

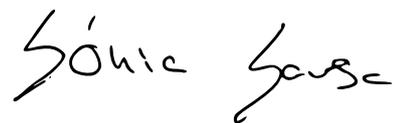
Cyprus University of Technology  
Tallinn University  
MSc Interaction Design

# **A TOOLKIT TO ENABLE THE DESIGN OF TRUSTWORTHY AI**

Master's thesis

Author: Stefan Schmager

Supervisor: Dr. Sónia Cláudia da Costa Sousa



Tallinn, 2021

# Author's declaration

I declare that apart from work whose authors are clearly acknowledged, this manuscript is a product of the author's original work, and it has not been published before, neither is not currently being considered for publication elsewhere for any other comparable academic degree.

This master thesis document has been supervised by Dr. Sónia Cláudia da Costa Sousa (Tallinn University, Estonia).

Author: Stefan Schmager

A handwritten signature in black ink, appearing to be 'Stefan Schmager', written in a cursive style.

Date: 04.05.2021

**Non-exclusive license to reproduce a thesis and make thesis available to public**

I, Stefan Schmager (date of birth: 05.05.1987), grant Cyprus University of Technology and Tallinn University a permit (a non-exclusive license) to reproduce for free and make public in the repository of Tallinn University Academic Library a piece of work created by me,

*“A toolkit to enable the design of trustworthy AI”*,

supervised by Dr. Sónia Cláudia da Costa Sousa. I am aware of the fact that the author also retains the rights mentioned above. I certify that granting the non-exclusive license does not infringe the intellectual property rights of other persons or the rights arising from the Personal Data Protection Act.



Heidelberg, 04.05.2021

# Acknowledgements

I would like to thank my supervisor Dr. Sónia Cláudia da Costa Sousa of the School of Digital Technologies, Tallinn University. She not only supported me with constant guidance and direction to ensure I keep my focus along the way. But she also provided expert feedback and knowledge on the research, contributing tremendously to the advancement of the research outcome.

Further, I would like to thank all the workshop participants, who provided their time, knowledge and experience from working in the field of interaction and user experience design. Their critical feedback and fruitful conversations helped to shape and co-design the toolkit for trustworthy AI.

And finally, I would like to express my gratitude to my family and friends who always supported, motivated and encouraged me throughout the whole Master's program.

# Abstract

Technological progress in artificial intelligence (AI) and machine learning (ML) has an enormous impact on our society, economy and environment. And although the urgent need for creating sustainable and ethical AI technology is admitted, there exists a lack of design tools and expertise to facilitate this advancement. This study investigates how to help designers design for the value of trust in AI systems. A literature review unveiled a myriad of ethical AI principles as well as gathered existing tools addressing the research area. Iterative reviews together with an expert on trust in technology evaluated these guidelines and a toolkit prototype containing 29 design principles had been created. Through multiple participatory design workshops the next iteration of the toolkit was co-designed in collaboration with design professionals. The result is an iterated toolkit comprising 16 principles relevant in the design for trust in AI systems, and providing tool suggestions for each principle.

# Table of Contents

<b>1. Introduction</b>	<b>11</b>
1.1. Research problem and significance	12
1.1.1. Research goal and motivation	13
1.1.2. Research questions	13
1.2. Research methodology	13
1.2.1. Research process	14
1.2.2. First pillar: Literature review	15
1.2.3. Second pillar: Design toolkit prototype	15
1.2.4. Third pillar: Prototype evaluation	16
<b>2. First pillar: Literature review</b>	<b>17</b>
2.1 Literature review methodology	17
2.1.1. Stage 1: Identification of the research question	17
2.1.2. Stage 2: Identify relevant studies	18
2.1.3. Stage 3: Study selection	19
2.1.4. Stage 4: Charting the data	20
2.1.5. Stage 5: Collating, summarizing and reporting the results	23
2.2. Theoretical contextualization	23
2.2.1. Artificial intelligence	23
2.2.2. Ethical AI	24
2.2.3. Human-centered AI	26
2.2.4. Ethical AI guidelines	27
2.2.5. Applying guidelines in practice	27
2.2.6. Trust in technology	28
2.2.7. Trust in AI	29
2.2.8. Trustworthy AI	30
2.2.9. Explainable AI	31
2.3. Existing tools & methods	31
2.4. Reflection	33
<b>3. Second pillar: Design of the toolkit prototype</b>	<b>34</b>
3.1 Research problem and strategy	34
3.2. Toolkit iteration 1: Process	35
3.2.1. Participants	36
3.3. Identifying design principles for trustworthiness	37
3.3.1. Collection of common ethical guidelines in the literature	37
3.3.2. Synthesizing ethical principles	38
3.3.3. Identification as design problems	39
3.3.4 Toolkit prototype	41
3.4. Reflection	43

<b>4. Third pillar: Evaluation &amp; iteration of the prototype</b>	<b>44</b>
4.1. Toolkit iteration 2: Process	44
4.1.1. Workshop setup	46
4.1.2. Workshop participants	46
4.2. Workshop activities	47
4.2.1. Workshop introduction	47
4.2.2. Mapping activity	47
4.2.2.1. Design process mapping	49
4.2.2.2. Trust quality filter	51
4.2.2.3. Optional tool suggestion	53
4.2.3. Collaborative activity	53
4.2.3.1. Contextual laddering	54
4.2.3.2. Toolkit prototype evaluation	55
4.3. Results	55
4.3.1. Removal of non-relevant principles	56
4.3.2. External redundancy mitigation	57
4.3.4. Semantic consolidation	58
4.3.3. Tools and methods suggestions	60
4.3.4. Evaluation	60
4.3.5. Tool identification	61
4.5. Reflection	61
<b>5. Discussion</b>	<b>63</b>
5.1. The results achieved	63
5.2. Limitations of the study	65
5.4. Ideas for further research	66
5.5. Conclusion	67
<b>6. References</b>	<b>69</b>
<b>7. Appendices</b>	<b>83</b>
7.1. Appendix 1: Existing tools	83
7.2. Appendix 2: Collections of ethical AI principles	85
7.3. Appendix 3: Exemplary card format of the toolkit	88

# List of Figures

<b>Figure 1:</b> PRISMA Flow chart of the study selection process	19
<b>Figure 2:</b> Framework to identify relevant principles for the design of trustworthy AI	36
<b>Figure 3:</b> Framework to validate prototype of the trustworthy AI toolkit	45
<b>Figure 4:</b> Simplified illustration of the Double Diamond design process model	49

# List of Tables

<b>Table 1:</b> Research process	14
<b>Table 2:</b> Identified key concepts from the literature review	21
<b>Table 3:</b> Toolkit prototype - including principles identified as design problems	41
<b>Table 4:</b> Likert scale evaluation choices	55
<b>Table 5:</b> Toolkit for the design of trustworthy AI	63
<b>Table 6:</b> Identified existing tools, methods and frameworks	83
<b>Table 7:</b> Collections of ethical AI principles	85
<b>Table 8:</b> Exemplary card format of the toolkit for the design of trustworthy AI	88

# List of Abbreviations

<b>AI:</b>	Artificial intelligence
<b>AI-HLEG:</b>	High-Level Expert Group on Artificial Intelligence
<b>HCD:</b>	Human-Centered Design
<b>HCI:</b>	Human-Computer Interaction
<b>HCTM:</b>	Human-Computer Trust Model
<b>ISO:</b>	International Organization for Standardization
<b>MAIEI:</b>	Montreal AI Ethics Institute
<b>ML:</b>	Machine learning
<b>PBC:</b>	Perceived behavioural control
<b>TAI:</b>	Trustworthy AI
<b>TAM:</b>	Technology Acceptance Model
<b>UNDG:</b>	United Nations Development Group
<b>UX:</b>	User Experience
<b>WCAG:</b>	Web Content Accessibility Guidelines
<b>XAI:</b>	Explainable AI

# 1. INTRODUCTION

---

Humanity is situated in the midst of a digital transformation. Technological progress, especially in the fields of artificial intelligence (AI), machine learning (ML) and autonomous systems, has a tremendous impact on society, the environment and the wider economical context. However, it is widely acknowledged amongst experts, that although there is an urgent need for finding ways and concepts to develop sustainable AI, the topic is vastly underrepresented (Pooley, 2020). And unfortunately, design as a profession, and as a craft, follows this trend and is vastly underutilised and misunderstood in the creation of these modern technologies. It seems that for every technological leap, design is overlooked and has to prove its value and contribution again until it is understood as a crucial asset in the process. The traditional view is that design is a technical and value-neutral task of developing artifacts that meet functional requirements formulated by clients and users (Van den Hoven et al., 2015). According to Peter-Paul Verbeek (Coeckelbergh, 2020, p.45) technology shouldn't be seen as a threat, but rather that humans are technological, meaning they have always used technology. And design should be looked at as a tool that services humanity. AI systems will impact all dimensions of our lives, from commercial and social interactions to relationships with the state, including dramatic structural transformations in the public sphere. Hence the design of such technologies can't be value-neutral. Systems using artificial intelligence (AI), machine learning (ML) and other advanced technologies are artifacts built by people to fulfill specific goals. The theories, tools and methods need to integrate societal, legal and moral values into the development process of these systems at all stages of the design and development process to ensure human flourishing and wellbeing in a sustainable world (Dignum, 2018). However, it seems that design practices aren't keeping pace and struggle to incorporate human values and algorithmic logic together into socially, economically and politically sustainable models. There is a lack of knowledge, skills and roles in the field of design to support and enable the creation of beneficial technologies. The range of opportunities is vast, but we need to take the chances to not miss our duty of influencing this transformation. The design and development of AI technology need to include ethical and human values into their processes to create human-centered products, services and systems. It is pertinent to shape the development of the technology while it is still possible.

## **1.1. RESEARCH PROBLEM AND SIGNIFICANCE**

Guidelines on how to build technology in ethical and sustainable ways are nothing new. For the field of artificial intelligence alone, recent studies have demonstrated that there exists an abundance of guidelines, created by research institutes, private companies and the public sector (Hagendorff, 2020; Jobin, Ienca & Vayena, 2019; Fjeld et al., 2020). But although such ethical principles and guidelines ought to shape the development and implementation of ethical technologies, their existence is not without criticism. Morley et al. (2019) have found that the tools and guidelines being developed and provided to address AI ethics are often difficult to map with regards to the categories or principles they could help to address. According to McNamara et al. (2018), studies suggest that guidelines have little impact on the practices surrounding AI development as they lack real implementation mechanics and assessment practices that would turn guidelines into more ethically aware development. This indicates that existing guidelines on ethical, explainable and trustworthy AI are too abstract and difficult to put into practice. In consequence, it's difficult for designers and developers of AI technologies to determine which ethical concerns they should be aware of, how these can present themselves and how they may be addressed (Ryan & Stahl, 2020). Another issue of the existing AI ethics guidelines is their aim to address a wide spectrum of potential stakeholders, from policymakers, users and developers to educators, civil society organisations, industry associations, professional bodies and more (Ryan & Stahl, 2020). But unfortunately, current guidelines are often difficult to understand as they tend to be written for technical users who constitute only one key user group. AI is depicted as a common and collective endeavour for which humanity cooperates. But more often than not it is argued that ethical problems can be addressed with technical fixes. Yet this perspective would position advanced modern technology, and AI in particular, as something outside of the social-relational structure and obscure the larger, often socio-technical problems. To make AI policies work, it is important to build a bridge between abstract high-level ethical principles defined by research institutes, companies or nation-states and the practices of technology development and use in their particular contexts (Coeckelbergh, 2020, p. 170). To overcome this misconception, humans need to be involved into the process to incorporate human values into the developed products, services and systems (Boden, 2016). And as designers need to be able to work on the fuzzy frontend of the development of AI technology, they need to be equipped with tools and methods to ensure practical implementation of theoretical AI guidelines.

### 1.1.1. RESEARCH GOAL AND MOTIVATION

The goal of this thesis is to develop a design toolkit which has the objective of enabling and catalyzing the design of trustworthy AI. The focus of this work will be on the concept of “Trustworthy AI” (TAI), which has relations to concepts like “trust”, “transparency”, “explainability” and more. The toolkit is meant to help designers consider theoretical AI guidelines and principles during their design process and act as supporting guidance to decide how the nature of an AI system, its context and implementation will affect the emergence of trustworthiness. It will contain a set of principles related to the formation of trustworthiness, and suggested methods and instruments for different phases of the design process. The aim is to provide assistance during the design process, helping designers to define the appropriate design for a given use case. By co-designing the toolkit with users and stakeholders involved in the process, it is ensured to create the most useful and beneficial toolkit for its later users.

