lacking in methods. Secondly, is identifying values-based conflicts and checking the functional components of it, in context of particular design choices. In this method, conflicts occur when not all specified values are implementable at the same time. Third is implementation and prototyping in which close attention is paid to the value conflict generated per functional component. This is an iterative process involving the earlier value sources, for ongoing feedback. Lastly, is values verification, in which with the initial list the values are compared with the result, desired values are embedded and undesired not. This process teaches us the different levels of values that can be integrated as well as the perspective of looking at value-based conflict in an iterative, functional and context specific perspective, making this trade off explicit. This appears to be a way to practically resolve value conflicts, however this method is only used in research context as far as found.

Envisioning cards (Friedman & Hendry, 2012)

Envisioning cards are a versatile toolkit that aims to discuss human values early in the design process as well as to put technological development in a wider socio-technical context and addressing it with a longer-term vision. It is based on the earlier discussed VSD. The cards have four so called "envisioning-criteria" namely stakeholders, time, values and pervasiveness, which are displayed on one side of the cards. With stakeholders is meant direct and indirect ones and consider implication for people one would not think of in the first place. Time is meant to stretch the timespan for which is looked at. Values is looked at the impact of technologies on human values. Pervasiveness looks at the new interactions that the rise of the new technology evokes. In this perspective is advised to look at for example geographic (google maps in urban areas), cultural (text messaging with blind people), demographic and many other factors. The other side of the card describes a focused design activity, with the words: think, identify, ask or sketch. 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). The tool is mostly used for educational purposes however is open to be used for inspiration, critique or heuristic evaluation and pointed out from the case studies that it catalyzes designers both humanistic as well as technical imaginations as well it is seen as a form of ethical reflection. The extracted insights form the toolkit are the 4 used "envisioning criteria" as well as the focused design activities used to make it easier to communicate abstract thoughts/opinions and make humans aware of the effects that their technology might have. To switch perspective and think about other opportunities it is handy. A point of critique from my personal perspective is the lack of the explicit making of one's own values, then still unconsciously unwanted values might be implemented in systems. Additionally, it does not give more concrete handles how to deal with discussion or the value-conflicts.

Insights methods and tools

From the analyzed methods and tools several insights can be drawn. Firstly, most of the methods focus at the beginning of the process, as they argued it will have a bigger effect on the outcome and process. Secondly, almost all methods account for both indirect and direct stakeholder values and consequences. Next to this it seems also a critique in many methods when it does not account for the designer's own values, as this can lead to unconscious value implementation or undesired value-tension later in the process. Thus, in this thesis it is important to take it into account for a more value aligned AI development.

Thirdly, as values can be rather abstract, most methods aim to translate these values or communicate these values in more concrete manners, however the manners are different. They differ, such as discussion, sketching, structured decomposition on diverse abstraction levels or conceptual and empirical research for deep understanding of the values both conceptual and in context. Therefore, it seems save to say the decomposition of values, doing research about them in context as well as the communication of them is crucial for a desired value-alignment, however the manner to do this can be one fitting to the IBM teams.

Fourthly, value-conflicts/tensions, seem to be addressed better when discussed explicitly and addressed explicitly as well. The value dams and flows method, sees every value conflict a functional constraint/opportunity, which is an interesting perspective on the problem. Also, the explicit empirical research about value-hierarchies is an interesting way to address value conflicts and might be an interesting way to proceed with value-alignment in AI.

Nevertheless, practically all methods have been tested limitedly in practice or in specific industries. Therefore, it might be extra interesting to look at development of support for companies together with the company, in this case IBM. Additionally, no method has been found that focuses on both value election of both designers and all other stakeholders, organization as well as resolving conflict between these in diverse industries.

Concluding, both core values of IBM and contextual values differing (per industry, client, team, individual) will be taken into account for design for a more value aligned AI in this thesis. The organizational perspective will be taken into account as this might be lacking in current tools and methods. Both for direct and indirect stakeholders' values will needed to be taken into account and explicit use of the values, their communication and decomposition to support the AI development team to design more value aligned AI applications. As well this thesis aims to add to the research field of value alignment in engineering, with the development of support with the (strategic) organizational perspective. Also, it aims to look at the value tension/conflicts from the different identified value levels and sources with a design perspective and the conflicts that can arise on the diverse levels as currently there seems to be a research area untouched upon.

Appendix K I Analyses interview and tool

Main findings interview with tool

Lack of technological knowledge of the team and the ethical implications it brings with it

"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)

→ **A need for team alignment on technical capabilities & ethical pitfalls**

Lack of moral motivation
" 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)
" if we are allowed to use it we should use it, it is part of the game haha.. " - interviewee about personal data (Manager)
→ **Increase intrinsic motivation for ethical decisions in projects of the team and especially the DS**

Lack incorporation of stakeholders and actual (societal) consequences
"For engineers, we can really easily be absorbed by technical challenges and forget about everything else, that is why engineers contributed to much terrible stuff, as the example of Volkswagen Emission Scandal. From engineering point of view, beautiful but actually a disaster. Engineers might need some check points in the mean time to make sure we are not too, ambitious to solve the technical issue."- Interviewee (data scientist)
→ **Integration of the consequences of the models predictions have in the ideation**

A lot of, roles and tasks are the responsibility of the data scientist and therefore a lot of pressure. Also the feature engineering and modeling decisions are made by the data scientist.
" Most of the decisions in the data preparation and modeling phase I make myself. I look at what works or does not work to improve the accuracy of the model" - Interviewee (data scientist)
"In the beginning I am often more asking and listening, trying to understand the problem, than I am the one in the data readiness assessment I am the one who asks if the data is ready, than in the feature engineering I am the main person modeling it the data enrichment and deployment and the last two I am supporting IT but they are making the call" - Interviewee (data scientist)
" my role? ahhaha it would be everything" - interviewee (Data scientist)
→ **Remove pressure from the data scientist & Highlight the importance of decisions in the modeling and feature engineering phase**

Value tensions or values are not consciously addressed
" ...and also for us as technical guys to be aware of values, and the higher impact, I don't think we ever thinking in this way" - Interviewee (Data scientist)
→ **Explicitly discuss & solve value tensions for desired outputs**

Much unexpected challenges occur during the process
" We did the rework and delivered it, but then they told that there are no models in place, so no increase, that is again a surprise, we kind of have to do everything from scratch" - Interviewee (Data Scientist)
" A big surprise was that the data baes was empty" - interviewee (Data Scientist)
→ **Support for dealing with (ethical) surprises**

Miscommunication between the different disciplines
" You create layers and layers of complexity, when one asks a question on a high level it is difficult to explain it without the complexity" - interviewee (Data Scientist)
→ **Support for communication**

Stakeholders
Most of the interviewees only took into account the core team and some of the internal people in the company that would use the system in the end. Just two people put the end customer as an indirect stakeholder. As appeared from the ethics literature, for more ethical outcomes it is important to integrate the stakeholders opinions both direct and indirect into account.

Values
Most data scientists choose mostly technical values for themselves and the model (such as: robustness). The data science consultants from IBM also choose the IBM values (such as serving client). The business owner chose much more business related values and wrote down new ones (such as entrepreneurship).

Overall it appeared that the data scientists really were appreciating their freedom in their work and did not want much control. Thus, a new value tension was discovered:
Responsibility/accountability vs autonomy freedom. On one hand most, data scientists and the manager did not take the responsibility for the ethical implication but on the other hand did not want to be controlled, empathizing with their freedom

Simplification vs Uniqueness/Veracity is also a value tension that appeared during the interview. This one is similar to the bias variance trade-off one found in literature. On one hand one does not want to oversimplify the world too much with the model. While on the other hand using it is also not good to

Probity (fairness) and accuracy; this one was not taken into account at all. The KPI's were all related towards accuracy so that is also the metrics it was tested for.
" Freedom from bias is a big thing ahah I am not sure they do it…" - Interviewee (Data scientist)

Socially desired vs historical data was also no attention paid to.
Explainability vs performance, appeared from the interviews and in some use cases it had a higher priority then in others. In the automation case, an illustrative quote is described:
" Explainability is nice but not the highest priority, but I don't have enough time to spend" - interviewee (Data Scientist)

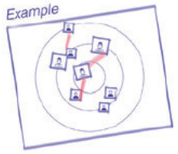
Input for provotypes
The following value tensions are tested with the provotypes, chosen with the lens of the research in mind (design for fairness):

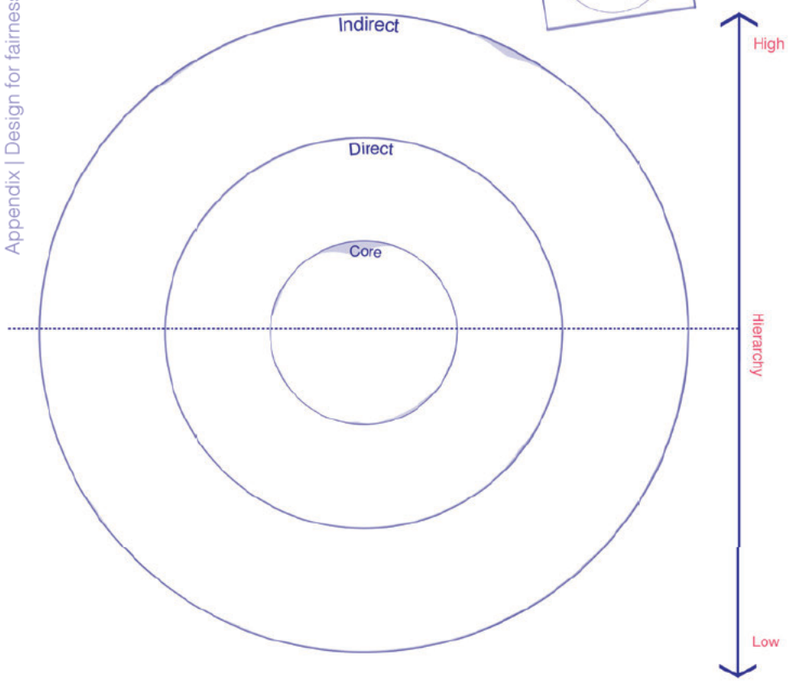
- **Socially desired value vs historical data**
- **Simplification vs Uniqueness Veractiy**
- **Responsibility/accountability vs autonomy freedom**
- **Probity vs accuracy**
- **Expainability vs performance**

1 Stakeholders

Could you map the stakeholders involved in the project (direct & indirect)

Example

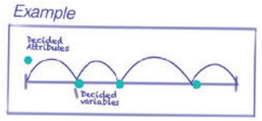




2 Process

Could you map the projects process entirely

Example



PROJECT IMPORTANT DECISION MOMENTS

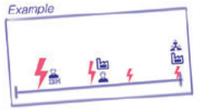
PERSONAL IMPORTANT MOMENTS

CHALLENGING MOMENTS

Tension

Could you map the tensions during the project

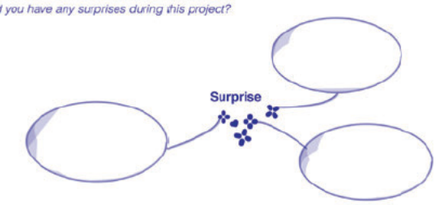
Example



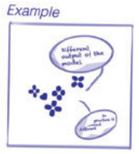
TENSION MOMENTS

3 Surprise

Did you have any surprises during this project?



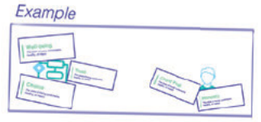
Example

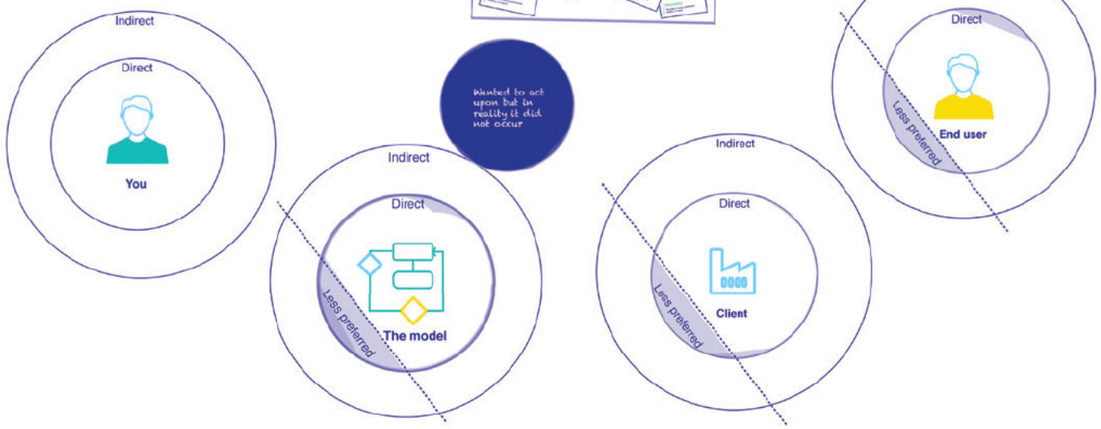


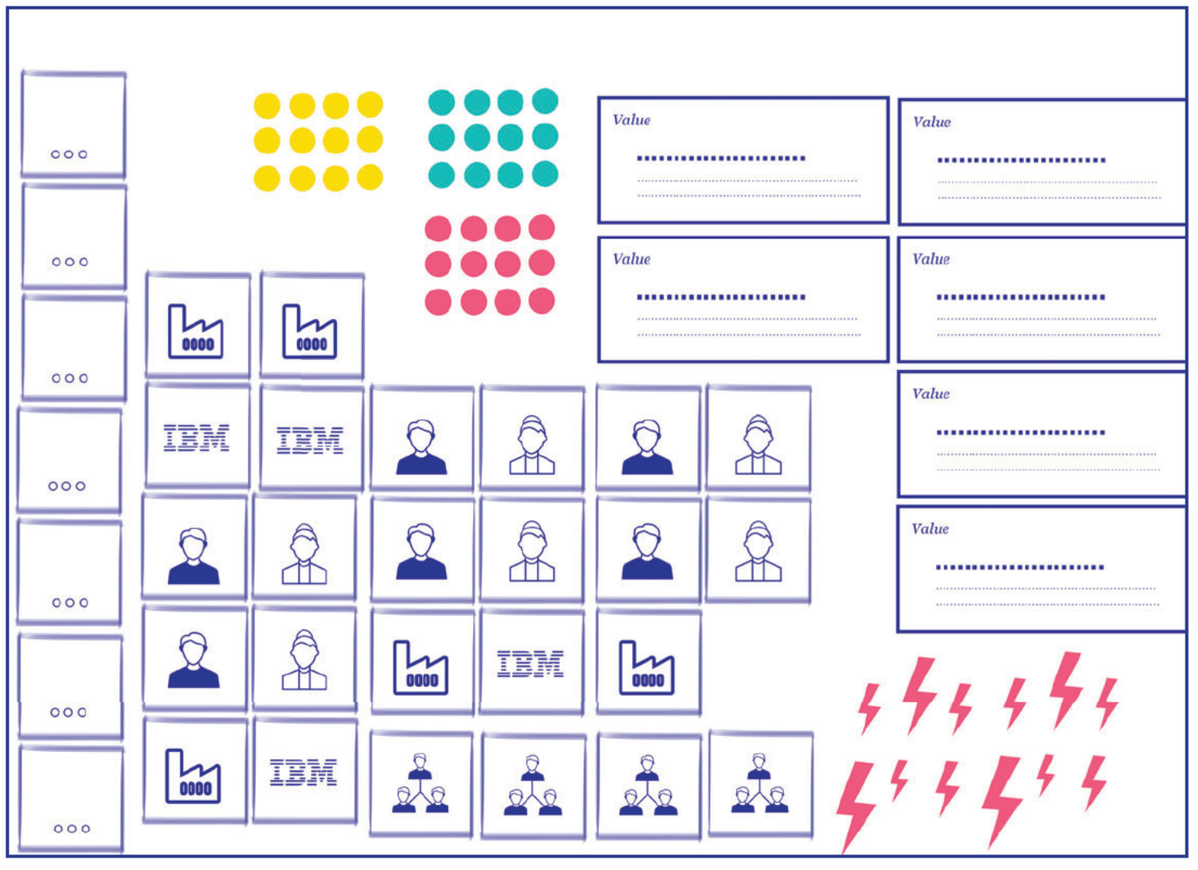
Value

Could you map the values that are consiously represented in the model? And the ones that are important to you?

Example







Interview guide

Values

Moral Value
Respect
including notions of respect for human rights

Moral Value
Trustworthiness
including notions of honesty, integrity, transparency, reliability, and loyalty

Moral Value
Responsibility
including notions of accountability, excellence, and self-restraint

Moral Value
Caring
including the notion of avoiding unnecessary harm

Moral Value
Citizenship
including notions of obeying laws and protecting the environment

Moral Value
Fairness
including notions of process, impartiality, and equity

Human Value
Well-being
the state of being comfortable, healthy, or happy

Human Value
Connection
the state of being related to someone or something else

Human Value
Self-expression
the idea of using your own thoughts and what is appealing to you, in order to express how you feel, showing difference

Value
Augmenting humans
the strength and believe of human-machine collaboration

Value
Trust
to believe that someone/something is good and honest and will not harm you, or that something is safe and reliable:

Value
Accountability
the obligation of an indiv. or organization to account for its activities, take responsibility, and to disclose the results in a transparent manner.

Value
Safety/Security
a state in which or a place where you are safe and not in danger or at risk

Value
Socially desired
for the social good with benefit for humarity

Value
Freedom from bias
free from undesired bias

Value
Awareness
knowledge and understanding of a particular activity, subject, etc.

Value
Community/ Belonging
acceptance as a member or part

Value
Autonomy/Freedom
the ability to perform tasks in complex environments without constant guidance by a user.

Value
Transparency
the quality of being done in an open way without secrets

Value
Explainability
transparent & explainable AI

Value
Informed consent
the process by which a person learns about and understands the purpose, benefits, and potential risks of a system

Value
Support
to agree with and give or recieve encouragement to someone or something because you want him, her, or it to succeed

Value
Environmental sustainability
the maintenance of the factors and practices that contribute to the quality of environment on a long-term basis

Value
Business sustainability
sustainable business success in terms of cash fow and profitability

Value
Serving client

Value
Robustness
strong and healthy; hardy; vigorous:

Value
Universal usability
something is universally applicable

Value
Privacy
the state or condition of being free from being observed or distubed by other people

Value
Calmness
the state or quality of being free from agitation or strong emotion

Value
Identity
the fact of being who or what a person or thing is and respecting that

IBM
Literature ICT
AI principles synthesis
Human
Moral

MEET THE AI TEAM

Personas

Tanja

Data scientist *(Consultant)*

Responsibility, trustworthiness, autonomy/freedom

Demographic
Single
Amsterdam
3 years of work experience
Data Science at VU

Behaviors & characteristics
Curious, likes to explore the new and is optimistic.
Strong technical perspective
Stubborn & Clever
Likes to work on specific tasks without being disturbed
Prefers numbers

Needs & Goals
Aims to lead the ideas to actual industrialization.
Seeks recognition in a rather new field
Designs the features as well as the modeling. In the entire process aims to bridge the business, clients, IT perspectives

Drawbacks
Often a bit naive, less thinking about the concequences and the time it takes in practice.
Not (much) trained in bridging perspectives but it is part of the job

"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"

"Most of the decisions in the data preparation and modeling phase I make myself. I look at what works or does not work to improve the accuracy of the model"

MEET THE AI TEAM

Personas

Niko

Business/internal client

Entrepreneurship, continous improvement

Demographic
Married and 2 children
Berkel en Roderijs
19 years of working experience
Marketing background

Behaviors & characteristics
Social, fast, functional wants to see immidiate benefits
Wants so see continous improvement
Gives direction where to go

Needs & Goals
Clear communication, likes social contact and meetings
Aims for fast results and performance

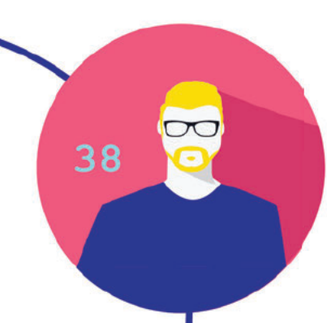
Drawbacks
Can be focussed on performance a lot

"... then the data scientist, with all respect " the nerd" just tells the possibilities, and we of course do not understand anything, from our side we want to know what it means for us and what kind of impact will it have on the different departments"

Bas

Manager DS team

Responsibility, trustworthiness



Demographic
Married & 1 kid
Haarlem
15 years of work experience
Business administration

Behaviors & characteristics
Does not like change too much
Prefers order of uncertainty
An agreement is an agreement
Strong technical perspective

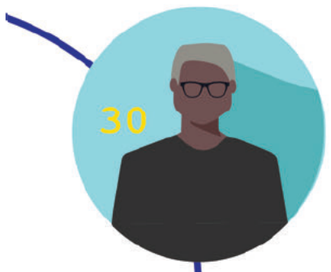
Needs & Goals
Convince the business departments
Seeks for commitment
Aims to go for industrialization

Drawbacks
Technological perspective might override other values

"If we are allowed to use it we should use it. Then it is up to the regulator to say you should not use the type of features, it is part of a game haha."