### 1.1.2. RESEARCH QUESTIONS

Based on the research goal, one main research question, as well as three sub-research questions can be defined:

**Main research question:** How to help designers design for trustworthy AI?

- *Sub-RQ1:* What current design practices exist to support the translation of existing trustworthy AI guidelines into practice?
- *Sub-RQ2:* How to create a toolkit that will enable designers to design for the formation of trustworthy AI systems?
- *Sub-RQ3:* How does the toolkit enable the design of trustworthy AI systems in the dimensions of “usefulness”, “satisfaction” and “efficiency”?

## 1.2. RESEARCH METHODOLOGY

As the methodological foundation for this study, a research through design approach is followed. It focuses on the contribution to the field of Human-Computer Interaction (HCI) and the creation of new knowledge. According to Zimmermann et al. (2007, p.497), by following research through design methodology, designers produce novel integrations of HCI research with the goal of creating a product that transforms the world from its current state to a preferred state. New opportunities for advancement in current technology that will have a

significant impact on the world are identified. Or as Pärnpuu (2020, p.10) describes it, “research through design aims to produce artifacts that become design exemplars, providing an appropriate conduit for research findings to easily transfer to the HCI community.” For proper implementation of the research through design approach, Zimmerman et al. (2007, p.499) suggest a formalized application, by applying four lenses to evaluate the contribution:

1. The contribution needs to be documented in a way that the process can be reproduced.
2. The way the specific subject matter is addressed needs to fulfill the characteristic of novelty.
3. The design research needs to be relevant, instead of self-indulging without any real-world impact.
4. The created knowledge needs to be understandable and usable for the further design community.

### 1.2.1. RESEARCH PROCESS

The structure of this research process consists of three pillars, which are based on each other. A literature review, an initial draft of a toolkit prototype and a participatory co-design iteration of the toolkit and its evaluation. Each of these process elements has a defined set of goals, a specified research method and expected outcomes. The outcome of each pillar builds the foundation for the next step in the process (Table 1).

**Table 1:** Research process

PILLAR	1. LITERATURE REVIEW	2. DESIGN TOOLKIT PROTOTYPE	3. PROTOTYPE ITERATION
INPUT	Research sub-question 1	Research sub-question 2	Research sub-question 2
			Research sub-question 3
GOALS	Identifying key concepts	Identifying relevant principles for the design of trustworthy AI	Co-designing next iteration of the toolkit with practitioners
	Identifying research gaps	Shortlisting principles to include in prototype	Evaluating toolkit
	Collecting existing tools & methods		
METHOD	Scoping review	Expert evaluation	Participatory design workshop
	Web search		

<b>PARTICIPANTS</b>	Researcher (Author)	Researcher (Author) Trust expert	Researcher (Author) Design professionals
<b>OUTCOMES</b>	Conceptualization (theoretical foundation)	Toolkit prototype	Iterated toolkit
	Collection of existing tools		Toolkit evaluation

### 1.2.2. FIRST PILLAR: LITERATURE REVIEW

The first pillar of the research procedure is a literature review, focusing on the topics of ethical AI, AI guidelines and current HCI design practices which support the translation of existing AI guidelines into practice. A scoping review approach is applied to address the broad topic, as many different study designs might be applicable. The literature review serves two purposes: First, it examines the extent, range and nature of existing research activity, to identify and define key concepts. Particular attention is directed towards identifying common principles for the design and development of trustworthy AI in the literature. Furthermore, it also aims to identify gaps in the literature (Arksey & O'Malley, 2005, p.6). The second goal is to assemble a collection of existing tools and methodologies for the translation of ethical AI guidelines into practice. This collection is supplemented with a dedicated search for available tools and methods, also outside of the academic spectrum and the narrow topic of trustworthy AI. The compiled collection builds the foundation for the second pillar of the research - the design of the toolkit prototype. In regards to the four lenses of the research through design methodology, the literature review fulfills the criteria of "process" and "extensibility" by outlining why the specific established review process has been chosen as well as the steps taken so it can be reproduced. In addition, the literature review states its relevance by demonstrating how the contribution of the work advances the current state of the art in the research community in the field (Zimmerman, 2007, p.499).

### 1.2.3. SECOND PILLAR: DESIGN TOOLKIT PROTOTYPE

The second pillar of the research procedure constitutes the identification of common and relevant principles for the design of trustworthy AI systems. Based on the literature review, the most commonly applied and deployed principles are collected and synthesised from a wide body of existing documents, guidelines and policies. In a next step, the identified

principles are assessed against the requirement, whether they can be formulated as a design problem. To be understood as a design problem, a scope of a principle needs to be applicable and influenceable from a design perspective. Only those principles which can be approached from a design angle are considered to be relevant for the design of trustworthy AI. In the next step, the remaining principles which are understood as design problems are assessed against the trust qualities identified by Gulati et al. (2019). If a principle can be associated with one of the three trust qualities - risk perception, benevolence or competence - it is further on considered as a relevant dimension for the design of trustworthy AI. The principles, which fulfill both evaluation criteria, the understanding as a design problem and to match a trust quality are relevant principles for the design of trustworthy AI. Further, the list of existing methods and tools is scrutinized towards their potential contribution to the design of trustworthy AI systems. The existing tools and methods are evaluated whether they are useful when addressing particular principles. The outcome of this research pillar will be an initial draft of a novel toolkit, that will be further iterated on in the next pillar, and therefore also fulfills the criteria of invention and extensibility.

#### **1.2.4. THIRD PILLAR: PROTOTYPE EVALUATION**

In the third pillar of the research, the initial prototype of the toolkit is iterated on in a participatory design workshop together with designers. The motivation of conducting a participatory workshop together with designers is to continuously involve the envisioned “users” of the toolkit in all stages of the toolkit development. The goal of the workshop is to co-design the next iteration of the toolkit, considering the dimensions “usefulness”, “satisfaction” and “efficiency” and to refine its composition. Ideally, within the workshop, the iterated toolkit would also be evaluated by the practitioners in the aforementioned dimensions. In regards to the lenses of the research through design methodology, a participatory design workshop constitutes an established format of co-designing. However, as the course of the workshop is of a dynamic nature, the criteria of process and reproducibility might not be fulfilled. Nevertheless, the outcome of the workshop will be an iterated and novel artifact, which fulfills the criteria of invention and extensibility. The fulfillment of the “relevance” criteria will be fulfilled by a critical evaluation of the toolkit in the workshops.

## **2. FIRST PILLAR: LITERATURE REVIEW**

---

As the goal of this Master's thesis is to contribute novel knowledge, methodology and tools to the field of Human-Computer Interaction (HCI) and to the advancement of designing for trustworthy artificial intelligence, a thorough literature review is pertinent. For this study, the contribution of the literature review is twofold. The first goal of the literature review is to understand the current state of the field of designing for trustworthy AI and to identify the key concepts within. The second goal is to identify tools, techniques and frameworks which already exist or are related to the design for trustworthy AI.

### **2.1 LITERATURE REVIEW METHODOLOGY**

For this study, the suggested framework from Arksey & O'Malley (2005, p.8) is adopted to conduct a scoping study. Hereby, it covers many different study designs within a broad scope. Colquhoun (2014) states that the "scoping review constitutes a knowledge synthesis which addresses an exploratory research question by mapping key concepts, types of evidence and gaps in research by systematically searching, selecting and synthesizing knowledge." According to Arksey & O'Malley (2005, p.6) there are various reasons to conduct a scoping study, including examining the extent, range and nature of research activity as well as identifying research gaps in the literature. Further, a scoping study is guided by a requirement to identify all relevant literature regardless of study design. These aspects render a scoping literature review as the most adequate choice for this research work.

#### **2.1.1. STAGE 1: IDENTIFICATION OF THE RESEARCH QUESTION**

In the first stage of the literature review process, the research question needs to be identified as it guides the way that search strategies are built. As described above, the main research question of this study is "*How to help designers design for trustworthy AI?*" The literature review contributes to this main research question in two ways. First, it examines the breadth, coverage and character of the existing research and it identifies key concepts by exploring the existing body of knowledge. It helps to understand the current state and directions of the research that has been done in the field. Further, it allows to synthesise relevant and pertinent research themes, as well as defining the limitations of this particular research endeavour. The

determined concepts will be charted and gaps in the research identified, which will establish the theoretical foundation of this work.

The second goal of the literature review is directly addressing the research sub-question 1: SRQ1: *“What current design practices exist to support the translation of existing trustworthy AI guidelines into practice?”*. The literature review will identify existing tools, techniques and frameworks to create a collection of available practices. This collection will constitute the possibility space and foundation for the following research methods. The motivation for collecting already existing tools and methods is to evaluate their suitability and adaptability for the specific use-case of designing for trustworthy AI. In the further course of this work, the list of existing tools will affect the outcome of research sub-question 2: SRQ2: *“How to create a toolkit that will enable designers to design for the formation of trustworthy AI systems?”*

### **2.1.2. STAGE 2: IDENTIFY RELEVANT STUDIES**

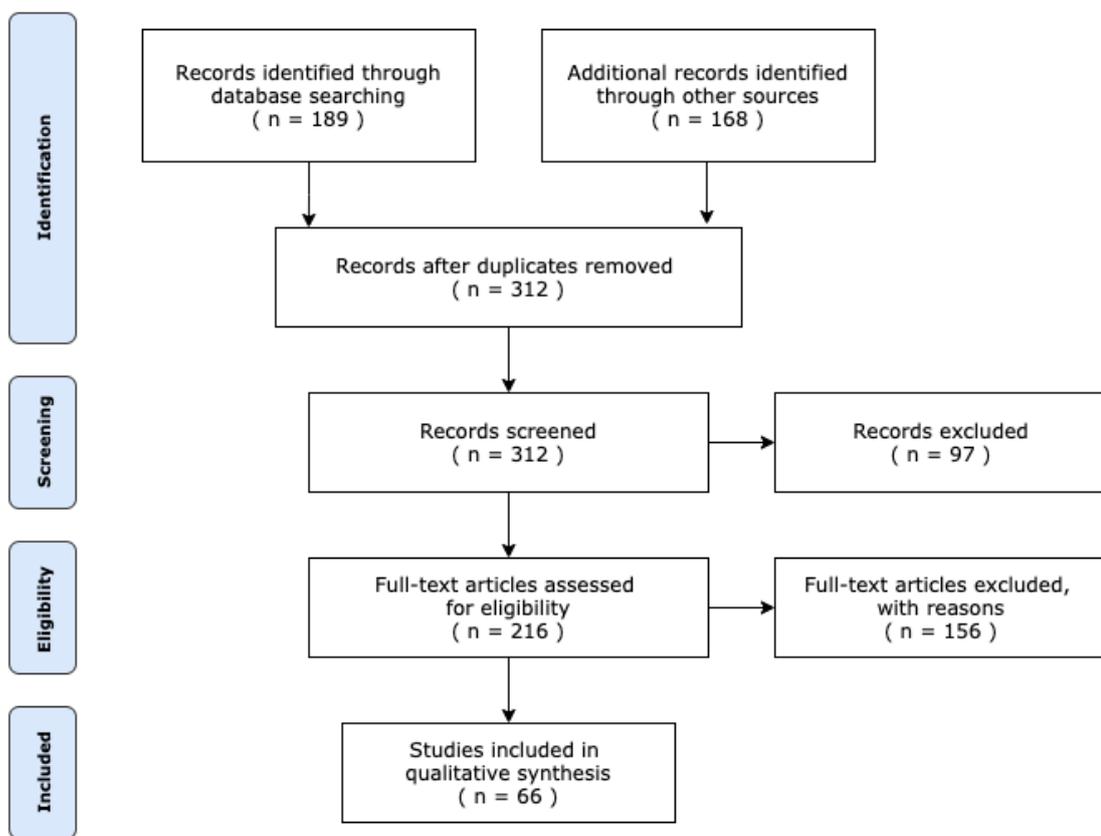
The second stage of the scoping study is to identify the relevant studies to include in the review, by defining inclusion and exclusion criteria. I relied on three search strategies, namely database search, web search and citation chaining. As a baseline, the time span for studies to include is set from 2021 back until the year 2000, with a focus on the more recent past, to address the empirical work concomitant with the latest research development. But although research on the field of designing for ethical and trustworthy artificial intelligence has increased in the more recent past, the time span was chosen to not miss potentially interesting speculative studies that had been conducted earlier. The focus of this work is on understanding the current state of research for trustworthy AI, as well as methods and tools to design for it. The defining search strings were *“design for trust”*, *“design for trust toolkit”*, *“design for trustworthy AI”*, *“designing for trust tool”*, *“ethical design tool”*, *“ethical design toolkit”*, *“human-centered AI”*, *“trust in AI”*, *“trust sensitive design”*, *“trust toolkit”*, *“trust-centered design”* and *“trustworthy AI AND ethics guidelines”*. A pilot search for studies was performed on the Google Scholar platform ([scholar.google.com](https://scholar.google.com)) as it provides one of the most comprehensive collections of available studies (Martín-Martín et al., 2020, p. 900). The refined search has been repeated on the ACM Digital Library ([dl.acm.org](https://dl.acm.org)) because it provides a more specific scope of peer-reviewed articles for the field of HCI and Interaction Design. Further, the search also included studies identified by the references stated in the

initial studies (Arksey & O'Malley, (2005, p.12). In total, the study search resulted in a list of 361 studies.