Stef

IT programmer



Demographic
Single
Amstelveen
6 years of experience
IT

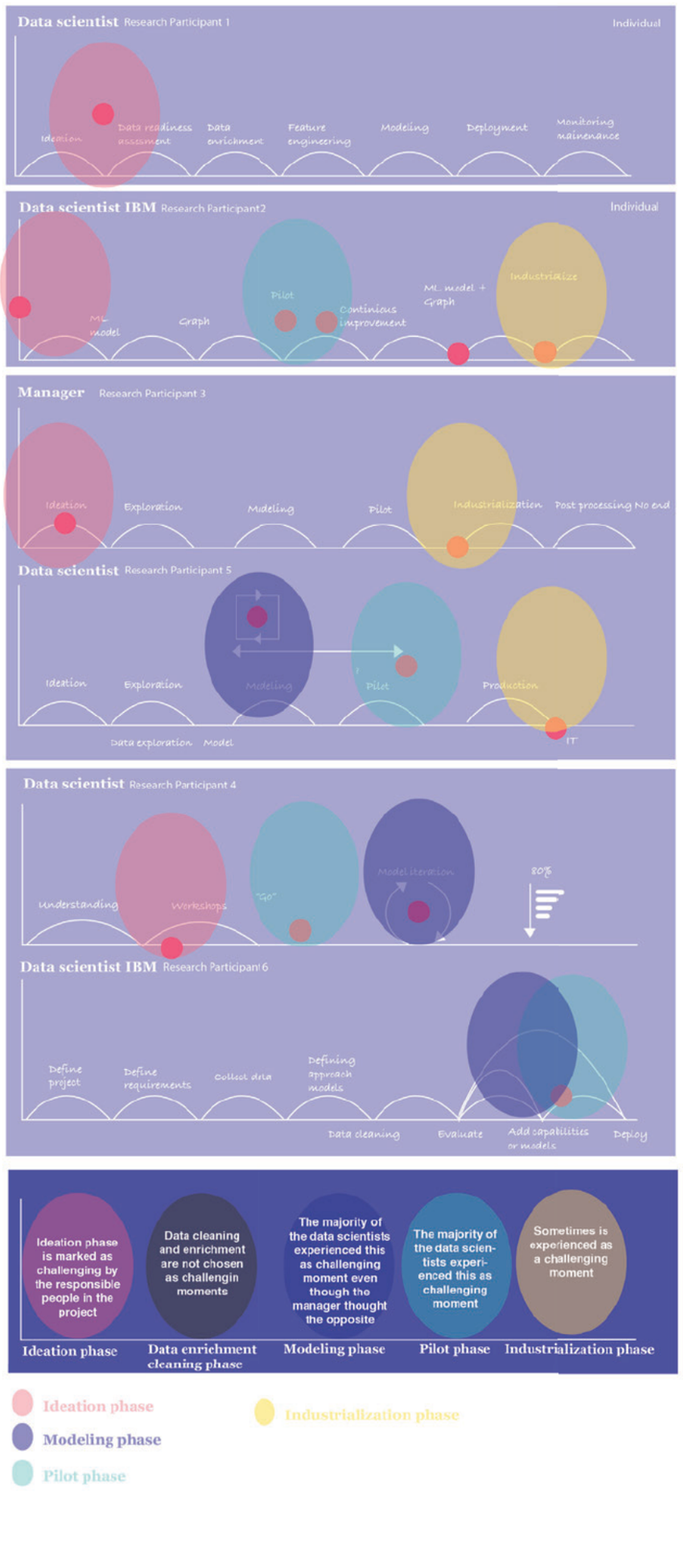
Behaviours & charactaristics
Wants clear communication and agreements
Wants the model to be robust and structured
Loves coding

Needs & Goals
Aims to make the models scalable and implemented in the current systems
Wants models in a manner that is implementable

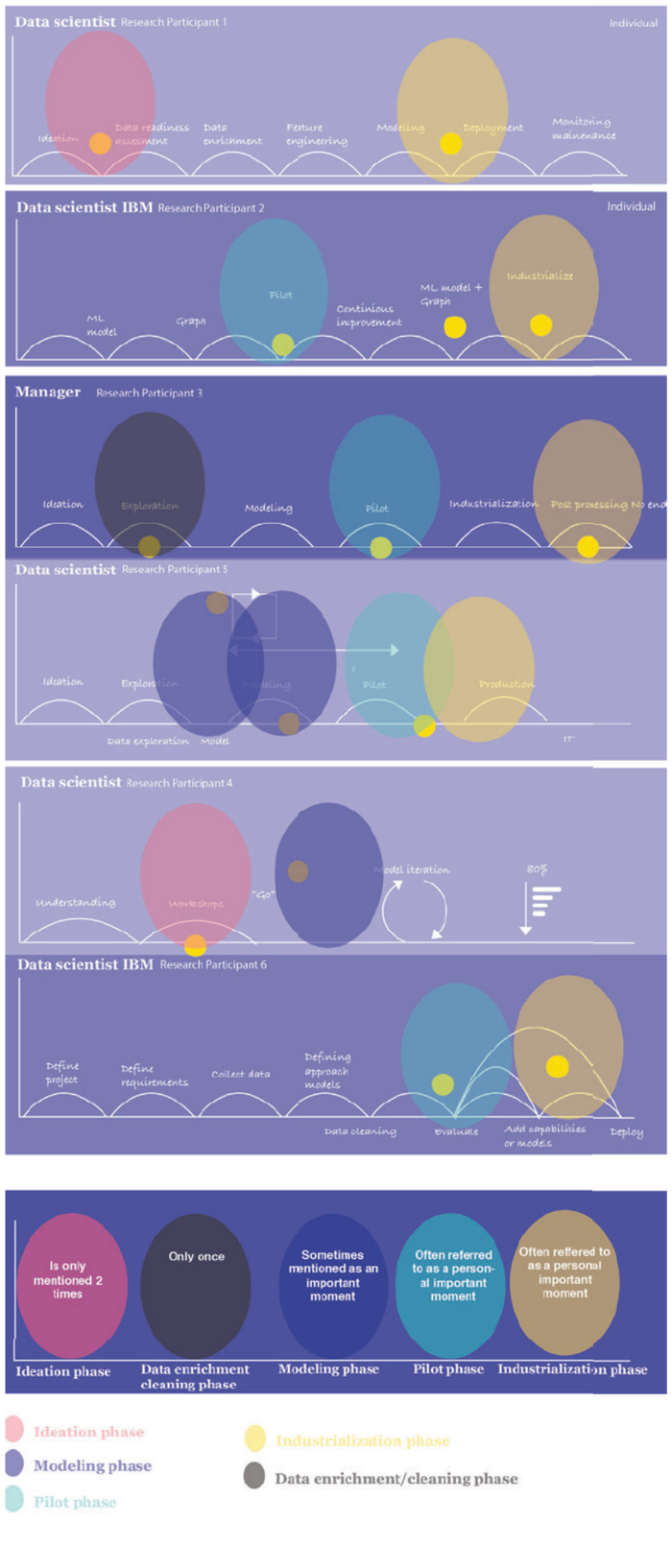
Drawbacks
Less open for new ideas due the drawbacks and difficulties in the implementation in current systems

Appendix J I Persona's based on interviews

Challenging moments

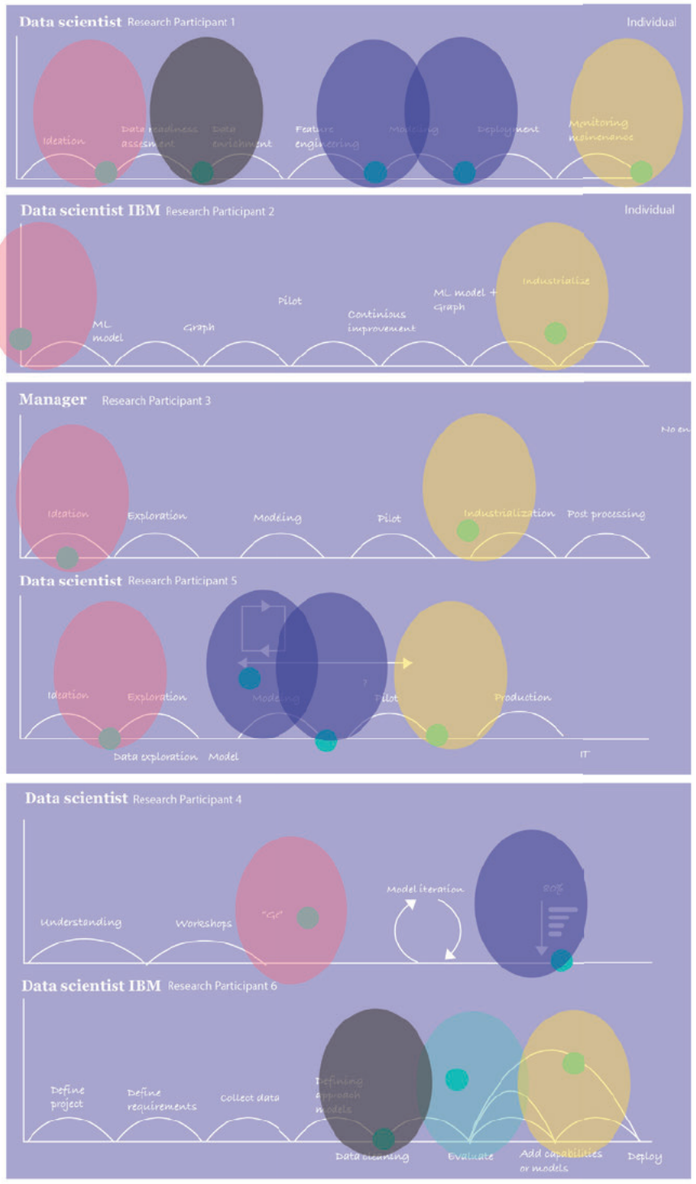


PERSONAL IMPORTANT MOMENTS



Moments of tension

PROJECT IMPORTANT DECISION MOMENTS



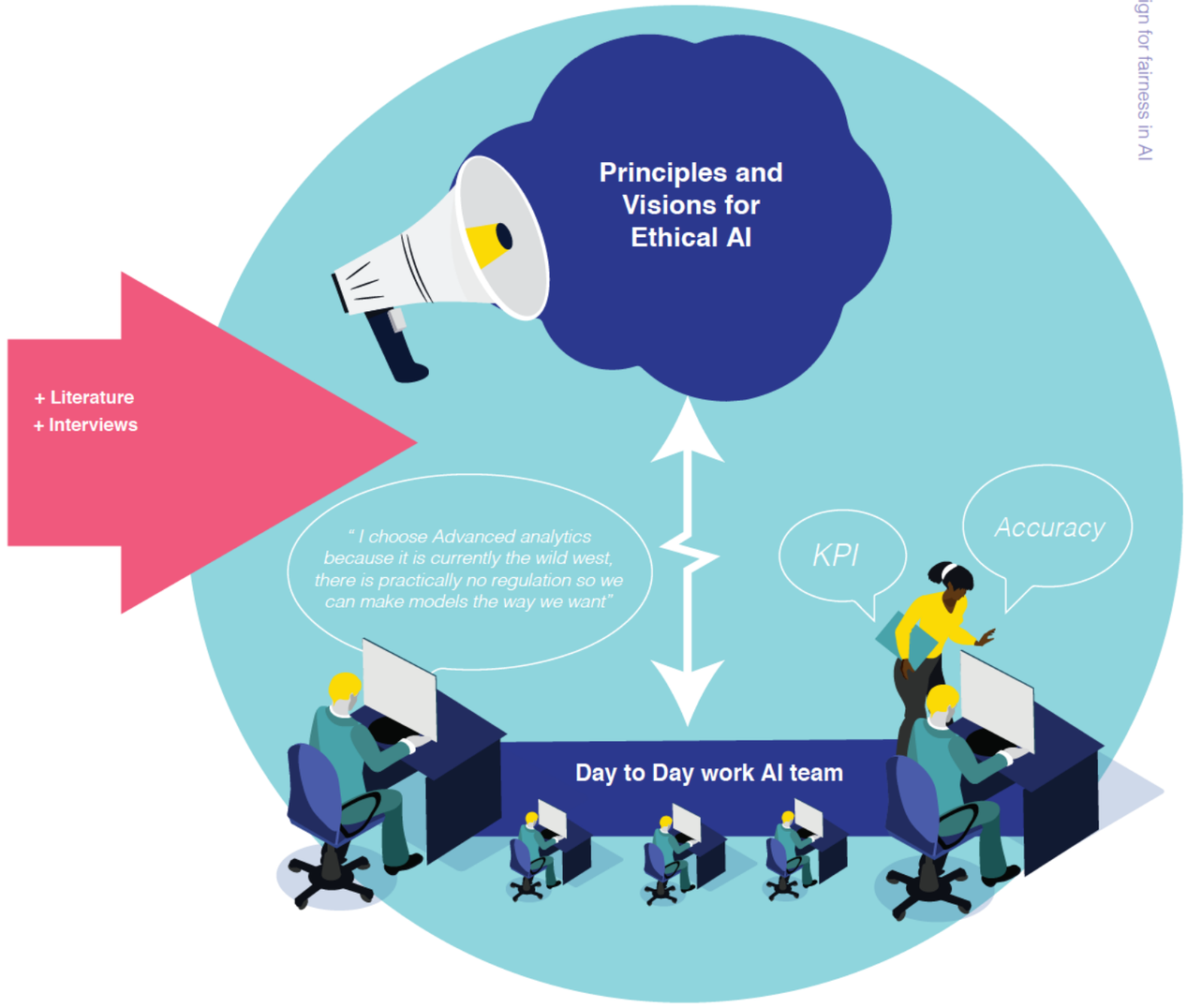


FIgure 6.12 | Visualizaiton of the challenge of ethical stategy and princples uptake in AI

FAIR PRICES.