### 2.1.3. STAGE 3: STUDY SELECTION

The study selection process went through several stages, following the PRISMA statement and PRISMA flow diagram by Moher et al. (2009).

Figure 1: PRISMA Flow chart of the study selection process (Moher et al., 2009)



In a first iteration, the identified articles were screened based on the abstract, whether they are relevant to trust in AI, ethical AI, ethical AI guidelines, human-centered AI or if they introduce a tool for the design of trustworthy AI. At the same time, the list of studies was also checked for duplicates that have been removed from the list. After the initial assessment, the total number of studies was reduced to a number of 312. During the pilot search, it became clear that a big part of available studies are addressing very specific and sensitive sub-fields for which artificial intelligence is becoming increasingly relevant, such as health, education,

military use and robotics. But as designing for those specific contexts also requires specific domain knowledge, studies from these fields were excluded as they go beyond the scope of this work. As Arksey & O'Malley (2005, p.10) state, adjustments to the search parameters can be made, once some sense of the volume and general scope of the field has been gained. This excluded 97 studies from the total list. In the final step of the study selection, the remaining articles were evaluated for eligibility for this study. The eligibility criteria for this selection process were, if a study can contribute to answering the research questions and whether it describes an existing tool or method to put theoretical AI guidelines into practice. In particular, this means that studies are evaluated if they fulfill any of the four inclusion criteria: contribution to the understanding of the current state of research on trustworthy AI, describing a method for designers to approach trustworthy AI design guidelines, discussing a current design practice for designing trustworthy AI or if it describes how to translate existing AI guidelines into practice. After this assessment, a total of 66 eligible, non-duplicate documents related to the topics were identified and went into the literature review.

#### **2.1.4. STAGE 4: CHARTING THE DATA**

The reviewed literature is charted in a tabular format, identifying the key concepts discussed in the literature (Table 2). This also includes the overarching concepts like ethical AI, ethical AI guidelines and their implementation, as well as adjacent topics, like trust in artificial intelligence systems. For each reviewed source, the discussed concepts have been mapped within the table. At the end of the process, the total count of occurrences within the literature has been calculated. Due to the breadth of the field of artificial intelligence, the focus is limited to a generic understanding of systems and products. Concepts that are mentioned less than five times have been dropped from the table, to keep the results relevant and provide a better understanding of specific focus areas in the current body of research.

**Table 2:** Identified key concepts from the literature review

	Ethical AI	Human-centred AI	Applying guidelines in practice	Applying guidelines in practice	Trust	Trust in AI	Trustworthy AI	Explainable AI	Existing Tool
AI Ethics Impact Group, 2020	X		X	X					X
AI-HLEG, 2020	X		X				X		X
Amershi et al., 2019		X	X						
Antonov & Kerikmäe, 2020	X						X		
Arnold et al., 2019					X			X	X
Auernhammer, 2020	X	X							
Bitkina et al., 2020						X			
Bostrom, 2014	X								
Dignum, 2018	X								
Doran et al., 2017								X	
Ehsan & Riedl, 2020		X						X	
Ferrario et al., 2019	X	X			X				
Fjeld et al., 2020	X		X	X			X		
Floridi, 2019	X		X				X		
Floridi & Cowls, 2019	X		X						
Floridi et al., 2018	X		X						X
Floridi et al., 2020	X	X							
Friedman & Hendry, 2012									X
Gulati et al., 2019					X				
Guszcza et al., 2020	X	X	X						X
Hagendorff, 2020	X		X						
Hagerty & Rubinov, 2019	X								
Hoffman et al., 2018					X			X	
Jobin et al., 2019	X		X						
Leijnen et al., 2020	X		X						X

Leong & Iversen, 2015									X
Leslie, 2019	X								
Lewis et al., 2020			X				X		
Li et al., 2008					X				
Liao et al., 2020		X						X	
Manders-Huits & Zimmer, 2009			X						
Mcknight et al., 2011					X				
McNamara et al., 2018	X		X	X					
Mittelstadt, 2019a	X		X						
Mittelstadt, 2019b	X		X						
Morley et al., 2019	X			X					X
Mucha et al., 2020									X
Nickel, 2015					X				
Raftopoulos, 2015	X								X
Rességuier & Rodrigues, 2020	X		X						
Riegelsberger et al., 2005					X				
Rossi, 2018					X		X	X	
Ryan, 2020			X				X		
Ryan & Stahl, 2020	X			X					
Sankaran et al., 2020						X			
Schmidt et al., 2020						X		X	
Shneiderman, 2020a		X					X		X
Shneiderman, 2020b		X					X		X
Shneiderman, 2020c		X					X		
Smith, 2019		X	X	X			X		X
Söllner et al., 2013					X				
Sousa et al., 2014					X				
Stouten, 2019	X		X	X					
Straus, 2020	X								
Sutrop, 2019	X		X				X		
Thiebes et al., 2020							X		
Tripp et al., 2011					X				

Uga, 2019				X		X	X		X
Umbrello & De Bellis, 2018									
van de Poel, 2020	X		X						
van Wynsberghe & Robbins, 2013	X								
Vermaas et al., 2010					X				
Wickramasinghe et al., 2020	X		X		X		X		
Xu, 2019	X	X				X			
<b>FREQUENCY IN LITERATURE REVIEW</b>	<b>32</b>	<b>12</b>	<b>23</b>	<b>8</b>	<b>14</b>	<b>5</b>	<b>15</b>	<b>7</b>	<b>15</b>

### 2.1.5. STAGE 5: COLLATING, SUMMARIZING AND REPORTING THE RESULTS

Considering the transformative force and profound impact across various societal domains, AI has sparked ample debate about the principles and values that should guide the development and use of advanced technologies (Jobin et al., 2019, p.2). In the literature, there is a strong body of research on the conceptual underpinnings of ethics and ethical guidelines for AI systems. Further, there are several studies and attempts to help practitioners translating theoretical guidelines into practical and actionable instructions. Also, the concepts of trust and trustworthiness in AI systems have been found to be discussed and considered in the sources of the literature review. The details of the identified key concepts from the literature review are discussed in the following section, providing an overview of the existing advancements in the field of ethical AI, Human-centered AI, and how to induce trust into AI systems by design.

## 2.2. THEORETICAL CONTEXTUALIZATION

### 2.2.1. ARTIFICIAL INTELLIGENCE

To approach the concept of trustworthy artificial intelligence or Trustworthy AI, it is important to situate it in the bigger picture of the development of artificial intelligence systems. According to Margaret Boden, a leading figure in the field of AI research, AI seeks to make computers do the sorts of things that minds do (Boden, 2016, p.1). These include

capabilities like perception, association, prediction, planning, motor control and reasoning. The High-Level Expert Group on Artificial Intelligence (AI-HLEG, 2019) defines AI systems as “software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behavior by analyzing how the environment is affected by their previous actions”. In general, AI can be distinguished between “narrow AI”, which is dedicated to specific, autonomous and potentially repetitive tasks, and “artificial general intelligence” (AGI), which possesses human-like capabilities or beyond when presented with complex or unfamiliar tasks (Sutrop, 2019, p.502). But since AGI constitutes its own, diverse and broad research field, this work will focus on the realm of narrow AI, which is focused on particular tasks and product value propositions. Nowadays, AI can be found in a vast number of applications, from home appliances over music software to driverless cars and autonomous weapons. But whereas the first two approaches can probably be seen as benevolent and neutral, the latter two applications are already hinting towards potential tensions in their nature. Inevitably, questions about responsibility and morality emerge. Those questions about ethics and values need to be addressed rather sooner than later as advancements are unstoppable.

### **2.2.2. ETHICAL AI**

The rapid advancement of AI technology is recognizable in various fields, thus also in the appraisal of the importance of ethical considerations in the development and use of AI. In 2014, Bostrom (p.17) concluded that AI offers only a few new ethical issues, which are not already present in the design of other technologies. Only the outlook to more humanlike AI algorithms would allude towards potential challenges and complications. No more than a few years later, this assessment seems outdated and almost ignorant. The literature review revealed a strong body of research on ethical AI and indicates a growing awareness and acknowledgment of the need for ethical and moral thoughts into the design and development of AI technologies. The concept of AI ethics has emerged as a response to the abundance of societal and individual perils caused by the abuse, misuse, poor design, or unintended consequences of AI systems. In their 2019 report “Understanding artificial intelligence ethics

and safety”, the Alan Turing Institute has provided an overview of the most common potential harms caused by AI systems (Leslie, 2019, pp.4-5). Mark Coekelbergh (2020, p.179) warns that if the project of ethical AI fails, we risk ethical, social and economic disaster with unpredictable human, non-human and environmental costs. But in contrast to the sensational image of a dystopian future, in which a superhuman intelligence has enslaved humanity or robots having obliterated mankind, AI already has a huge positive impact on the life of many of us. AI systems are used in healthcare, public safety and transportation to support humans in their work, ensure security and reliefs in a myriad of contexts (Stone et al., 2016). However, it is important that future systems must be introduced in ways that build trust and understanding and further respect human and civil rights (Dignum, 2018). The need for ethical considerations in the development of intelligent interactive systems is becoming one of the main influential areas of research. Coekelbergh (2020, p.8) describes that the recent spectacular breakthroughs in AI, have created a sense of urgency on the part of ethicists and policymakers. Social roles may be filled by AI algorithms, implying new design requirements like transparency and predictability. Dignum (2018, p.2) characterizes the relation of AI and ethics on three levels:

- **Ethics by Design:** the technical integration of ethical reasoning as part of the behavior.
- **Ethics in Design:** the integration of regulatory measures that support the analysis of ethical implications.
- **Ethics for Design:** encompassing the codes of conduct, standards and certification processes to ensure development integrity.

Gamberlin (2020) describes a new role that has been appearing more and more often in the latest discussions: the role of the AI ethicist. She offers a preliminary description of the role and what it means to be an AI ethicist. Bietti (2020) argues that the concepts of ethics and morality in relation to technology are more and more at risk of being instrumentalized, either by the industry in the form of “ethics washing,” or by scholars and policy-makers in the form of “ethics bashing”. She states that “[...]the more ethics is used in tech circles as a performative façade, the more it is instrumentalized and voided of its intrinsic value” (p.218). Consequently, in their paper “Ethicist as Designer” (2013), van Wynsberghe & Robbins state that ethical considerations need to be integrated at an earlier stage into the design process before a product or service is getting developed or even introduced. In their perspective, ethics ought to be pragmatic and the ethicist should be considered a designer in the process of

technology development, who subscribes to a pragmatic view of ethics in order to bring ethics into the research and design of artifacts. Another issue that needs to be recognized is described by Hagerty & Rubinov, 2019, p. 2, who identified that the current analyses of AI in a global context are biased towards western perspectives and that there is a lack of research, especially outside the U.S. and Western Europe. But to approach AI ethics seriously, they need to be addressed with a cross-cultural understanding (Hagerty & Rubinov, 2019, p.19).

### **2.2.3. HUMAN-CENTERED AI**

Human-centered AI provides a new perspective on the design and development of AI systems. Many writings on AI ethics are emphasizing human values and human-centricity (Auernhammer, 2020; Ferrario et al., 2019; Guszczka et al., 2020). The Ethics Guidelines of the High-Level Expert Group of the European Commission state that a human-centric approach “in which the human being enjoys a unique and inalienable moral status of primacy in the civil, political, economic and social fields” (AI-HLEG, 2019, p.10). It is the evolution of bringing human-centered design into the field of AI as it aims to bridge the gap between ethics and practical application by providing specific recommendations to create products and services that augment, amplify, empower, and enhance humans. Human-centered AI research strategies emphasize that the next frontier of AI is not just technological but also humanistic and ethical. According to Ben Shneiderman (2020c), the concept of human-centered AI reverses the current emphasis on algorithms and AI methods, by putting humans at the center of systems design thinking. It emphasizes the user experience and measuring human performance, aiming to empower people, rather than to emulate them. This mental shift could result in a safer, more understandable, and more manageable future, mitigating existing fears of AI’s existential threats and raise people’s belief that they will be able to use technology for their daily needs and creative explorations (Shneiderman, 2020c, p.117). And there are further signs that such a shift is emerging and evolving. Besides the European Commission and its High-Level Expert group, there are initiatives from MIT, UC Berkeley or Stanford University, which established “Human-centered AI” institutes (Xu, 2019, p.42). According to Auernhammer (2020), it is this type of commitment that highlights the need and the potential of designing human-centered and ethical AI systems to play a pivotal role in the development and use of AI technology for the well-being of people.

#### **2.2.4. ETHICAL AI GUIDELINES**

One of the measures taken to address the nascent concerns about malicious, uncontrollable or hostile technology has been the development of ethical guidelines on the development of artificial intelligence. Various organizations, private companies as well as research institutions, have produced guidelines for the ethical development and use of AI systems. These guidelines comprise normative principles and recommendations aimed to harness the “disruptive” potentials of new AI technologies (Hagendorff, 2020). The comprehensive analysis of a corpus of 84 existing AI ethics guidelines, conducted by Jobin, Ienca & Vayena reveals a convergence of five prevalent ethical principles, namely transparency, justice and fairness, non-maleficence, responsibility and privacy, which appear in more than half of the analyzed guidelines. However, the analysis also shows a substantive divergence in relation to how these principles are interpreted, why they are deemed important, what issue, domain or actors they pertain to and how they should be implemented (Jobin, Ienca & Vayena, 2019, p.7). Fjeld et al. (2020) have identified eight themes, 47 principles within a body of 36 sources. But their value for practical implementation is often half-baked. The term “ethics washing” is used in this context, meaning polishing your public image on false grounds, says Anna-Mari Rusanen, one of the driving forces behind the “Ethics of AI” online course, developed and provided by the University of Helsinki (helsinki.fi, 2020). AI ethics is failing in many cases as it lacks reinforcement mechanisms as well as practical recommendations for actions (Hagendorff, 2020). And in cases where ethics is integrated into institutions, it currently mainly serves as a marketing strategy. Empirical experiments even show that the mere reading of ethics guidelines has currently no significant influence on the decision-making of software developers. To make AI policies work, it would be imperative to build a bridge between abstract high-level ethical principles defined by research institutes, companies or nation-states and the practices of technology development and use in particular contexts (Coeckelbergh, 2020, p. 170).

#### **2.2.5. APPLYING GUIDELINES IN PRACTICE**

As the AI guideline analyses by Fjeld et al. (2020), Hagendorff (2020) and Jobin, Ienca & Vayena (2019) illustrate, nearly all of the formulated guidelines consider similar values to be crucial requirements for the development of “ethically sound” AI applications. Yet, how the development of ethical AI technology should work and how to implement these precepts is

currently still uncertain and fuzzy. The lack of specific and verifiable principles endangers the effectiveness and enforceability of ethics guidelines. The AI Ethics Impact Group (2020), a joined research initiative, has developed a framework that offers concrete guidance to decision-makers in organizations developing and using AI. The framework offers directives on how to incorporate values into algorithmic systems, and how to measure the fulfillment of values using criteria, observables and indicators combined with a context-dependent risk assessment. It introduces different tools and methods to operationalize abstract principles and to classify application contexts, e.g. through a risk matrix. Via the risk matrix tool, the ethical sensitivity in relation to an application context can be determined (AI Ethics Impact Group, 2020, p.35). The risk matrix tool developed by Krafft & Zweig (2019) is a schematic visualization for identifying categories - classes - of an algorithmic decision-making system according to its risk potential. In his paper “Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-centered AI Systems” (2020a), Ben Shneiderman proposes a set of recommendations, divided by levels of governance – team, organization and industry, which aim to increase the reliability, safety and trustworthiness of Human-centered AI systems. Floridi et al. (2018) have defined an ethical framework for a good AI society, presenting a synthesis of five ethical principles and a list of 20 concrete recommendations to assess, develop, incentivize and support the concept of “good” AI. The authors state that in order to create a “Good AI Society”, the ethical principles they described need to be embedded in the default practices of AI.

#### **2.2.6. TRUST IN TECHNOLOGY**

The focus of this work will be on the concept of “trust in AI” or “Trustworthy AI”. Trust is commonly defined as an individual’s willingness to depend on another party because of the characteristics of the other party (Rousseau et al. 1998). Mayer, Davis, and Schoorman (1995, p. 712), argued that trust is “the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party”. Trust in technology has been discussed extensively in the past as it is believed that by focusing on trust in technology, the determination of what makes a technology trustworthy, irrespective of the people and human structures that surround the technology, is more achievable (McKnight et al., 2011, p.2). Further, Söllner et al. (2013) have found that trust helps to reduce risk, uncertainty and anxiety associated with technological interaction. And Lankton et al. (2015)

identified trust as being crucial in supporting user adoption and maintaining a gradual and steady relationship with the system. The initial trust formation is particularly relevant, in the technology context, as users must overcome perceptions of risk and uncertainty before they are willing to use and adapt to novel artifacts. Li, Hess & Valacich (2008) have found that cognitive and calculative bases of trust have a significant impact on the initial trust formation. Tripp, McKnight & Lankton (2011) assert that the predictive ability of different trust measures is significantly impacted by the technological context in which they are applied. Further, Andras et al. (2018) distinguish three different levels of trust, namely inductive trust, social trust and moral trust to describe a trust relationship. Inductive trust is derived from personal past experience and is therefore based on estimation and expected outcome. The authors state that to obtain trust in a human-machine relationship that is not already established, hence can't be inductive, a system has to rely on the initial trust formation. To support the initial trust formation, which is also based on the levels of social and moral trust, an AI system needs to be transparent. With the concept of "trustworthy AI", an attempt in recent research is made, to shed light on the perceived "black box" of artificial intelligence decisions and to enable the emergence of trust.

### **2.2.7. TRUST IN AI**

In their comprehensive literature review "Human Trust In Artificial Intelligence: Review Of Empirical Research", Glikson & Wooley (2020) state that success of integrating AI into the organizational context critically depends on workers' trust in AI technology. They define Trust as particularly relevant to the human-AI relationships because of the perceived risk embedded in them, due to the complexity and nondeterminism of AI behaviours. The European Commission's High-level Expert Group on AI (AI-HLEG) has adopted the position that we should establish a relationship of trust with AI and should cultivate trustworthy AI (2019, p. 35). According to Xu (2019), especially the so-called "black-box phenomenon", common for deep learning technologies, can result in users questioning decisions from the system. Such reflexive skepticism can affect users' trust and decision-making efficiency, which in turn will also affect the adoption of AI solutions. Ferrario et al. (2019) also see the current advancements in the field of Trust in AI critical, as they are worrying that statements like "Trust is a prerequisite for people and societies to develop, deploy and use AI" (AI-HLEG, 2019) are not providing a clear setting for the above discussion. Further, they argue that the overall level of awareness in society on topics like AI is still quite low.

According to their analysis, most users of AI-powered products and services are not aware of the presence of AIs. In their paper “In AI We Trust Incrementally: a Multi-layer Model of Trust to Analyze Human-Artificial Intelligence Interactions”(2019), Ferrario et al. propose an incremental model of trust that can be applied to both human-human and human-AI interactions. Ryan (2020) even argues that AI doesn’t have the properties to be trusted since it doesn’t possess any emotive states or can be held responsible for its actions.

#### **2.2.8. TRUSTWORTHY AI**

Trustworthy AI (TAI) is grounded on the concept that trust is a fundamental prerequisite to create societies, economies, and sustainable development. Thiebes, et al. (2020) derive that individuals, organizations, and societies will only be able to realize the full potential of AI, if trust can be established in its development, deployment, and use. Francesca Rossi, a member of the AI-HLEG concludes in her work, that to fully gauge its potential benefits, a system of trust needs to be established, both in the technology itself as well as in those who produce it. The development of several high-level principles has laid the foundation to guide AI towards a positive impact. The necessary next step is to put such principles to work and create robust implementation mechanisms (Rossi, 2018, p.132). In their work “Trustworthy Artificial Intelligence”, Thiebes et al. (2020) propose five principles for the development of TAI, namely beneficence, non-maleficence, autonomy, justice, and explicability. By analyzing how the involved parties interact with each other during the development and co-creation process of AI, they identify tensions between the current state of AI development, deployment, and use and the five proposed TAI principles. Shneiderman (2020c) proposes a two-dimensional framework of Human-Centered Artificial Intelligence (HCAI) to seek high levels of human control AND high levels of automation. The framework separates the levels of automation and autonomy from the levels of human control. The author claims that by applying the new guideline, it is more likely to produce computer applications that are reliable, safe and trustworthy. Mark Ryan on the other hand states that the concept of trustworthy AI is not accurate. He suggests either changing the term “trustworthy AI” to “reliable AI” or removing it altogether, as trust can only occur between trusted parties, whereas he describes AI as a systematic group of techniques (2020, p.2765).

### 2.2.9. EXPLAINABLE AI

The idea of explainable AI (XAI) is to address questions such as "How does it work?" and "What mistakes can it make?" and "Why did it just do that?". Explanations, in the form of post-hoc interpretability, can help establish rapport, confidence, and understanding between the AI agent and the user. This is especially relevant when it comes to understanding failures and unexpected AI behavior (Ehsan & Riedl, 2019). The core idea of explainability is meaning-making, which relies on the user's ability to make sense of the inner workings of a system. This process is a relational process where the alignment of situated epistemologies of the user and the machine needs to take place (Ehsan & Riedl, 2019, p.2). Doran et al. (2017) assert that to achieve trustworthiness and an evaluation of the ethical and moral standards of a machine, explanations should provide insight into the rationale of the AI system and enable users to draw conclusions based on them. They developed three notions to classify systems and their levels of comprehensibility.

- **Opaque Systems:** the inner mechanisms like input and output are invisible to the user.
- **Interpretable Systems:** users can see, study and understand the input and output mappings.
- **Comprehensible Systems:** emit symbols of explanations, e.g. words or visualizations, which allow users to relate properties to the input and outputs. However, users are responsible for compiling and comprehending these symbols.

Although interpretable and comprehensible systems already provide some means to explain and understand the output of systems, both approaches lack the ability to formulate human-understandable reasoning that explains the decision-making process of a model. Hoffmann et al. (2018) discuss the challenges with measuring explainable AI, focussing on different methods of evaluation, the quality of explanations in terms of satisfaction and understandability as well as how curiosity motivates the search for explanations.

### 2.3. EXISTING TOOLS & METHODS

The second goal of the literature review was directly addressing the research sub-question 1: SRQ1: "*What current design practices exist to support the translation of existing trustworthy AI guidelines into practice?*". The motivation for collecting already existing tools and methods was to evaluate their suitability and adaptability for the specific use-case of

designing for trustworthy AI and to provide hands-on assistance. However, the scope of the collection was deliberately kept broad, to capture a variety of tools and methods. Despite focusing only on those which already address the specific topic of “design for trustworthy AI”, the search also considered adjacent and related fields. For instance, did the search also consider tools that are addressing the topic of “Design for Trust”, or “Ethical Design”, without a specific focus on AI, as there is a potential of adapting the method for a specific use-case? Further, the collection also considered data ethics tools, as they could also be relevant in addressing some of the aspects when designing for trustworthy AI.

### **Example - Design for Trust toolkit**

The “Design for Trust” toolkit created by the “SRI International” non-profit research institute (2020) provides a comprehensive set of principles and methods to help designers and technologists to include trust considerations in their processes. It contains activities as well as worksheets addressing the value of trust from a design perspective, which hold potential to also be applicable in the design for trustworthy AI systems.