Coffee place 2025

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

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

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

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

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

D.P.Simons 2019

FAIR FACTS.

Airport control 2030

1I What is your first reaction?

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

3I Do you consider this as a fair use of the statistical fact? Why yes/not?

4I What values important to you might be affected? Try to think of at least 2.

5I If the values are different for the different stakeholders, how should you resolve those values contrasts in this scene?

D.P.Simons 2019

RESPONSIBLE SPENDING.

Paying in 2027

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

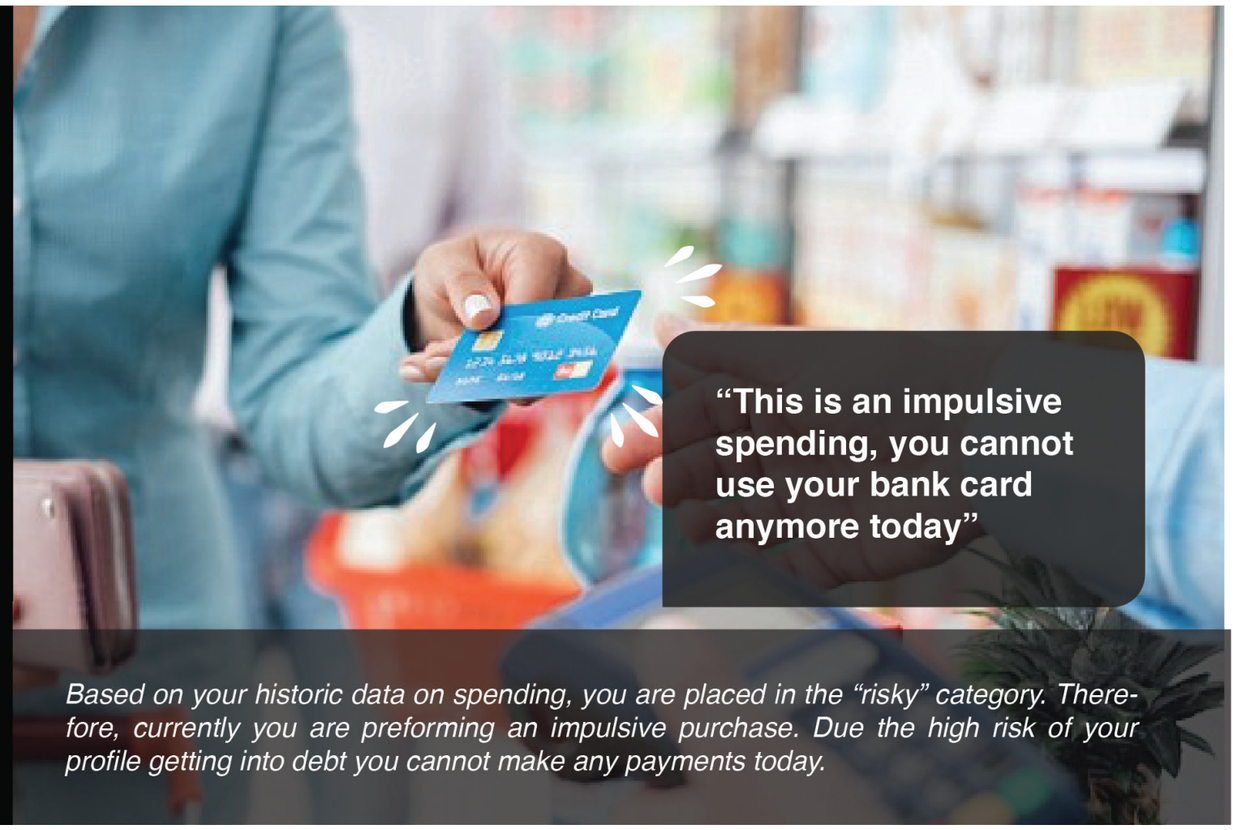
2| Would you like to work on making a system like this? Why yes/not?

3| Do you believe the bank should take this responsibility? Why yes/not?

4| What values important to you might be affected? Try to think of at least 2.

5| If the values are different for the different stakeholders, how should you resolve those values contrasts?

D.P.Simons 2019



OPTIMIZED PERFORMANCE.

Your Office by 2030

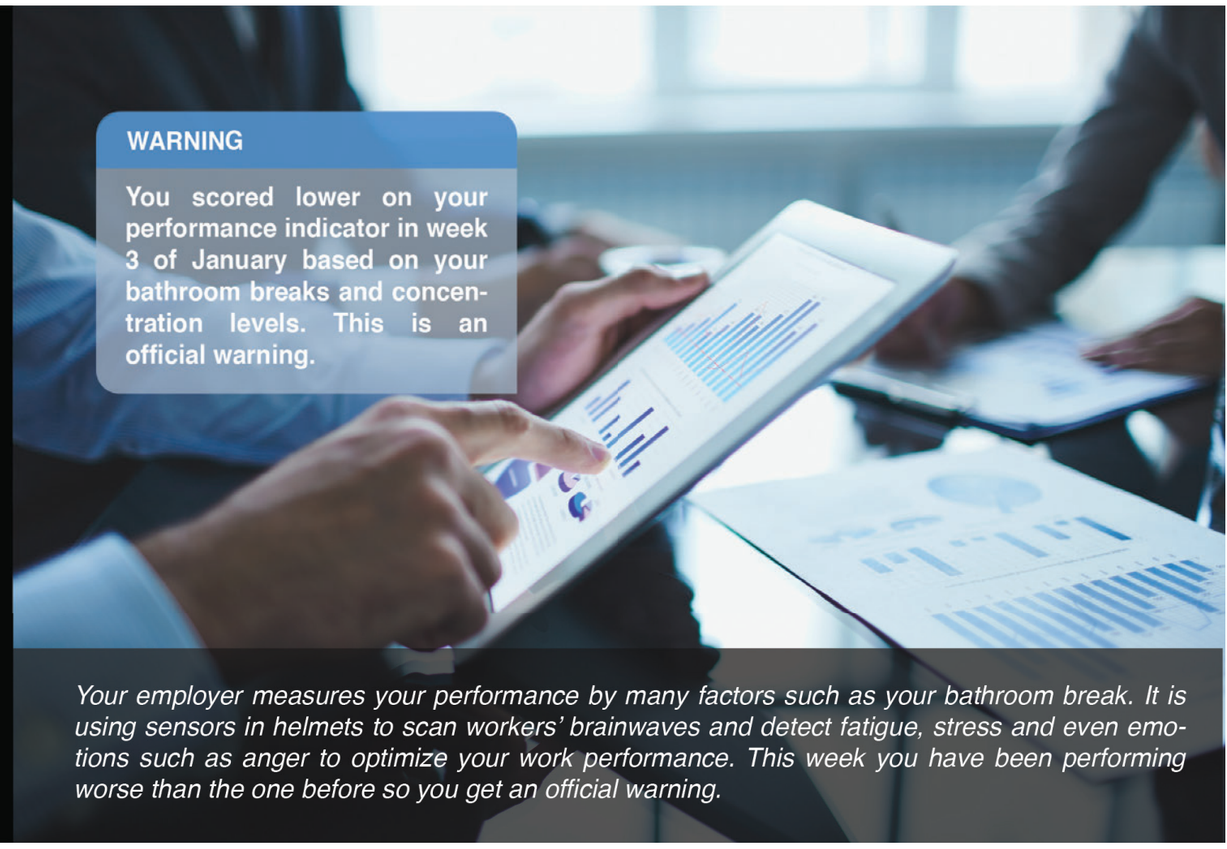
1| What is your first reaction?

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

3| What values important to you might be affected? Try to think of at least 2.

4| If the values are different for the different stakeholders, how should you resolve those value contrasts?

D.P.Simons 2019



RESPONSIBLE FREEDOM.

Skiing holiday 2030

1| What is your first reaction? How would you continue?

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

3| Do you consider going on this slope your personal responsibility? Why yes/not?

4| What values important to you might be affected with this system? Try to think of at least 2.

5| If the values are different for the different stakeholders, how should you resolve those values contrasts?

D.P.Simons 2019

Appendix M I Analysis of the provotypes

5.3.2 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 (un)conscious values

Thus, for the first aspect I proposed to ask a question concerning a scenario, such as to finish a scenario in a manner they would like? Or/and what value would be most important for them in a such a scenario?

Secondly, the way the **value tensions are resolved by the current AI team** is unknown for this project and in research in general. To gain a richer understanding concerning these value trade-offs and how these are currently resolved in AI development, is targeted by the provotypes. In this manner also, distinction between values trade-offs important for the whole team or just for one certain role.

For the second aspect I propose the type of scenario's concerning their reaction and how they would resolve the situation in order to discover the way they resolve the val\ue tensions. Also, I asked them to make a hierarchy of scenarios they prefer the most till the least.

Thirdly, I will test the **extracted process** from the interviews. For the third aspect I will send the extracted process and ask for remarks and feedback, testing my analyses of the tension points and decisions moments

The provotypes

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 were shown to al participants a switch was made between the responsible spending and responsible freedom per participant. The provotypes were personalized in name usage and small details to increase the empathy with the scenarios.

- Socially desired value vs historical data
- Simplification vs Uniqueness Veractiy
- Responsibility/accountability vs autonomy freedom
- Probity(fairness) vs accuracy
- Expainability vs performance

Results and findings

The answers of the provotypes are all read, analyzed, summarized and compared to the answers between the different participants. I performed this analysis with the lens of the research questions and goals in mind.

Goal 1 Values

For the first goal of the provotypes, extracting/confirming values from the interviews and new ones, the results were surprising. Some participants stayed within the scope of technical values or service related ones (such as customer loyalty), where other participants stretched their imagination and named more human and personal values (such as adventurous).