### **Example - Data Ethics Canvas**

According to their creators, the “Data Ethics Canvas” is “a tool for anyone who collects, shares or uses data. It helps identify and manage ethical issues – at the start of a project that uses data, and throughout. It encourages you to ask important questions about projects that use data, and reflect on the responses” (The Open Data Institute, 2019). And although data collection and processing is often not directly considered to be a “design matter”, asking important and relevant questions is. Therefore the “Data Ethics Canvas” tool was added to the collection.

This broader range is also reflected in the search strategy consisting of the defined search terms, inclusions and exclusions. During the literature review a collection of 34 existing tools, techniques and frameworks have been identified to create a collection of available practices (Appendix 1). The collection would already provide value on its own as a repository of tools, methods and techniques available for design professionals to use. However, as the goal this work is to create a comprehensive and actionable toolkit, reflected in research sub-question 2: SRQ2: *“How to create a toolkit that will enable designers to design for the formation of trustworthy AI systems?”* as well as research sub-question 3: SRQ3: *“How does the toolkit enable the design of trustworthy AI systems in the dimensions of “usefulness”, “satisfaction”*

and “*efficiency?*”, the collection rather contributes to the possibility space and provides a practical foundation for the envisioned toolkit.

## **2.4. REFLECTION**

From the literature review, it became apparent that the current understanding of AI ethics and its related sub-fields continues to reach deeper and deeper into our lives—often in unexpected ways that challenge the very foundation of our collective notions of society (MAIEI, 2021). Academia, private companies and Nation-states discuss how algorithms now inform decisions ranging from the seemingly inconsequential to those that have a profound, direct effect on our lives, and how they ought to be designed and developed. The literature review unveiled a myriad of research on the field of ethical AI and guidelines about the development and use of AI. Extensive reviews on the body of these guidelines have been conducted by Jobin et al. (2019), Hagendorff (2020), Fjeld et al. (2020) and Glikson & Woolley (2020). However, if you take a closer look and seek explicit help and guidance on how to implement such guidelines, the numbers are dwindling – even more so, if the particular lens of designing for trustworthy AI is applied. Although trust as a concept is acknowledged as a fundamental principle within the field of AI development (Ryan & Stahl, 2020, p.14), there still remains a gap when it comes to tools and methods to help designers working on AI systems. This lack of tooling aggravates the above-discussed underutilization of design as a profession in the development of AI systems, products and services and underpins the relevance and need for helping professionals designing for trustworthy AI.

### **3. SECOND PILLAR: DESIGN OF THE TOOLKIT PROTOTYPE**

---

The overarching objective of research sub-question 2: SRQ2: *“How to create a toolkit that will enable designers to design for the formation of trustworthy AI systems?”*, is to ensure that designers are able to embrace the challenges of modern technology development. To achieve this, they need to be equipped with the right tools and methods. As identified above, the current problem isn't a lack of guidelines around ethical and humane AI development, but rather the means to bridge the chasm between these abstract high-level ethical principles and the practical application of them (Coeckelbergh, 2020, p. 170). Further, to examine the emerging interaction, behavior, and needs of the users of AI technology, which in turn allows assessing the impact on people, experimentation is needed. However, experiments require designers and design researchers to decide which ethics perspective need to be included and how each perspective is assessed and integrated into the process (Auernhammer, 2020, p.4). Therefore it is imperative that frameworks, methodologies and tools are developed to provide practitioners the necessary means to approach.

One of the fundamental values for the development of artificial intelligence, is the value of trust. Trust is an essential principle for interpersonal interactions and it constitutes a prerequisite for a society to work. It is therefore indisputable that trust needs to be acknowledged as a key requirement for the ethical deployment and use of AI (Ryan & Stahl, 2020). The High-Level Expert Group on Artificial Intelligence even uses trustworthiness as the overarching paradigm for their ethical AI guidelines (AI-HLEG, 2019). However, there aren't many details, specifics or instructions available on how the design and development of “trustworthy AI” should be approached from a design perspective. Sutrop (2019) states that “if it is important that people trust AI systems, it is not enough to establish and articulate the purpose of achieving trustworthy AI. It is imperative that we also think about how to build trust in AI”. To fill this gap embodies the underlying research objective of this work.

#### **3.1 RESEARCH PROBLEM AND STRATEGY**

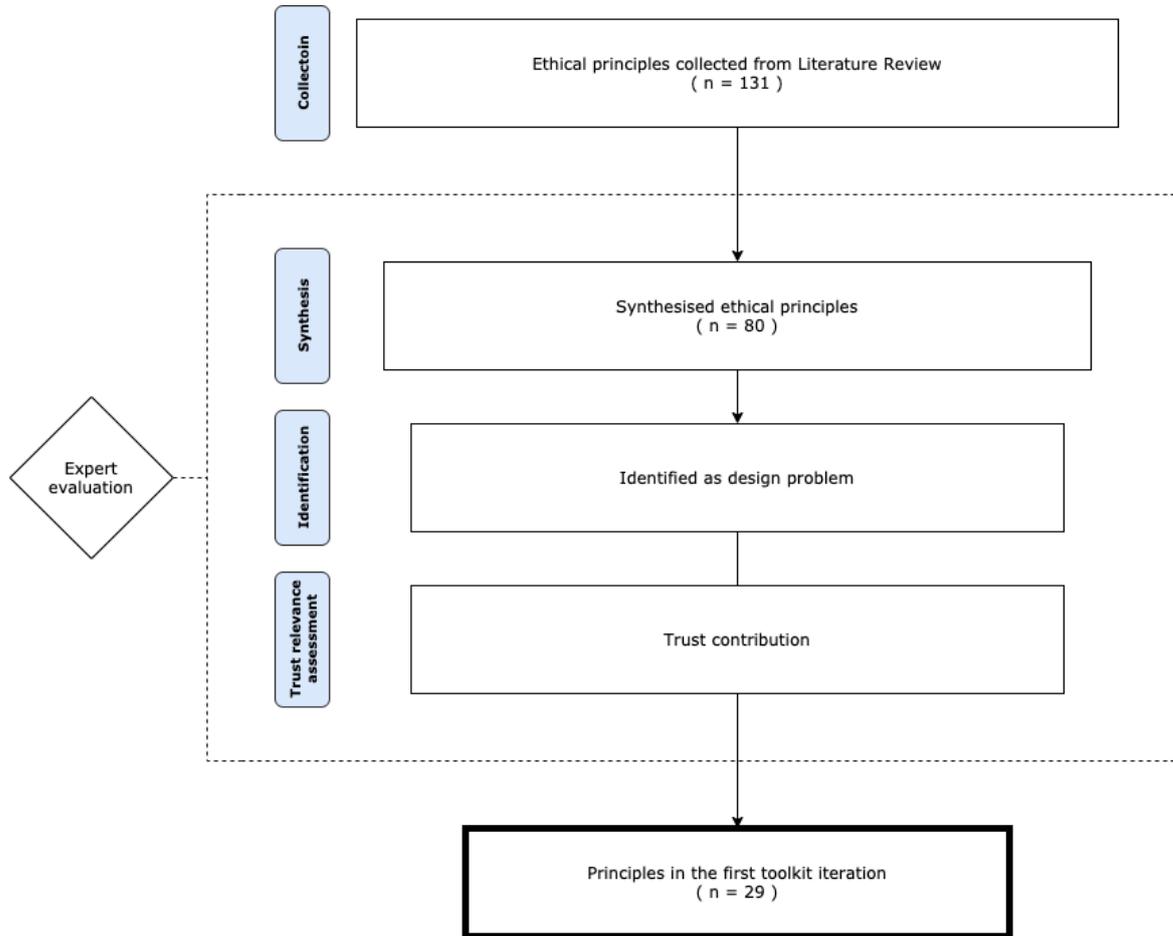
As the objective of this research is to advance the current state of knowledge and tooling for designers working with AI systems, a research through design constitutes its conceptual foundation. Applying a research through design methodology, explorations are grounded in real knowledge produced by the design researchers. Through an active process of ideating,

iterating, and critiquing potential solutions, design researchers continually reframe the problem as they attempt to make the right thing. According to Zimmermann et al. (2007, p.497), the target output is a concrete problem framing and artifacts like prototypes and design process documentations. The research process is composed of two main iterative parts. In the first iteration, the most common ethical principles from a wide body of ethical AI guidelines are identified based on the literature review and synthesized into a comprehensive list of agreed-upon principles. Furthermore, the list of principles is assessed for the first time by the researcher in collaboration with an expert in the field of trust in technology, whether they qualify to be understood and addressed as a design problem for the value of trust. In the second phase, the qualified principles are evaluated by design professionals in participatory design workshops. Within the workshops, the design professionals are tasked to assess if a principle relates to one or more phases in a design process. To ensure comparability of the results of the workshops, a formalized design process model was predetermined for the participants. Furthermore, the principles are assessed whether they contribute to one of three trust-inducing qualities, namely “Benevolence”, “Competence” and “Risk perception”. For each validated principle, useful and appropriate methods are suggested. Additionally, the design professionals are asked to highlight if a principle is already addressed in other user interface or human-computer interaction guidelines, to avoid redundancy. And lastly, the participants are also encouraged to share any tools and methods they think could be useful to address a specific principle.

### **3.2. TOOLKIT ITERATION 1: PROCESS**

The goal of the toolkit is meant to help designers consider the relevant theoretical AI guidelines and principles related to the value of trust during their design process. The toolkit will act as supporting guidance to decide how the nature of an AI system, its context and implementation will affect the emergence of trustworthiness. The idea of the toolkit is to contain a set of principles related to the formation of trustworthiness, and suggested methods and instruments for different phases of the design process. To determine the relevant principles for the emergence of trust, a multi-phase framework was developed.

**Figure 2:** Framework to identify relevant principles for the design of trustworthy AI



### 3.2.1. PARTICIPANTS

The first iteration of the toolkit was shaped by the author of the thesis in collaboration with a subject matter expert on the topic of “trust in technology”. The author is a professional designer and researcher with over a decade of experience. Its tenure includes multiple years as an in-house designer in software companies, design agencies, working as a freelance designer as well as participating in academic design research projects. The subject matter expert is a trust researcher and Associate Professor of Interaction Design at Tallinn University’s School of Digital Technology, who has published plenty of articles and papers on the topic of trust in technology. Further, she also developed the Human-Computer Trust Model (HCTM) together with her colleagues and iterated and validated the model in several studies. Involving a subject matter expert added invaluable knowledge and relevant perspectives for the highly complex topic of trust and trust assessment.

### **3.3. IDENTIFYING DESIGN PRINCIPLES FOR TRUSTWORTHINESS**

Francesca Rossi states in her article “Building trust in artificial intelligence” (2018), that to harness the benefits entailed by AI, it is inevitable to create a system of trust, which considers the technology side as well as those who produce it. The notorious issues of biases, the lack of transparency and explainability, malicious data handling and questionable data policies are, at least partially, approachable from a design perspective. But to understand which of the principles qualify as “design problems”, a general overview of the existing ethical AI guidelines is a precondition.

#### **3.3.1. COLLECTION OF COMMON ETHICAL GUIDELINES IN THE LITERATURE**

As a result of the conducted literature review it became apparent that there exists an abundance of guidelines on the ethical and moral design, development and use of AI systems. But a comparative study, let alone deeper analysis of all currently AI ethics guidelines would go beyond the scope of this study. Therefore, this thesis draws upon existing, reviewed studies of such nature. In 2019, Jobin et al. undertook the ambitious endeavour to create a comprehensive overview of the available AI ethics guidelines at the time. Altogether, 84 documents had been reviewed, their content described and convergence as well as common themes across these guidelines had been identified. In 2020, Ryan and Stahl built upon the robust categorisation of ethical principles from Jobin et al. (2019), analysing 91 sources in total, including the 84 guidelines from the Jobin et al. (2019) study, plus 7 additional sources. Their goal was to move beyond the high-level ethical principles that are common across the AI ethics guidance literature and provide a description of the content that is covered by these principles (Ryan & Stahl, 2020). They created a categorisation system of eleven principles, with 61 sub-principles. In the same year, the Berkman Klein Center for Internet & Society at Harvard University, led by Jessica Fjeld et al. (2020), published a report called “Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI”. The report states that guidelines for ethical, rights-respecting, and socially beneficial AI develop in tandem with the underlying technology. Therefore an urgent need to understand these guidelines is obligatory. The report analyzed thirty-six prominent AI guideline documents, discovering eight themes which suggest an emergence of sectoral norms, and outlines 47 principles elicited from these documents. Another relevant study had been conducted by Morley et al. (2020), building upon the work by Hagendorff (2020) and

Floridi & Cowls (2019), connecting five high-level ethical principles and 23 identified tangible system requirements. In summary, the starting point for the analysis of relevant ethical guidelines for the design of trustworthy AU was composed of a corpus of 131 guidelines.