Overall, the data scientists, as expected, had a more technical perspective on the diverse scenarios. Here is an illustrative answer of one of them as a response towards the airport woman fast lane example:

" 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

"The problem here is that the outcome is 0 (not criminal) or 1 (criminal). Thus, it is not a question to be stricter or not but it is a question of 'Who will be checked'. At a more general level, I would be okay to work in a world where the first filter is provided by statistics indeed as long as it can be explained to users. The fact that it discriminates Man/Women is not really a problem (as long as it would also discriminate other features whether these features were significant)." - Manager

These quotes show a strong technical manner of thinking and less the ethical lens.

It seems to be education concerning ethical features and the social impact resulting from these systems needs to be integrated and made clear in their processes. Nevertheless, participants also showed also the trade-off between technically very interesting and desired systems (optimized performance):

"No. even if "technically speaking", the project is re-

ally 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)

This shows a reflection of the implications of certain systems may have is made by some of the participants. The scenarios make the context and the consequences much more relatable, which is seen in the provotypes results compared to the interviews.
Scenarios seem to be a good manner to relate to the actual end user as in this case it is the participant who is the end user.
Additionally, I noticed that some participants dislike the situations in which they are part of the group which is treated less preferable. An illustrative example:

" No, because I think I am being screwed again and I do not want to develop something I have no faith in." DS

From this example is extracted that when a scenario is more personal and relatable, the participants seem to dislike the situation/value it more. **Therefore, making the final design more personal, might stimulate ethical reflection.** However, in the answers of the skiing example provotype, most participants agreed with the idea although they really like skiing of the slopes. Still they agreed upon taking their own responsibility as a skier. **This shows that they make the trade-of between personal benefit as well as societal benefit, choosing in this case societies benefit. However, it was clearly stated that in this case that the transparency and the awareness are highly important in this case to not limit one's freedom and make it clear that the skiing person takes his/her own risk, the right of the awareness of own responsibility.**

" 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

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. Concluding, privacy is experienced as fair and the violation of privacy as unfair.
Managerial/business positions answered more in an organizational fashion and less in an individual one. Also, the responsibility was transferred to other parties in society.

" The market context is equal. This would support fraud. I think the current system of the taxes salary depended is more effective." - Business owner

Goal 2 Resolving value tension

Relating to the second research goal of the provotypes, I extracted new or more nuanced value tensions. This was done by comparing the different values they mentioned in the scenarios, extracting values from the sentences as well as clustering of the similarities and differences between them. From the analyses of answers, the following value tensions are derived, diverse from the ones identified before: **freedom/privacy & safety/control, simplification/ optimization & authenticity, individual good & collective good.**

Also, different aspects of fairness appeared from the provotypes, equality is a recurring word used by the participants. This is one of the perspectives to consider fairness. In the provotypes most of the participants did not believe this perspective was fair, as they are affected negatively, their group or because it takes extreme forms closer to communism. Also, its mentioned in for example economic equality we already have tax differences so the rest should not account for it (putting the responsibility for economic equality towards the government). **The "deserved" perspective on fairness reoccurs to be the more preferred one** (described in the fairness chapter).

" Wealth. In my opinion if we would start living our lives this way wealth would be no existent as you cannot become wealthier than the next person. Equality. This would bring everyone on the same level of wealth thus creating equality but probably at cost of productivity" - Data Scientist

The question how value tensions were resolved was experienced as difficult by the participants. Some did not answer this question. **Some changed the data that was used for the model to make it acceptable, others put the responsibility on more political levels or put certain restrictions on a system for them to work with it.**

" I would not mind working on such a system as long as the boundaries of coverage are transparent and clearly communicated to everyone." - Data Scientist (Skiing example)

One participant changed the autonomy of the system to make it more acceptable changing from autonomous to supportive decision making. In other **examples the explainability, transparency or awareness of the responsibility are added to make the system fairer**. **These examples of changing features of the AI system to make it more acceptable can serve as input for the solutions space of resolving value tensions.**
Remarkable is that from the previous identified value tensions, the fairness vs accuracy did not reoccur, also historical data value vs socially desired did not come back in the answers. The performance of the models was not really mentioned in general as one of the tensions that might occur. I believe the reason for this is that they were mostly answering the provotypes being the end user and less as the maker of the model. From the interviews and literature, the tensions are highlighted as very relevant relating to the models fairness. Therefore, I take them into account and perhaps more explanation and education is needed to with the AI teams to relate to these value tensions.
3 Relating to the third research goal, testing the extracted process, decision moments and tensions no feedback was given. Therefore, I see this process as validated and as a good representation of th

teams.

Case

Global Insurance company - Dutch establishment

Team: 1 data scientist, 1 IT person, 1 business owner, 1 Machine learning team manager, Head of team and data analyst.
Client: the business department
People using the system: Pricing team (Internal department insurance company)
End-customer: who pays for the policy
Duration project : 6 months
Project goal: Improve pricing for the premium customers for car insurance and increase prof t
Firstly, the deciding on the premium getting from the customers involves building mathematical models and there could be other machine learning models

End deliverable: for all the policies that are given we give the price form the model and, sometimes just the probabil ty score that they will laps etc. this kind of deliverables

Data accessible: all data an insurance company has about a person buying a car and the state of the car when bought.

Current state:

" If we are allowed to use the data we should use it it is part of the game"

AI Dish
Ethical and technological alignment & understanding

Ingredients

Data from humans
Protected features
Non-protected features
Non-Human data
Where from?
Link
Sensitive ingredients?
If used checks in the process are needed
By law

Who is in the kitchen?
Role?
Data scientist
IT developer
Manager
Business owner

Fire
Computing power

Appliances
What algorithms are used?
Why this choice is made?
Positive aspects
for society, individuals organization the model
Negative aspects
for society, individuals organization the model

Recipe
Supervised learning | **Unsupervised learning** | **Reinforcement learning**
Type of model
Why?
Positive aspects
Negative aspects

For who?
Who will be using the model
Who will reading the results of the model
Who will be updating the model
Who will be impacted by the models results

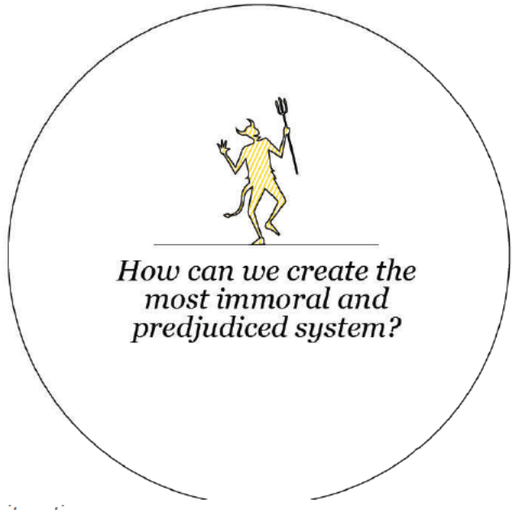
Dish
Goal
Implications

1st iteration AI Dish

I Extremely **injust & immoral & predjudiced & unfair**

1 | *Write down the project goal concisely*
2 | *Come up with really evil unfair approaches and ideas for the model - Release your worst side & the evil in you*
3 | *Describe why these ideas are unfair.*
4 | *Extract the attributes for the unfair reasons*

Project goal: .....



For who?

Who is most affected by these ideas? Which parts of society? Or specific individuals? Organizations?

Attributes:

Why is it so immoral, predjudiced and unfair?

Because of the ingredients?

Or the recipe?

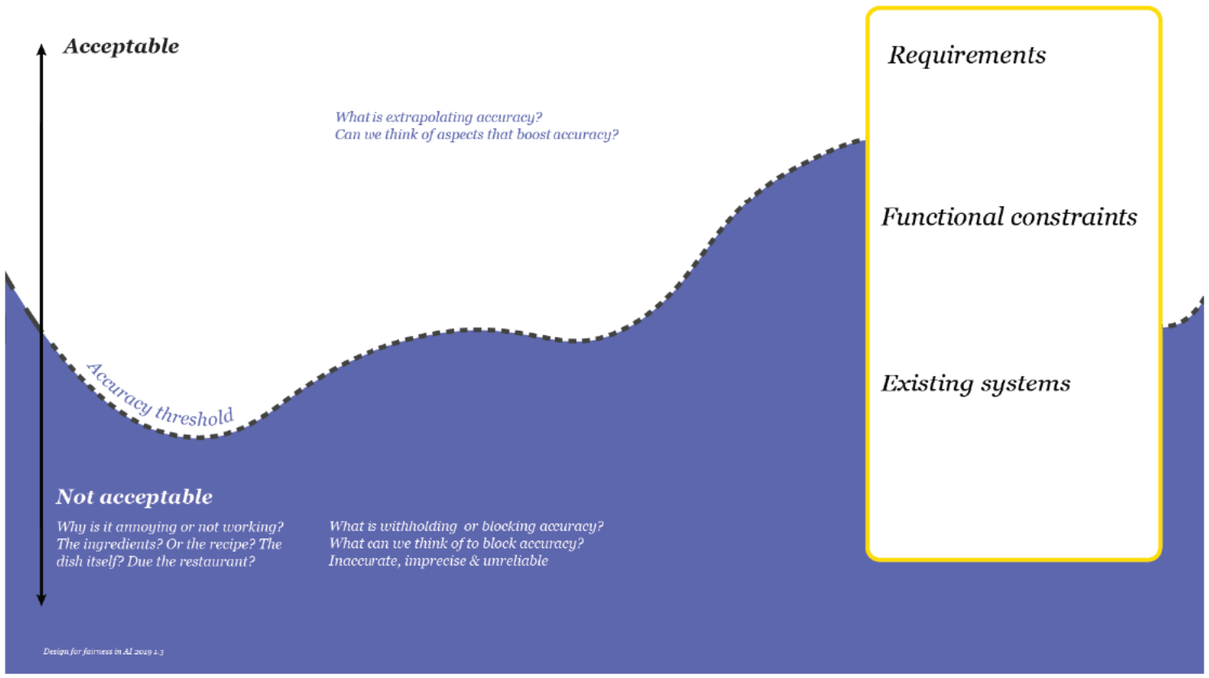
The dish itself?

Due the restaurant?