### 3.3.2. SYNTHESIZING ETHICAL PRINCIPLES

As the body of ethical guidelines was a composition of multiple sources, the collection needed to be synthesized, deduplicated and cleaned up to achieve a coherent list of principles. For the synthesis, the principles had been compared by name, description and semantic meaning, to mitigate misunderstanding and redundancy.

#### **Example - Explainability**

All three of the used sources include a principle called “Explainability”, which provided a straightforward suggestion on how the principle should be named. Further, the provided descriptions have been analyzed and compared, to ensure that there is alignment in meaning between the sources.

- *“AI should be explainable to external algorithmic auditing bodies to ensure the technical and ethical functionality of their AI.”* (Cerna Collectif, 2018, in Ryan & Stahl, 2020, p.6)
- *“The translation of technical concepts and decision outputs into intelligible, comprehensible formats suitable for evaluation.”* (Fjeld et al., 2020, p.42)
- *“The ability to explain both the technical processes of an AI system and the related human decisions.”* (Morley et al, 2020, p.2151)

#### **Example - Showing**

For the principle of “Showing”, the name comparison did not yield any results, as the term “Showing” was only used in the Ryan & Stahl study. However, the analysis of the descriptions revealed close similarity between the principle of “Showing” and the principle of “Notification when interacting with an AI” from the Fjeld et al. (2020) report.

- *“It should also be clear to the end user that they are interacting with an AI system, rather than a human.”* (Ryan & Stahl, 2020, p.7)
- *“Where an AI has been employed, the person to whom it was subject should know.”* (Fjeld et al., 2020, p.45)

The same process has been applied across all 131 principles from the collection, to standardize naming, content and meaning. However it's relevant to mention, that the descriptions and definitions provided in the different studies aren't written by the authors of the comparative studies, but they are the descriptions from the original sources. This means that multiple descriptions existed for each of the principles. Although the selection of the most suitable description for a principle has been conducted with rigor and diligence, there hasn't been any extended analysis on the coherence of the gathered definitions within the scope of this thesis. The synthesis has been conducted based on the assumption of scientific accuracy from the authors of the comparative studies. After the synthesis process of the three sources, the list of ethical AI principles was reduced to a total number of 80 principles.

### **3.3.3. IDENTIFICATION AS DESIGN PROBLEMS**

To focus the toolkit specifically on the concept of designing for trustworthy AI, the list of 80 general ethical AI principles needed to be scrutinized in regards to which of the principles would qualify to be understood as general design principles contributing to the emergence of trust. However, this endeavour wasn't straight-forward, as aptly described by Blackler et al. in their latest paper analysing 20 years of discussion on how to define "design". In the article it is stated that "ever since the industrial revolution, a solid, common understanding of what design is and does has proven nearly impossible to establish" (2021, p.42). As there was no commonly agreed upon definition available, which could be applied to decide which of the ethical AI principles would qualify as design principles for trust, multiple non-structured filters have been deployed. The identifications were the result of extended conversations, evaluations and collaborations between the author of this thesis, a professional designer and a subject matter expert on the topic of trust in technology. During the process of elimination in relation to other areas and sub-fields, each of the principles was reviewed whether it would be approachable from a different perspective than design, and how more or less suitable this perspective would fit the principle in comparison to the design discipline. Further, there was an initial assessment of the relevance for the value of trust. In this step, principles were assessed whether could would be understood to contribute to the emergence of trust when designing an AI system. If a principle was considered to be not, or only marginally influenceable by the broader understanding of the responsibility of a designer, or was found to not relate to the trust value, a principle had been discarded from the list.

### **Example - Open source data and algorithms**

According to the report from the Berkman Klein Center (Fjeld et al, 2020, p.43), the principle of “Open source data and algorithms” is a common and familiar concept in technology governance. It helps to avoid monopolies in regards to data, platforms or other dimensions and encourages the sharing of the benefits of AI development to the greatest extent. And although the idea of using open source systems and algorithms or publishing a developed system as an open source system can be influenced by design decisions, it is rather understood as a mindset and required acceptance and willingness from business stakeholders, as well as engineers and data scientists. Therefore the principle of “open source data and algorithms” didn’t meet the requirements for being considered a design principle and had been eliminated from the list.

### **Example - Disclosure**

In the study from Ryan & Stahl (2020) the principle of “Disclosure” is described as “AI should go through internal and external auditing to ensure they are fit for purpose, but the organisation also needs to be able to explain and justify the use of their AI. Organisations should allow for independent analysis and review of their systems” (Amnesty International/Access Now, 2018) in Ryan & Stahl, 2020, p.7). Based on the experience from the author of this thesis, decisions about the disclosure of technology and processes are subjected to matters of intellectual property rights, licensing agreements and other binding obligations which usually can’t be affected by designers.

### **Example - Resources (energy)**

One of the principles during the literature review was “Resources (energy)”, which was described as “AI should be created in a way that ensures effective energy and resource consumption, promotes resource efficiency, the use of renewable materials, and reduction of use of scarce materials and minimal waste (European Parliament, 2017, in Ryan & Stahl, 2020, p.15). Even though resource efficiency could be arguably understood as addressable from a design perspective, the trust contribution could not be determined in the assessment discussions. Therefore the principle was removed from the initial list of principles.

### 3.3.4 TOOLKIT PROTOTYPE

These evaluation filters have been applied for all of the remaining 80 principles to further reduce and specify the list of principles into a concise and relevant list of items that can be understood as design principles. As with every subjective, less formalized process, the decisions haven't been clear and precise for each of the principles. There have been vague and ambiguous cases, where a distinct definition and decision weren't straightforward. For these cases, the heuristic of keeping the list of principles as lean as possible has been defined and applied. This means that in cases where a decision couldn't be made unambiguously, the principle was also eliminated from the list. By this multi-layer elimination process, the list had been further reduced to 29 principles, which constitutes the prototype of the toolkit to enable the design of trustworthy AI (Table 3).

**Table 3:** Toolkit prototype - including principles identified as design problems

PRINCIPLE	DESCRIPTION
Accessibility	AI should be accessible to those that are often socially disadvantaged (such as those with vision problems, dyslexia or mobility issues) (Ryan & Stahl, 2020, p.10).
Beneficence	AI should “compliment the human experience in a positive way (Unity Technologies, 2018 in Ryan & Stahl, 2020, p.13).
Communication / Autonomy	End users should be provided with accurate information to ensure that they are not manipulated, deceived, or coerced by AI. AI organisations should ensure that end users are informed, not deceived or manipulated by AI and should be allowed to exercise their autonomy. AI organisations need to ensure that the “principle of user autonomy must be central to the system’s functionality (Ryan & Stahl, 2020, p.14).
Consent	The use of personal data must be clearly articulated and agreed upon before its use (UNDG, 2017 in Ryan & Stahl, 2020, p.14).
Consideration of Long Term Effects	Deliberate attention to the likely impacts, particularly distant future impacts, of an AI technology during the design and implementation process (Fjeld et al. 2020, p.58).
Dignity	AI should be developed and used in a way that respects, serves and protects humans physical and mental integrity, personal and cultural sense of identity, and satisfaction of their essential needs (HLEG-AI, 2019, p. 10, in Ryan & Stahl, 2020, p.15).
Diversity	Organisations implementing and using AI should encourage a diversity of opinions throughout every stage of its use (Smart Dubai, 2019 in Ryan & Stahl, 2020, p.9).
Explainability	The translation of technical concepts and decision outputs into intelligible, comprehensible formats suitable for evaluation. (Fjeld et al. 2020, p.42).
Fairness	There should be steps in place to ensure that data being used by AI is not unfair, or contains errors and inaccuracies, that will corrupt the response and decisions taken by the AI. To ensure the Artificial intelligence ethics guidelines fairness of AI, their design should be fit for purpose, identify impacts on different aspects of society (ICO, 2017 in Ryan & Stahl, 2020, p.7).
Human Agency	AI systems are designed and implemented with the capacity for people to intervene in their actions (Morley et al., 2020, p.2151).
Human Oversight	The “ability to opt out of automated decision” principle is defined as affording individuals the opportunity and choice not to be subject to AI systems where they are implemented (Fjeld et al. 2020, p.54).
Impact assessment	The objectives and expected impact of AI must be assessed and documented in the development stage (Algo.Rules, 2019 in Ryan & Stahl, 2020, p.7).

Inclusion	Attention should be given to under-represented and vulnerable groups and communities, such as those with disabilities, ethnic minorities, children and those in the developing world. Data that is being used should be representative of the target population and should be as inclusive as possible (HLEG-AI, 2019, p. 10, in Ryan & Stahl, 2020, p.8).
Justification	The purpose for building the system must be clear and linked to a clear benefit —system’s should not be built for the sake of it (Morley et al., 2020, p.2151).
Non-bias	Developers should examine unfair biases at every stage of the development process, including training data used, potential human biases and bias derived from the results of algorithmic processes and should eliminate those found (Ryan & Stahl, 2020, p.8).
Non-discrimination	AI should be designed for universal usage and not discriminate against people, or groups of people, based on gender, race, culture, religion, age or ethnicity (Cerna Collectif, 2018 in Ryan & Stahl, 2020, p.8).
Non-subversion	AI systems should be used to respect and improve the lives of citizens, rather than “subvert, the social and civic processes on which the health of society depends” (Future of Life Institute, 2018 in Ryan & Stahl, 2020, p.11).
Plurality	AI developers should consider the range of social and cultural viewpoints within society and should attempt to prevent societal homogenization of behaviour and practices (University of Montreal, 2017 in Ryan & Stahl, 2020, p.9).
Precaution	Those who develop AI must have the necessary skills to understand how they function and their potential impacts (Algo.Rules, 2019 in Ryan & Stahl, 2020, p.11).
Privacy	Users should have control and access to data stored about them (The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2019, in Ryan & Stahl, 2020, p.12).
Responsibility	The notion that individuals must be conscientious and thoughtful when engaged in the design of AI systems and the crucial role that researchers, engineers and developers play as “architects of our digital society (Fjeld et al. 2020, p.57).
Reversibility	It is important to clearly articulate if the outcomes of AI decisions are reversible. The ability to undo the last action or a sequence of actions allows users to undo undesired actions and get back to the ‘good’ stage of their work” (Personal Data Protection Commission Singapore, 2019, p. 16; Clark, 2019 in Ryan & Stahl, 2020, p.9).
Showing	It should be clear to the end user that they are interacting with an AI system, rather than a human. Further, where an AI has been employed, the person to whom it was subject should know (EPSRC, 2011, Ryan & Stahl, 2020, p.7).
Solidarity	It is important to consider if the AI supports rich and meaningful social interaction, both professionally and in private life, and not support segregation and division, within the design and development process (Ryan & Stahl, 2020, p.16).
Stakeholder participation	To develop systems that are trustworthy and support human flourishing, those who will be affected by the system should be consulted (Morley et al., 2020, p.2151).
Transparency	The principle of “transparency” is the assertion that AI systems should be designed and implemented in such a way that oversight of their operations are possible (Fjeld et al. 2020, p.42).
Trustworthiness	End users should be able to justly trust AI organisations to fulfil their promises and to ensure that their systems function as intended. Building trust should be encouraged by ensuring accountability, transparency and safety of AI (Deutsche Telekom, 2018; Institute for Business Ethics, 2018; Microsoft, 2018a, 2018b; Sony, 2018; NITI Aayog, 2018 in Ryan & Stahl, 2020, p.14).
Understandability	AI organisations should understand how their AI works and explain the technical functioning and decisions reached by those technologies, whenever possible (European Parliament, 2017 in Ryan & Stahl, 2020, p.6).

### 3.4. REFLECTION

During the second pillar of the research process, an initial list of 131 ethical principles elicited from the literature review had been compiled. The list was an amalgamation of three previous studies, which had thoroughly mapped and analyzed the current corpus of principles and guidelines on ethical AI, identifying convergences, common themes and normative

implications (Ryan & Stahl, 2020; Fjeld et al., 2020, Morley et al, 2020). The initial list went through a first review, detecting overlaps and duplications in the principles from the different sources. In the next iteration, the tidied-up list was then reviewed and assessed by the author of the thesis, a professional designer, together with a subject matter expert on the topic of trust in technology. The evaluation criteria were defined as “comprehensible as a design problem” and “contributes to the emergence of trust”. The evaluation happened in multiple expert reviews with the aim of eliminating principles which would not meet the requirements of the envisioned toolkit for the design of trustworthy AI. After the elimination process, the prototype of the toolkit consisted of 30 design principles and descriptions which had built the foundation for the third pillar of the research.

## **4. THIRD PILLAR: EVALUATION & ITERATION OF THE PROTOTYPE**

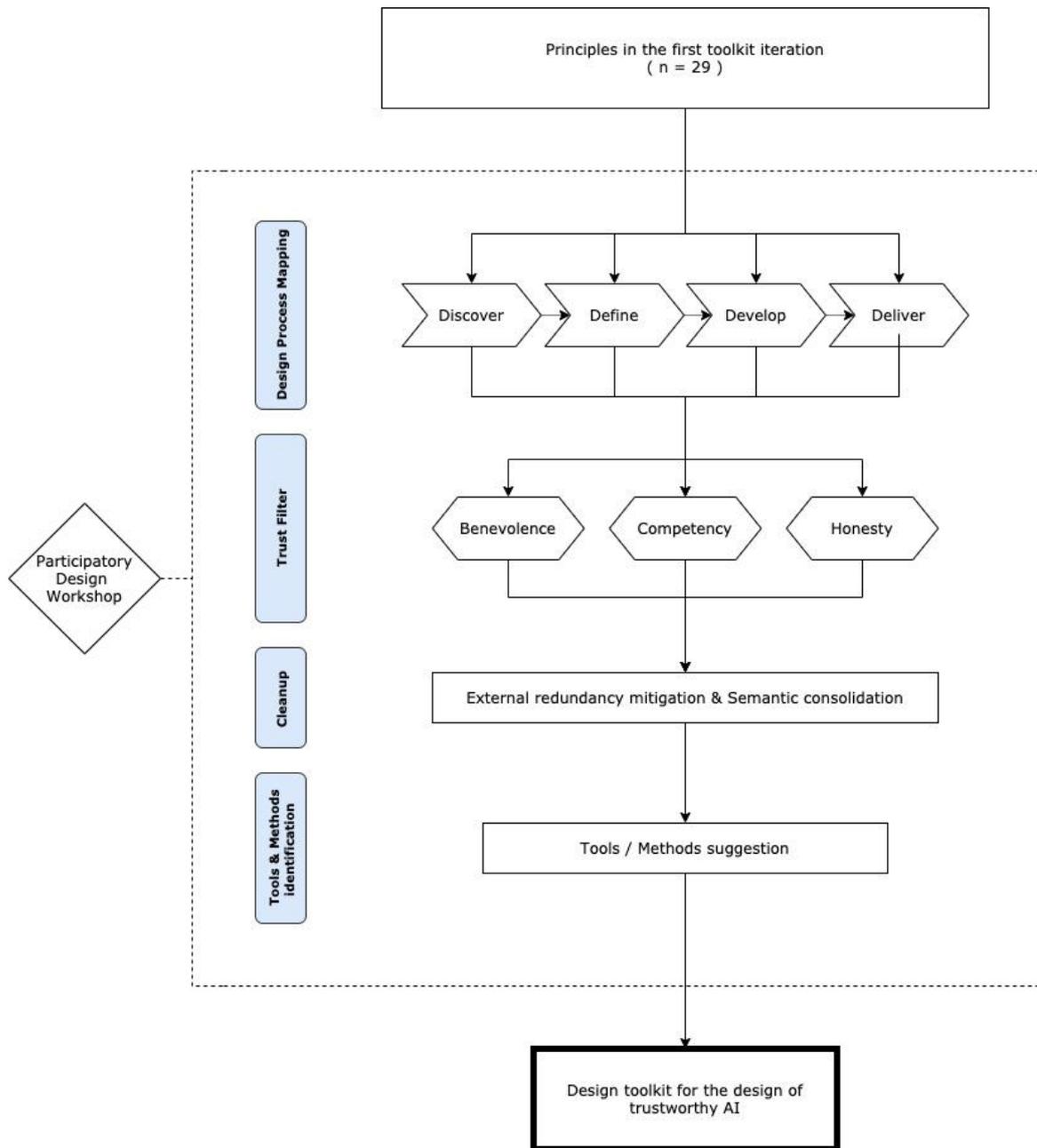
---

To further pursue the goal of answering research sub-question 2: *“How to create a toolkit that will enable designers to design for the formation of trustworthy AI systems?”*, the initial prototype of the toolkit was evaluated together with design professionals in several participatory design workshops. Design workshops are a variant of the participatory design concept which consolidates creative co-design methods into organized sessions for several participants to work with designers (Martin & Hanington, 2012, p.62). The objective of conducting participatory workshops together with experienced designers was to ensure continuous involvement of the envisioned “users” of the toolkit already during the toolkit development. The goal of the workshops was to assess the identified principles and to co-design the next iteration of the toolkit. According to Sanders & Stappers (2008, p.6), co-design is understood as a collective creativity which is applied across the whole span of the design process and refers to the collective creativity of collaborating designers.

### **4.1. TOOLKIT ITERATION 2: PROCESS**

Jan Auernhammer states, that “designers need to recognize their role, ideology, and socio-economical processes in which they are embedded for the design of AI systems beneficial for society. An HCD approach that can overcome the isolated viewpoint of the designer is Participatory Design” (Auernhammer, 2020, p.6). The backbone of participatory design workshops is manifested by the concept of activity-based research. And although designing and conducting the activities, as well as organizing and running the design workshops can be very labor intensive, they also can yield tremendously useful and insightful findings.

**Figure 3:** Framework to validate prototype of the trustworthy AI toolkit



Design workshops encompass methods with the unifying philosophy to allow face-to-face involvement of users via co-design engagements (Martin & Hanington, 2012, p. 62). Co-Design as a concept can be used in conjunction with many other tools, as they can be adapted for a co-creative setting (Stickdorn & Schneider, 2012, p.198). According to Sanders & Stappers (2008, p.12), Co-Design carries the notion that “the person who will eventually be served through the design process is given the position of ‘expert of his/her experience’, and plays a large role in knowledge development, idea generation and concept development”. This

is especially relevant in the context of developing a toolkit for designers as it allows the craftswoman and craftsman to shape their own tools.

#### **4.1.1. WORKSHOP SETUP**

The workshops were organized as one-on-one online sessions, using appointment scheduling, video conferencing and collaborative work applications included in the Google Classroom Suite, specifically Google Calendar ([calendar.google.com](https://calendar.google.com)), Google Meet ([meet.google.com](https://meet.google.com)) and Google Sheets ([sheets.google.com](https://sheets.google.com)). Under different circumstances, the workshops would have been conducted in-person and simultaneously with multiple participants attending at the same time. But the current situation of the global COVID-19 pandemic didn't allow for such a fruitful collaborative setup. Therefore the structure and process of the workshops had been adapted to better accommodate for local and social distancing as well as the general strain in the population. Prior to the workshops, there was a pilot study conducted to run through the planned session and test all the associated materials. The pilot study yielded feedback on the timeframe of the workshops, the clarity of the instructions as well as the composition of the worksheets. Furthermore, it also provided practice in running the workshop and increased confidence in the research design from the author (Sanders & Stappers, 2016, p.166).

#### **4.1.2. WORKSHOP PARTICIPANTS**

For the participatory design workshops, three professionals from the field of designer had been recruited. The profiles ranged from a Product Design Lead to a Master student in Human-Computer Interaction (HCI) to a Usability Testing Expert and Digital Business Expert in User Experience and Service Design. The professional tenure varied between the participants, but all of them had multiple years of experience working in the design field. The product design lead brought around nine years of experience in various product and user experience (UX) design roles. The Usability Testing and Service Design Expert had a decade of experience working in different UX research roles in software companies as well as multinational corporations. The Master's student had multiple years of experience working as a web designer in different roles.

## **4.2. WORKSHOP ACTIVITIES**

The workshops in this research proposition consisted of three parts. The first part is a short introduction, explaining the structure of the workshop, the activities and the material. This is followed by an individual mapping activity, in which the participants were working alone, at their own pace and time. The final part is a collaborative activity, where the results of the individual activity were compared, discussed and further refined.

### **4.2.1. WORKSHOP INTRODUCTION**

The workshops commenced with a short introduction to the participants, explaining the research goal, the structure of the workshop, the provided material and the tasks they were supposed to perform during the different activities. Further, the participants were asked whether they would consent to have the session recorded, under the premise that the recordings will be deleted after the research has been completed. Also, some basic information about the participants has been collected, namely their current role as well as their number of years of experience as a design professional. Lastly, any questions the participants had before the activities were addressed.

### **4.2.2. MAPPING ACTIVITY**

After the introduction, the participants had to leave the video conferencing call to work on the individual activity by themselves. As the goal of the activity was to better understand and capture unique individual perspectives, it's recommended to ask participants to work individually (Sanders et al., 2010, p.197). In addition, this also helped to mitigate any kind of bias or influence from the researcher, even in unconscious ways by making the participants feel observed during the activity. The individual activity consisted of a mapping exercise, which was an adaptation of the Card Sorting method. Card Sorting is a participatory design technique to explore how participants relate concepts to each other (Martin & Hannington, 2012, p. 26). Furthermore, according to the participatory design framework defined by Sanders et al, 2010, p.197, activities involving card-based artifacts are useful for the purpose of understanding the perception and experience of participants. In this particular case, the traditional card-sorting method was adapted to fit the need and purpose of the research goal. During the individual activity, the participants needed to decide if a design principle could be

related to at least one phase of a design process, if it would contribute to one of three trust qualities and if they could suggest an existing method or tool which would be useful to address a principle. To allow for this kind of multi-dimensional mapping, the participants didn't have to organize "principle-cards" into different categories, but the mapping happened in a more formalized process by selecting from predefined options in a Google spreadsheet. Although this interaction is less immediate and deviates from a traditional card sorting exercise, there are multiple reasons for this adaptation. Initially, it was considered to use an online whiteboard tool like "Miro" ([www.miro.com](http://www.miro.com)) "Mural" ([www.mural.com](http://www.mural.com)) or "Conceptboard" ([www.conceptboard.com](http://www.conceptboard.com)), which would have allowed direct manipulation of digital representations of the "principle-cards". However, the still relatively large number of items (29 design principles + 3 trust qualities), plus their respective descriptions, would have made this interaction cumbersome and potentially confusing in such digital whiteboard environments. In addition, there was a risk that participants would spend too much time exploring the tools if they wouldn't be "supervised" during the activity, thereby lengthening the process unnecessarily. Other specific online card sorting tools like "Optimal Sort" ([www.optimalworkshop.com/optimalsort](http://www.optimalworkshop.com/optimalsort)) were explored, as well. Yet again, the peculiarities of the research, in particular the quantity and length of the accompanying descriptions of an item, as well as the multi-dimensional mapping, rendered such tools inexpedient for the purpose. Considering the requirements of the research, combined with the structure of the research tasks, a spreadsheet approach was decided to provide the best option.

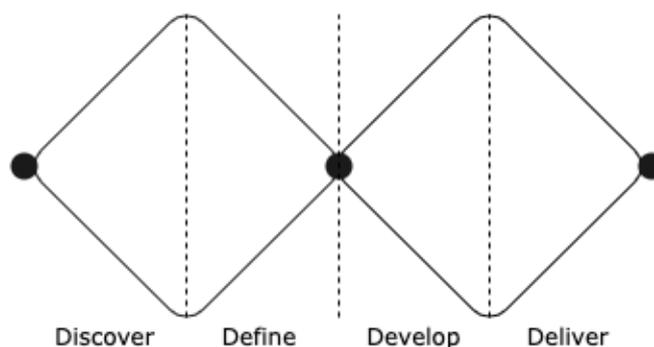
The worksheet was created in Google Sheets, containing an introduction sheet, two task sheets and three material sheets. The worksheet was shared with the participants so that they could interact with the spreadsheet on their own devices. The introduction sheet contained a short summary of the research goal and a task description, which mirrored the explanation given during the introduction session. Further, it also contained links to the respective sections in the worksheet. The first task sheet contained the list of the 29 principles identified during the prototype design. Each row of the task sheet contained one principle, one selectable field for the design process mapping, one field for the trust quality mapping, one field for suggesting methods or tools and one field for adding optional comments.