Evil AI canvas for probity 1st iteration

I Threshold

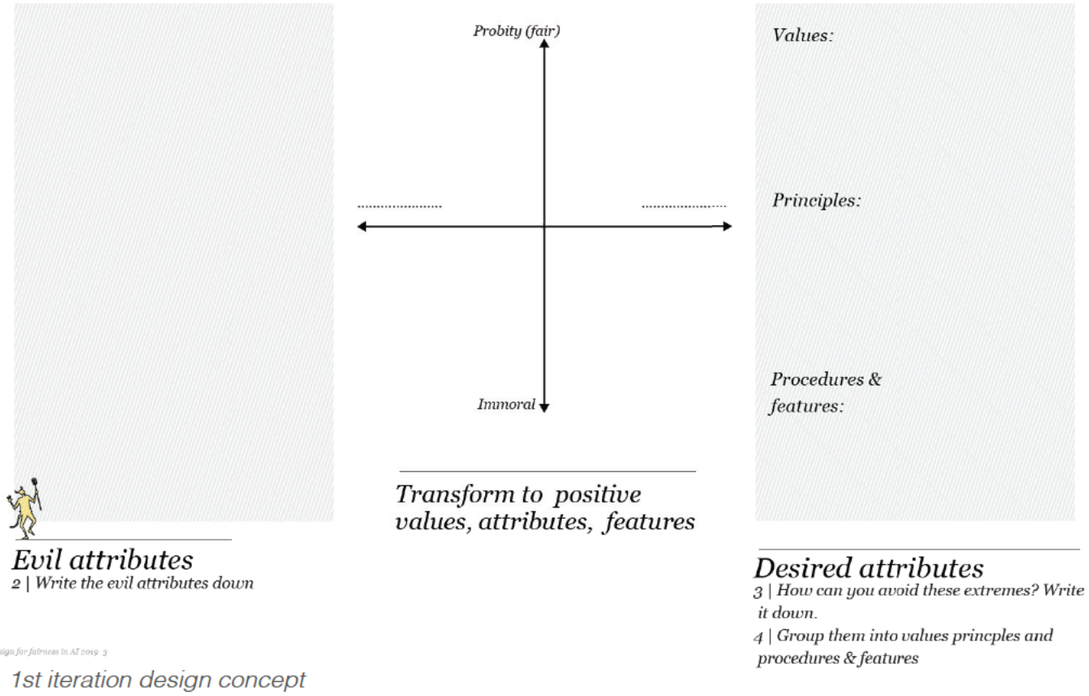
1 | *Which are really unacceptable? And why?*
2 | *Are there ideas that would be still acceptable? and why?*
3 | *If not try to ideate around ideas that contain that aspect.*



Threshold canvas for accuracy 1st iteration

II Angels Advocate

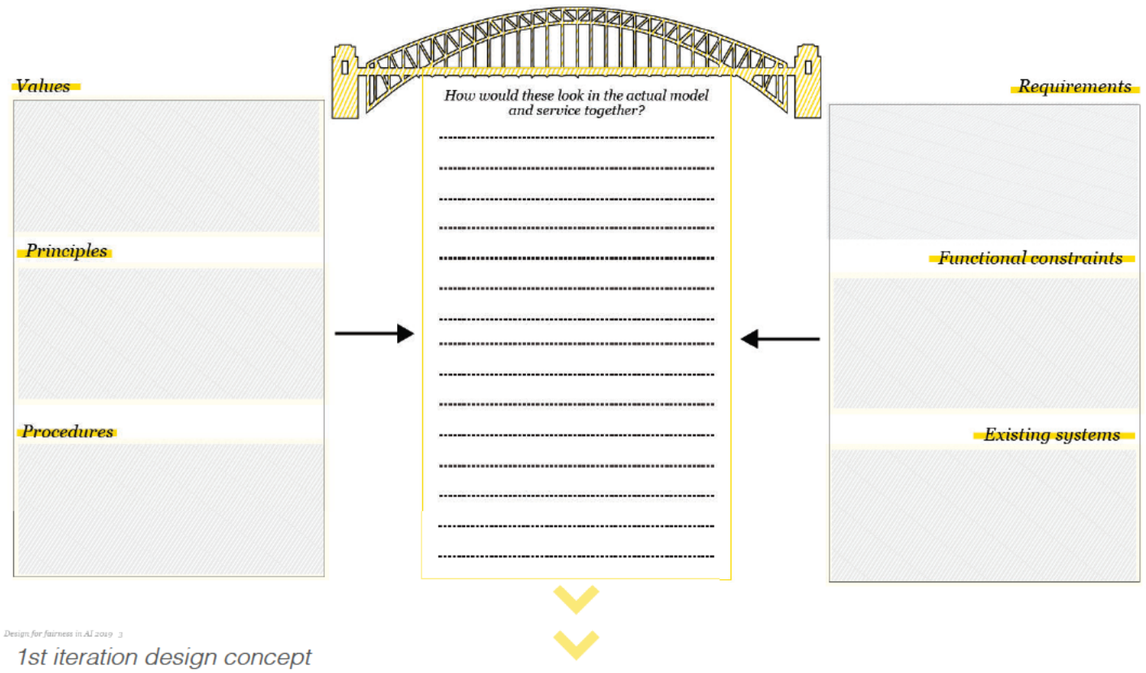
1 | *How to avoid these unfair ideas/attributes? Try to think of principles, features, attributes you want the model to have to avoid these.*



1st iteration design concept

III Overpass

1 | Pick the ideas which spark your imagination of both scenarios and put them on the sheet together with the attributes
2 | How can you avoid these extremes? Write it down in the middle
3 | Are there conflicting ideas? Try to discuss these what is in this projects context more important?



1st iteration design concept

AI Dish

Testing 1 with designers

The metaphor and explanation was experienced as very helpful to relate to the topic. The advice is given to continue further with the metaphor further in the workshop. For example, in the clustering for attributes in the evil exercise this is advised. When answering the question, why is this unfair? It would be easier to answer with, it is unfair because of the ingredients or unfair because of the recipe.

" The AI dish, works really well and really appeals, it works immediately"
- Participant

Testing 2 with computer scientists

The AI dish worked well for the aimed goals. Even though it was a short workshop during the workshop also reflection moments relating to the AI dish happed. After the session informal interviews with the participants and the designer assisting the facilitation were held. Some insights are represented by quotes from these interviews and translated into call for actions.

Reflection on data:
M: "It is in the data, but I do not use the sensitive ingredients.....but wait maybe the algorithm still can pick it up"
Z: "yeah in that case I think it contains sensitive ingredients" - participants (computer science)

Visually strengthen the reflection:
"The AI dish metaphor was nice, it would be helpful to put examples with the text, so people understand it clearer. It would be nice to have it more visual, that when you look back to it you immediately see what you need, visually you can work this out. Maybe it could be in a logical structure, the people/team first later the other ones."
- second facilitator (designer)

Example answers as guidance:
" Really love it, really good. It brings everyone on the same page. Only the dish
itself was not clear maybe you can write it is the output. I also liked it is structured." - participant (computer science)

Actions based on the second session

- Overall really liked, related to and appreciated (validated)
- Make it more visual so one can clearly see from a distance what is what part
- Put numbers on the boxes for order (clearer for the tech-oriented people)
- Keep the structured manner while making it more visual
- Different word for utilities
- Put example answers to guide people

These points of feedback will be integrated in the next iteration of the tool. The next step is validation with professionally working AI team within IBM.

IV Choreography

1 | Write the stages of the pathway:

Restaurant visitor — End user Actions — *How will he/she be impacted concerning his/her actions?*

Touchpoints — *How will the touchpoints be impacted?*

Line of interaction

The model Actions — *Where in the pathway does the model interfere towards the outside world? Will it take any decisions and what kind?*

The service provider — Front stage staff — *Will the model impact the internal processes? In what manner?*

Line of visibility

The kitchen staff — Backstage staff — *What actions are taken when to make the model according to the previous created features?*

The model — *Where in the pathway does the model interfere internally? Will it take any decisions and what kind?*

The kitchen processes — Support processes — *Is there any type of support needed when the model is deployed? When something goes wrong who will fix it?*

1st iteration design concept

First iteration evil cards (test 1)

Second iteration evil cards (test 2)

VI Surprise

1 | Pick an surprise card, what would happen with the models scenraio/service? What implications will it have for who?
2 | Does the service need some adjustments to be prepared?

Surprise	Implications	How to solve?
Place the unexpected scenario card here	*The end users?* *Society?* *Organization?* *Employees?*	

1st iteration design concept

Appendix O I First workshop iteration analyses

Testing session 1

The analyses of the session and the informal feedback interviews is subdivided by the separate canvases and presented in the following sections.

Evil AI:

The translation from the goal to really evil ideas is not completely clear. The designers did not use the Evil AI cards. I propose to have something more provocative in between the exercises, something that stimulates bad thinking. For some, the value such as inaccuracy was not completely clear. I already was thinking to shortly describe the values briefly on a card so everyone would have the same understanding.
One participant mentioned to really like releasing the worst side of himself.

J: " Evil thinking provides funny ideas"

All the designers were thinking from the end consumers perspective instead of the internal client, this is remarkable(not surprising), compared to the computer scientist' interviews earlier. This gives a really different perspective in the evil ideas as they were all unfair for the end-customer.
Also asking a few times during this phase of the workshop: "why is this unfair or inaccurate?" really helps the participants to reflect deeper and come up with the attributes.

Overpass:

The overpass has 2 different steps. During the session appeared it makes the exercise less clear and I better can divide the two steps separately. First give an exercise to change the evil things to desired attributes. Second to bridge the two different value attributes into principles or features that will go together.

L: " I think so too, because on one hand you have the unfair and the other one the inaccurate, then you need to firstly have an extra step to change them into positive and then afterward bridging them"

Value pathway:

The value path way was experienced as quite difficult. Participants mentioned it was a lot to think about at once. Although they also mentioned that if it is your project and you know more about the topic it is easier. They proposed to make clearer steps of what to think first, then second etc.

L: " I cannot imagine what will happen along the way"

Surprise cards:

The surprise was experienced as very good and interesting and mentioned as a good tool for validation, critical thinking and depth in the idea by representing surprises from real life cases.

M: " I think it gives much more depth, because this is just a journey, the perfect journey. I think the surprise cards give more depth"

Overall reflection:

Overall they really liked the topic, to think about it and the visual style of the workshop with a clear flow. The participants mentioned, non-designers might need more stimuli to step out of their normal thinking habits. For example, an energizer practicing association and disassociations could help.

Testing session 2

Evil:

Refection
Z: " The common thing here we treat people what they do not know"

Z: " Here are we more talking about the application i.s.o. the system itself"

The ingredients and categorization:
M: " we can have a really dis-balanced data set, the data set Is disbalanced on age I just saw"
Z: " I really like this session it really brings projects together"

Threshold:

Not clear text:
" What do you mean with the first what?"

Referring back to the AI dish:
M: " haha so I would say the ingredients haha, so yeah the data, is important"
G: "Hahah I think then also the recipe then?"

Angel's:

Reflection and translation into implementation
M: " What I learned in one of my ethics courses even if people consented than it still can be a problem if they do not get it"
Z: " Ah oke, then maybe we have even a button saying that the meeting is not going well"
M: " but if people consent and they know what they are consenting to, that is really fair"

Overpass:

Reflection
Z: ".. (consent) it is the purpose of this for this thing to come up"
M: "yes and that should also right"

A bit too fuzzy/difficult:
Z: " I am going to give back to Dasha as, are a little bit too vague to my taste"

Both goals a bit overlapping
Z: " but did not we put these already in the angels one"

Extra guidance needed: after I asked some questions and gave some examples more ideas in new and diverse directions came up.

Surprise

" 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."

Actions based on 2nd session

Evil:

- Overall liked and appreciated
- Change the structure of the evil part, already include evil ingredients, evil recipe. Although it was on the sheet it was not currently understood in that manner
- Keep the evil cards but change the words to one similar group
- Attributes and features are in data terms the same, find new words or explain these.

Angels advocate:

- Not really understood as it was both ways

- Iterate on this idea more and see how it can be changed, or unnecessary

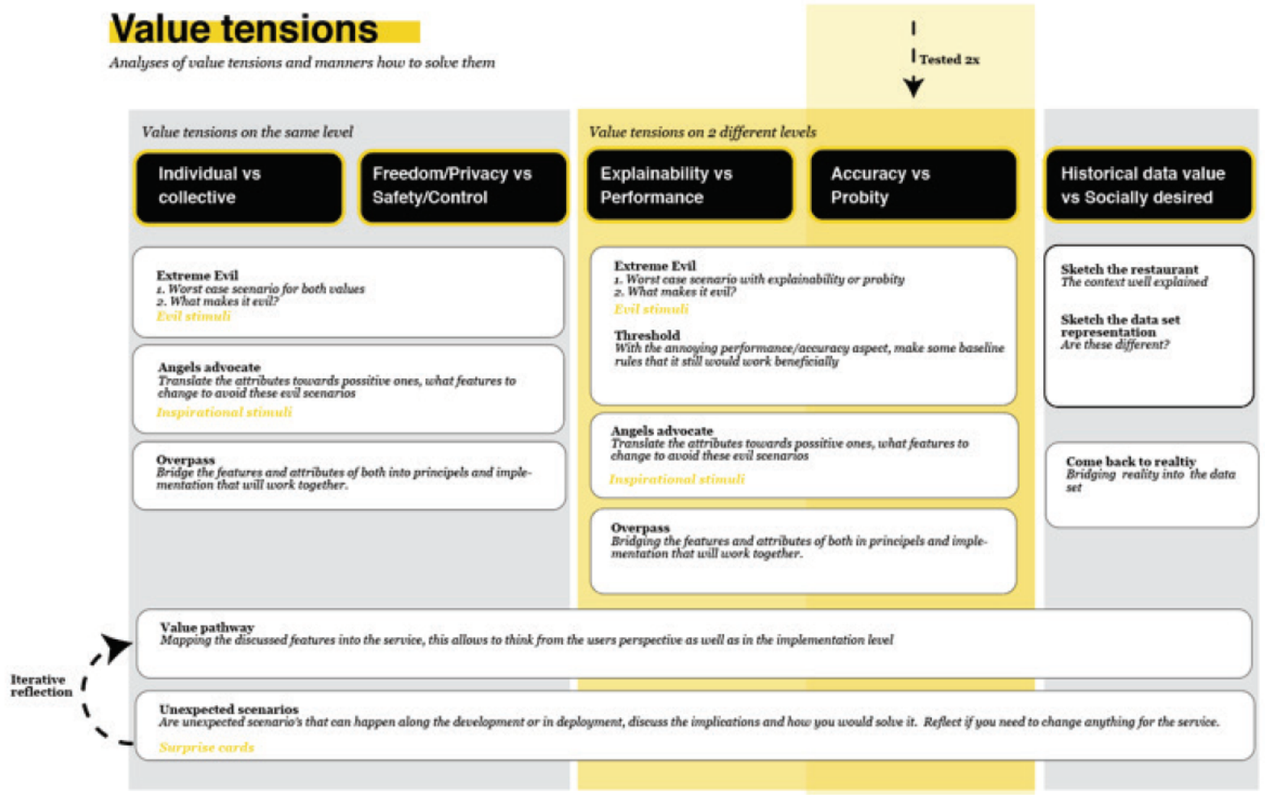
Threshold:
- Is related to as relevant but currently not too clear, therefore change the structure with numbers and formulate the goals clearer
- Formulate it in a manner in which the participants will write aspects they have an influence on.

Overpass:
- Too difficult and needs ideation around it

Surprise:
- Really appreciated
- Good reflective exercise
- I had the idea to do it with dices and make it a game.

Overall:
- With the first presentation win the trust and respect of the data scientists, with knowledge they do not know yet
- Put all the sheets on a wall in a line so easy go back to reflection
- Clear structure on the sheets, put numbers etc.
- Clear definitions of all words used as in different disciplines
- An example which resembles the input and output of all the sheets
- For the real session more, time is needed
- Multidisciplinary team is needed for a richer output
- Energizer when people do not know each other
- Ask maybe tips with someone who is a more experienced creative facilitator to spark creativity
- Remove sheets that are not essential from the flow as it is a bit much.
- The facilitator should ask the right questions and guide the users and have an active role during the session

Appendix P I Value tensions idea overview



In this figure the chosen value tensions with the shape workshop structure are visualized. Different type of values will need a different fashion to solve tensions between them due its nature. For example, a value such as accuracy has different dimensions then a value such as collective benefit. Thus, these are addressed differently.

Appendix Q I Nudging for fairness

Nudging for fairness

Goal: Actual checks & implementation of the shape workshop
Support during the AI process for a more fair AI System

Need

Currently there is a lack of actual implementation of ethics in the day to day work of the AI team. 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. Many ethical important decisions are made in these phases by one person. This thesis puts forward that during these phases actual support for the implementation of output of the shape workshop is needed in these identified moments.

Theoretical background

A nudge described by Leonard et al., (2008) "*A nudge, as we will use the term, is any aspect of the choice architecture that alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives. To count as a mere nudge, the intervention must be easy and cheap to avoid. Nudges are not mandates. Putting fruit at eye level counts as a nudge. Banning junk food does not.*"
The need for the 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). They have poor feedback. The effects are delayed (at the end of the development process) and the relationship between the choice made in de 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.

Design

The work of van Lieren et al. (2018), proposes nine strategies of behavioral override that serve as input for support for the ethical consultant role. These create moments of reflection and more conscious decision making (Van Lieren et al, 2018).
These are altered and tailored towards AI development for a fairer AI with small examples.
it is the task of the ethical coach 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 value pathway at scrum meetings/stand ups in the current AI development process. See the next page for the first iteration of the ethics fulfillment cheat sheet.

1. Add small friction

Adding small extra tasks in process to alter repetitive tasks is a strategy to make the AI team think once more about the choice. Examples of this are changing small things in an assignment every time one does it or implement extra ones with a small effort.

2. Increase decision moments in the process

By adding small extra decision moments in the process of the AI team, it supports them to think and reflect ethically about the project. These could be integrated in the sessions that the teams currently also have. The ethical aspect can be a reoccurring part on the agenda. The AI dish and the value pathway could be continuously hanging in the space in which the AI team works in order to easily reflect upon it.

3. Highlight loses and therefore active choice

In the AI development process, the loses and gains per decision can be made clearer by for ex-ample asking the data scientist to write down the choices he/she made with the loss and ad-vantage of it. This can be discussed in the current meetings that they have.

4. Personal ranking

If more ethical measures for projects assessments as well as individual assessments will be developed, then these could be compared life in order to stimulate more ethical choices. Another direction is to compare how often one asks an ethical question and make it a game.

5. Make commitment with an action plan

Partially the output of the shape workshop is an action plan with commitment. An idea would be to translate this into personalized for the project context, Hippocratic oath.

6. Checklists to easy remember Information

Implement checklists for checking AI models on algorithmic biases, incomplete training data and redundant encoding. Also, checklist can be implemented for relevant features decided upon during the shape workshop. It would be the role of the ethical consultant to summarize it in an easy to remember list. Also, to remind the AI team of the list and

7. Real-time feedback real time of consequences

Look back to the AI dish and value pathway during scrum meetings, what are the changes that the team decided upon afterwards and what consequences will this have on the fairness of the model?

8. Create Personalized feedback

Use the AI team's personal data for feedback on altering the AI system.

9. Create reminders & alerts

Small notifications and alerts made by the ethical consultant could support the AI team to re-member ethical decision moments and to check for sources of unfairness such as algorithmic bias.

Appendix | Design for fairness in AI

Ethics Fulfillment Cheat sheet

Support for different types of implementation of the shape workshop for the ethical consultant.

Rational Overrides (Van Lieren et al., 2018)	*Ethics building blocks*		
	Ethical people *Stimulate moral motivation, knowledge and responsibility (individual level)*	**Ethical processes & tools** *Stimulate explicit processes & decisions (team level)*	**Ethical company** *Stimulate ethical outcomes (company level)*
Add small friction	e.g. For the data scientist, add small friction in the programming system for reflection	e.g. Add small friction in the scrum meetings	e.g. Add small friction in the company processes
Increase decision moments in the process	e.g. Increase moments of ethical reflection for the data scientist individually	e.g. Add small decision in the scrum meetings	e.g. As a company, increase the decision moments and explanation behind the project choices
Highlight losses and therefore active choice	e.g. Highlight the disadvantages of certain choices made by the data scientist	e.g. Create a small exercise that shows the losses and disadvantages of a decision	e.g. Propose two different types of solutions, of which one is more ethical. Show the disadvantages visually
Rank Personally		e.g. Rank people based on the ethical questions & ideas they ask / propose	e.g. Rank people personally based on an ethical KPI
Create commitment with action plan	e.g. Make a personal hippocratic oath based on the shape workshop	e.g. Create a mutual commitment plan/ type of creative "contract" based on the shape workshop	e.g. Create company wide commitment for the ethical outcome of a project black and white
Create checklists to easy remember information	e.g. Make a personal checklist for the data scientist based on the shape workshop	e.g. Create a checklist based on the shape workshop that is adressed in the scrum meetins	e.g. Create companywide checklists for manager to check the ethical aspects of a project
Create real time feedback & real time concequences	e.g.	e.g. Look back at the scrum meetings to the value pathway, what kind of implications do these new decisions have? and for who?	e.g. VIsualize concequences of a certain decision
Create personalized feedback	e.g. Make the data scientist check the system with their own data	e.g. Create a randomized data picker from one person of the team	e.g. Create companywide KPIs for ethical aspects of AI projects
Create reminders & Alerts	e.g. Make notifications in the feature engineering/modeling phase to test for things as algorithmic bias	e.g. Make notifications in agendas of the team for moments of ethical reflection	e.g. Create companywide reminders for managers to include the ethical aspects in AI projects.

To check for

Data preparation, ideation & understanding :

- Algorithmic bias
- Incomplete data set
- Incomplete training data
- Subjective measurement of data
- Carefully choosing
- target variables

Modeling & feature engineering

- Redundant encoding
- Reinforcement feedback loops
- Self-fulfilling predictions

Evaluation & deployment

- Inconclusive evidence
- Untransparency
- Reinforcement of prediction

Interview guide English

This interview is concerning a thesis with Tu Delft and IBM, about the topic ethics in AI, more specific value-alignment and fairness. All information from this interview will be anonymized and the interview itself will be not shared with others. I would like to ask to record the interview for analyzing purposes only.

1 Process & Project – 15 min

I would like to ask you to be brief in answering the following questions.

1.1 Could you tell me briefly the assignment of this project?
- 1.1.1 What is the aimed deliverable of this project?

1.2 Could you describe me your role in this project?
1.2.1 For which activities do you feel responsible? Is it the same as is expected from you?

1.3 Could you map with who you are working on this project? (team) exercise 1
- 1.3.1 What are the roles?
- 1.3.2 Who reports to who? (hierarchy?)

1.4 Could you describe and map the process this project? Exercise 2.1
- 1.4.1 Could you use the stickers to map the relevant decision moments?
- 1.4.2 Who was involved in these decisions?
- 1.4.3 Could you use the stickers for important moments for you personally?

1.4.4 Could you map some challenging moments for you during the project?

1.5 Could you map the diverse tensions during the project? Exercise 2.2

2 Values - 15 min

2.1 Did you encounter any surprises during this project? Exercise 3.1

2.2 Could you pick values (if any) that are relevant for you personally in this project? Exercise 3.2
2.2.1Could you pick values (if any) that are relevant for the end-user in this project?
2.2.2 Could you pick values (if any) that are relevant for the client in this project?
2.2.3 Could you pick values that are consciously imbedded in the model?

3 Fairness – 10 min

3.1 Do you have any guidelines concerning the use of protected/unprotected features of data?

3.2 Could you explain me, based on which data categorization is made?
- 3.2.1 When in this process do you choose it?

3.3 How do you choose the measurement variables?
- 3.3.1 When in this process do you choose them?

3.3.2 Do you experience tension between the real-world complexity and the translation to the model? If yes where in the process?

3.4 Do you check for redundant encoding?
- 3.4.1 If so, then when?

3.5 Which real-life decisions will be based on this system?

3.6 How do you test this system?

3.7 What data was involved? (Please map) exercise 4
- 3.7.1 Who was the provider of the data?(Please map) exercise 4

3.8 What machine learning techniques do you use?

Extra

4.1 Do you sometimes use support like methods or tools during the process?
- 4.1.1 Why do you (not)?
- 4.1.2 What would be a convincing reason for you to use a tool/method in your project?

4.2 Do/did you experience any ethical challenges in this project?
- 4.1.1 If so do you try to resolve these? If so then how?
- 1.4.2 Do you reflect on your project?

Appendix S I The diverese challenges of value alignment

Challenges of value alignment

The value alignment problem is a challenging one (Ydkowsky, 2016). The following challenges are grouped per value alignment stage, to clarify the challenge which is addressed in this thesis. Identifying values will be given the most attention due the scope of this thesis. Firstly, the challenges of value-alignment will be described, after which will be touched upon some of the existing tools and methods that aim to tackle these.

Challenges in value alignment in Stage 1 I identifying values

Describing values

(in philosophy and psychology literature)

From a philosophical perspective values inform the foundation of ethical decision making. Values are also what we believe in and people believe in many things, in their mom, in the pope or prime minister for example. People find it hard to describe the values we have (Borning and Muller 2012; Van den Hoven, Vermaas, Van de Poel, 2015. p 84). As earlier mentioned, values can differ, for example, per individual, team, company, industry, country, culture etc. It is already experienced as complex to explicitly describe values to one's peers and evidently, teaching or programming these into code and an AI system occurs as a significant challenge. However, making the values explicit in the process, it argued to be crucial for innovation (Van den Hoven, Vermaas, Van de Poel, 2015)

Whose values

One of the difficulties of value alignment lies in whose values need to be programmed into the AI system: those of the end users; the AI teams, etc. (Kasenberg, and Scheutz 2017; Estrada, 2018; IEEE, 2018; Roos et al, 2018). There is not one set of universal values that can be programmed into the system. (IEEE, 2018; Arnold, Kasenberg, and Scheutz 2017; Estrada, 2018: Roos et al, 2018). And some argue that for proper alignment, participation in human moral communities is required (IEEE, 2018). The fields of philosophy, ethics and psychology have put much research effort in determining ways how discover whose values need to be integrated. However, the field is still divided. Some argue for a set of "core" values in this case moral values (Friedman, 2012), while others argue for situational and contextual values due the differences per context, culture (Borning & Muller, 2012). No clear strategy or solution for whose values to integrate in technological systems seems to be agreed upon in research.

Value tension

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)? (Friedman et al. 2013) 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). Human beings can resolve these by accepting trade-offs or see norms/values in more hierarchical relationships (IEEE, 2018), however for machines this is a complex task. So how should be address the value trade-offs in design and its implementation? An example of a value trade-off in a system: is an open calendaring system which supports group activities, awareness and presence over one's individual privacy.

In other words, not only which/whose values need to be integrated into an AI system, but how in a certain context, people prioritize norms/values as well. Thus, there is a need to do empirical research towards hierarchical relations and trade offs in certain industries and communities.

IEEE research institute pointed out, fixed hierarchical relations of values often do not fit. Thus, context specific value tradeoffs would be more suited. To achieve this user input will play a crucial factor to understand the subtle context specific differences that will fuel the value trade-off hierarchies in AI-systems. Also, in their paper is described that these value tradeoffs performed by an AI system should be transparent in order to give explanation and clarity about these to the involved stakeholders. Not addressing value tension in an explicit way can lead to a lack of appropriation by disadvantaged groups or even more drastic consequences such as system sabotage (Flanagan et al. 2005).

Within value tension, a specific challenge from computer science is increasing the value alignment challenge complexity, described by ¬¬both Stuart Russel and Eliezer Yudkowsky called: edge instantiation. Often AI systems aim is optimization, when optimizing something hard enough, one ends up in an undesired solution space. When a system optimizes a function as much as possible with an objective K depending on how much n is being optimized, this often leads to setting the remaining variables to very extreme values, this can lead to very undesirable outcomes (Russell and Norvig, 2016), such as the paperclip example.

Updating for future changing values and norms

Values, norms and morals are not static. They change in reaction to social progress, novel legal measures and other developments (Verbeek, 2012; IEEE, 2018). An example of this is the change of moral decision making is one with the rise of the echoscope concerning pregnant women. Firstly, due the visual view of the embryo and its hart, it strengthens the connections one has as a parent with the baby (as one has a visual image of the "human"). Secondly, it allowed to see things as autism before the baby is born. Leaving the moral decision to process with it the pregnancy with the parents, whereas before one was not able to make the decision as a parent (there are cases when the child sued their parents for not performing an abortion). This changed the moral decision space and therefore also some values we have in our lives as if a life is worthy to live.

How to account for these changing values in AI systems is a question that more and more researchers are aiming to tackle (Rossi, 2018). Humans have the quality to update new values and norms, they do observation and are sensitive to collective change of norms and naturally respond to feedback. For AI systems this is a complex task, which until now remains a question. IEEE recommends having a transparent way the system can change and alter norms, for traceability reasons, and to control how it learns when aiming to further research this area (IEEE, 2018). Creative strategic foresight might shed a new perspective on this challenge and new spaces for opportunity development.

Challenges in value alignment in stage 2 I Implementing values

Unforeseen instantiation

It is very difficult and slow to search for all possible possibilities that can happen with AI and embedding the values, and therefore complex to account for all the situation an AI needs to react according to certain values. (Yudkowsky, 2016).

Translation of value into design

It is challenging to translate some vaguer values as " fairness" into a code as the AI system does not know what it means. Therefore, it needs to be explicitly programmed and that is currently a challenge many AI teams are facing. (Soares & Fallenstein, 2014). This area how to get from values into design is understudied in literature (Van den Hoven, Vermaas, & Van de Poel, 2015).

Challenges in value alignment in stage 3 I Evaluating values

Context disaster

A crucial aspect in AI development is the context in which it is used. It can happen that one tests an AI application during development and it shows great results. However, when the context changes in which the AI works the results might change. Something which can be very beneficial

in situation 1 is not beneficial in situation two, however the program is the same (Yudkowsky, 2016). Therefore, the way values and AI's work are tested needs to be adjusted per context.
Nick Bostrom describes therefore "safety test objection" to AI catastrophe scenarios:" Safety test objection: An AI could be empirically tested in a constrained environment before being released into the wild. Provided this testing is done in a rigorous manner, it should ensure that the AI is "friendly" to us, i.e. poses no existential risk."

Nearest unblock strategy

Another complexity concerning evaluation of values is when AI's capacity will exceed. For now, the values and the code and perseverance of " wrong" behavior, however when AI will have more capacity or become " more intelligent" then this type of encoding values or preserving it with a code overlay might not work anymore. (Yudkowsky, 2016).

No one standard matrix or checklist

Due to the two above mentioned challenges lead towards another, that is hard, if it is even possible, to make one universal checklist for AI value evaluation. Per specific application and context, it is different.

To conclude from the above-mentioned challenges is that goals that are simply to specify (and programmed/learned to an AI system) cannot account for contextual ramification of the real world human values and goals. (Yudkowsky, 2011). Humans want many different things, in specific ways, context specific(Soares, 2015). Nevertheless, research and practice put effort into addressing these challenges with a variety of support offered for practitioners which will be discussed in the following section.

Appendix T I Approved project brief

2930

DESIGN FOR our future

TU Delft

IDE Master Graduation

Project team, Procedural checks and personal Project brief

This document contains the agreements made between student and supervisory team about the student's IDE Master Graduation Project. This document can also include the involvement of an external organisation, however, it does not cover any legal employment relationship that the student and the client (might) agree upon. Next to that, this document facilitates the required procedural checks. In this document:

- The student defines the team, what he/she is going to do/deliver and how that will come about.
- SSC E&SA (Shared Service Center, Education & Student Affairs) reports on the student's registration and study progress.
- IDE's Board of Examiners confirms if the student is allowed to start the Graduation Project.

USE ADOBE ACROBAT READER TO OPEN, EDIT AND SAVE THIS DOCUMENT
Download again and reopen in case you tried other software, such as Preview (Mac) or a webbrowser.

STUDENT DATA & MASTER PROGRAMME
Save this form according the format "IDE Master Graduation Project Brief_familyname_firstname_studentnumber_dd-mm-yyyy".
Complete all blue parts of the form and include the approved Project Brief in your Graduation Report as Appendix 1 !

family name: Simons
initials: D.P. given name: Daria
student number:
street & no.:
zipcode & city:
country:
phone:
email:

Your master programme (only select the options that apply to you):
IDE master(s): IPD / DfI / SPD
2nd non-IDE master:
individual programme: - - (give date of approval)
honours programme: Honours Programme Master
specialisation / annotation: Medisign / Tech. in Sustainable Design / Entrepeneurship

SUPERVISORY TEAM **
Fill in the required data for the supervisory team members. Please check the instructions on the right !

** chair: Elisa Giaccardi dept. / section: HICD
** mentor: Lianne Simonse dept. / section: PIM
2nd mentor: Zoltan Szlavik
organisation: IBM
city: Amsterdam country: Netherlands

comments (optional)

Chair should request the IDE Board of Examiners for approval of a non-IDE mentor, including a motivation letter and c.v..

Second mentor only applies in case the assignment is hosted by an external organisation.

Ensure a heterogeneous team. In case you wish to include two team members from the same section, please explain why.

TU Delft

Procedural Checks - IDE Master Graduation

APPROVAL PROJECT BRIEF
To be filled in by the chair of the supervisory team.

chair Elisa Giaccardi date 19 - 10 - 2018 signature

CHECK STUDY PROGRESS
To be filled in by the SSC E&SA (Shared Service Center, Education & Student Affairs), after approval of the project brief by the Chair.
The study progress will be checked for a 2nd time just before the green light meeting.

Master electives no. of EC accumulated in total: 49,5 EC
Of which, taking the conditional requirements into account, can be part of the exam programme 31 EC
List of electives obtained before the third semester without approval of the BoE

(✓) YES all 1st year master courses passed
( ) NO missing 1st year master courses are:

name D. Hamster date 16 - 11 - '18 signature

FORMAL APPROVAL GRADUATION PROJECT
To be filled in by the Board of Examiners of IDE TU Delft. Please check the supervisory team and study the parts of the brief marked **.
Next, please assess, (dis)approve and sign this Project Brief, by using the criteria below.

- Does the project fit within the (MSc)-programme of the student (taking into account, if described, the activities done next to the obligatory MSc specific courses)?
- Is the level of the project challenging enough for a MSc IDE graduating student?
- Is the project expected to be doable within 100 working days/20 weeks ?
- Does the composition of the supervisory team comply with the regulations and fit the assignment ?

Content: (✓) APPROVED ( ) NOT APPROVED
Procedure: (✓) APPROVED ( ) NOT APPROVED
comments

name A Huwae date 27 - 11 - 2018 signature

Initials & Name Student number
Title of Project