#### 4.2.2.1. DESIGN PROCESS MAPPING

The individual activity for the participants consisted of two mapping tasks and an open and optional tool suggestion. The first task was to determine if a principle could be mapped onto at least one phase of a design process. This exercise intended to learn whether the participants would agree with the initial judgment that a principle could be understood as approachable from a design perspective. The mapping onto design process phases was designed to avoid a binary choice resulting from a generic question like: “Do you understand this principle as a design-related principle?”, which would have allowed for a Yes/No answer. By enforcing a closer examination of a principle, the aim was to achieve higher confidence in the judgment.

There seem to be as many variations of design processes as there are writers about it. According to Jones (1992, Buxton, 2007, p.231) “there is little support to the idea that designing is the same under all circumstances, [...] the methods proposed by design theorists are just as diverse as are their descriptions of the design process”. There exist a myriad of design process models and definitions, from the well-known examples like the Design Thinking model developed at Stanford University, the resulting Human-Centered Design process propagated by IDEO (2015) to other variations, like the processes described by Archer in 1965 or Lawson in 1990. For the individual activity of the workshops, the so-called “Double Diamond” design process model provided the dimensional framework of the evaluation (Design Council, 2005).

**Figure 4:** Simplified illustration of the Double Diamond design process model  
(Adapted from [www.designcouncil.org.uk](http://www.designcouncil.org.uk))



In its basic form, the Double Diamond process model describes four distinct phases, in which it maps the divergent and convergent stages of the design process, showing the different modes of thinking that designers use over the course of a project (Design Council, 2007, p.6).

### **Discover**

“The objective of the Discover stage is to act as a ‘phase of divergent thought’, where the designers and other project team members keep their perspectives wide to allow for a broad range of ideas and influences. In this stage of the design process, the company is asking a question, posing a hypothesis or identifying a problem by analyzing market data, trends and other information sources” (Design Council, 2007, p.8).

### **Define**

“The Define stage should be thought of as a filter where the review, selection and discarding of ideas takes place. This is where findings from the Discover stage are analysed, defined and refined as problems, and ideas for solutions are pitched and prototyped” (Design Council, 2007, p.14).

### **Develop**

“At the Develop stage the project has been taken through a formal sign-off, which has given the corporate and financial backing to the development of one or more concepts that have addressed the initial problem” (Design Council, 2007, p.19).

### **Deliver**

"The Deliver stage of the Double Diamond design process is where the final concept is taken through final testing, signed-off, produced and launched. It will result in a product or service that successfully addresses the problem identified during the Discover stage. It will also include processes for feeding back lessons from the full design process to" (Design Council, 2007, p.23).

The Double Diamond process model was extensively scrutinized by a study conducted by the UK Design Council in 2007, to learn more about the design processes used by design departments of world-leading companies. The study studied the design processes at eleven companies, what elements they involve, and how these processes take a product or service from an idea through to implementation and launch. There are more complex variants of the

Double Diamond process available, but to keep the mapping exercise as simple as possible, the elementary version was used.

The task given to the participants of the workshop was to decide if a design principle could be related to at least one of the design process phases. It was specifically mentioned that a decision doesn't have to be exclusive, meaning that if a principle could be related to multiple phases, the participant should indicate this in the worksheet, for example making a note in the field for comments. Furthermore, it was also highlighted to the participants that if they feel a principle could not be mapped to any of the design phases, this would be a perfectly legitimate and valid option.

#### **4.2.2.2. TRUST QUALITY FILTER**

The second task of the individual activity was to decide whether a principle could contribute to one of three trust qualities defined in the Human Computer Trust Model (HCTM). The HCTM was initially developed by Sousa et al. in 2014, with the initial version consisting of seven principles, namely motivation, willingness, competence, benevolence, predictability, honesty and reciprocity, which would predict trust in user technology interaction. In a 2017 study, the HCTM was evaluated by Gulati et al. with the Estonian i-voting system. It was found that the significant attributes in relation to trust formation resulting from that study were competence, benevolence and honesty. In another study from 2018, testing the HCTM on Apple's intelligent assistant Siri, it was found that the attributes competence, benevolence, reciprocity and risk perception have an effect on the trust users have while interacting with technology (Gulati et al., 2018). In their latest HCTM validation from 2019, the HCTM was further refined and it was found that specifically the trust qualities risk perception, benevolence and competence were statistically significant in relation to trust formation in user technology interactions (Gulati et al., 2019).

#### **Risk perception**

Risk perception is a subjective assessment on the part of the end user of the probability of a specified type of incident happening when using a technical artefact and how concerned they are with the consequences of their action (Gulati et al, 2019, p.4). It is the extent to which one party is willing to participate in a given action while considering the risk and incentives

involved. The higher a risk is perceived, the less there is a willingness to interact. And vice versa, the lower a perceived risk, the higher the willingness to interact with a system.

### **Benevolence**

Benevolence describes a user's perception that a particular system will act in their best interest and that most people using the system share similar social behaviors and values. According to the social response theory, people using technology understand it as a social actor (Gulati et al., 2017, p.9). The emerging relationship is governed by the same social rules applicable to interpersonal relationships. Users of technology expect certain human-like behavior from it, i.e. the technology would act in their best interest and not try to deceive them. Benevolence is understood as technology being able to provide adequate, effective and responsive help to the end user which eventually helps the user attain specific goals related to their interaction (Gulati et al., 2017). When individuals perceive that technology would help them and act in their best interest, there is a likelihood of higher continued use and fostering a higher level of trust with that technology.

### **Competence**

Competence of a system is a direct representation of whether or not it has all the features and the functionalities to perform its intended tasks (McKnight et al. 2011 in Gulati et al., 2019, p.5). The system is capable of doing what the user needs it to do, performs reliably and delivers accurate results. The ease of use is associated with a system in that it is perceived to perform its tasks accurately and correctly - the extent to which the technology performs its functions properly (Gulati et al, 2017). According to the Technology Acceptance Model (TAM), a user would interact with and accept the technology when they perceive it to be useful, easy to use and perform as it says (Davis, 1989). If technology is perceived as easy to use, the perceived behavioral control (PBC) would also be higher, meaning it's likely that users will use and interact with the technology and the higher the competence that a user would assign, the greater would be the trust of the user with that system. Essentially, if a user perceives technology to be competent (i.e. having all desired functionality to achieve a particular outcome), then there is a high likelihood that they would place trust and act on the advice and recommendations offered by the technology (Gulati et al., 2017).

Provided with descriptions of the trust qualities, the participants should decide if a particular principle would contribute to one of the three trust qualities. Again, deciding that none of the trust qualities would relate to the principle, was a valid option.

#### **4.2.2.3. OPTIONAL TOOL SUGGESTION**

In addition to the two compulsory mapping tasks, the participants were also asked to mention any tool or method they would think could be beneficial when working with a particular principle. This additional request served two purposes. First, it would open up the possibility to find more existing tools in relation to the research sub-question 1: SRQ1: *“What current design practices exist to support the translation of existing trustworthy AI guidelines into practice?”* In addition, it would also provide additional indication and direction on which of the tools identified during the literature review, would be similar to the tools suggested by the participants. With similarity it is meant that a tool would share either purpose, structure or scope with another tool from the list.

#### **4.2.3. COLLABORATIVE ACTIVITY**

The final stage of the workshop was a collaborative activity, together with the researcher, with the goal of evaluating the toolkit prototype and co-creating the next iteration together. The collaborative generative research session built upon the results from the individual activity. According to Sanders et al. (2010, p.197) the individual expressions set the stage for successful collaboration in later activities, as it's often in the collaborative act of making, telling or enacting that innovation occurs. A key feature of running generative sessions is to combine the participatory activities with verbal discussions (Martin & Hanington 2012, p.94). In preparation for the collaborative activity, the author had also completed the mapping exercise separately. The toolkit prototype served as an artefact to trigger an engaged and comfortable conversation between the participant and the author. Therefore the activity was structured by going through each of the principles, discussing the respective decisions made on each mapping-dimension, as well as trying to find a consensus for the next iteration of the toolkit.

#### 4.2.3.1. CONTEXTUAL LADDERING

To get a holistic understanding of how participants have understood a particular principle, how they decided on a mapping and also why they did so, a variation of the contextual laddering technique was applied. Contextual laddering is a one-on-one interviewing technique that helps researchers to understand coherences and contexts in the answers given by the participants. It originated in consumer research and relies heavily on Means-End Theory to understand attributes, consequences and values in relation to a specific product or service. For the laddering, the interviewees are asked multiple “Why-” questions, allowing the interviewer to explore the links among attributes, consequences and values, as it was found that these factors influence purchasing decisions (Gutman, 1982). For this research, the technique was adapted to shift its consumer research focus, to a design methodological one. Instead of probing for the dimensions “*Attributes*” > “*Consequences*” > “*Values*” as the relevant stages for the laddering, the dimensions of “*Relevance*” > “*Consequences*” > “*Values*” were used. As the conversation would unfold, the constructed “ladder” would reveal the understanding of a principle and how a participant describes its utilization and benefit for the design of an AI system. The first question asked to the participants was if they had mapped a design principle to one or more design process phases, and if so, which one they had chosen. This question established the baseline for the laddering, as it manifests a general impact of a principle in relation to the design of an AI system. Following the first question, the participants were asked, why they had mapped the principle to one or multiple design process phases. The aim of this second “rung” of the latter was to learn about the consequences and benefits a participant would perceive about using that principle in the design process. Phrasing the question as a “Why”-question instead of a closing question, opened up the discussion space and avoided a Yes/No answer, which would have stifled the conversation. At the third “rung” of the laddering, the follow-up question was chosen depending on the answer from the previous conversation. The goal of this last stage was to elicit the underlying values a participant would associate with a specific principle. In addition to the collaborative mapping comparison, there was a final evaluation of the toolkit conducted together with the participants to identify and learn about the perceived usefulness, satisfaction and efficiency of the toolkit prototype.

#### 4.2.3.2. TOOLKIT PROTOTYPE EVALUATION

In addition to the mapping discussion, the collaborative activity also included a short evaluation task, in which the participants should rate the toolkit prototype on a Likert scale in the dimensions of “usefulness”, “satisfaction” and “efficiency”. This rating was implemented to elicit a final perception and evaluation of the idea of the toolkit, as well as its current state. At the end of the collaborative mapping comparison and the tool suggestion discussion, the participants were asked how they would rate the prototype of the toolkit in regards to the dimension of “usefulness”, “satisfaction” and “efficiency”.

**Table 4:** Likert scale evaluation choices

<b>USEFULNESS</b>	<b>SATISFACTION</b>	<b>EFFICIENCY</b>
Very useful	Very satisfactory	Very efficient
Somewhat useful	Somewhat satisfactory	Somewhat efficient
Not that useful	Not that satisfactory	Not that efficient
Not useful at all	Not satisfactory at all	Not efficient at all

Likert scales are a proven research method for measuring opinions, attitudes, beliefs and satisfaction with products (Sharp et al., 2007, p.314). In its usual form, a Likert scale provides an odd number of options, with a neutral, midpoint choice (Oppenheim, 1992, p.195ff). For this evaluation, an adapted Likert scale was used with an even number of options. The decision to use an even number of options was essentially twofold. First, by enforcing a choice, it helped to identify a general tendency of the perception. Due to the small number of participants, there was a risk of ending up with non-evaluable feedback if participants would end up choosing the neutral midpoint option. Removing the neutral choice helped to overcome this risk. Further, it also eliminated possible misinterpretations of the meaning of the scale midpoint (Troy, 2014).

### 4.3. RESULTS

During the first part of the collaborative activity, the mapping results between the mapping done by the participant and the one done by the author were compared. For each principle, the mappings have been discussed to validate the general understanding as a design problem, the idiosyncrasies of a principle in relation to a design process phase and the attribution to a trust quality. For these conversations, the laddering technique allowed deep introspection and

contemplation on fundamental values as well as potential considerations about consequences and the elicitation of underlying values.

#### **4.3.1. REMOVAL OF NON-RELEVANT PRINCIPLES**

The first thread of the collaborative activity was the evaluation of each principle in regards to its relevance as a design principle. As mentioned above, the mechanism to identify and assess the relevance of a principle consisted of the general design process mapping. The mapping exercise should prompt the participants to perform a critical analysis of the principle. The laddering technique would help further to unearth deeper understandings and valuations of each principle. If it would not be possible to attribute a principle to one or multiple design process phases, the relevance of a principle would be put into question and a principle removed from the toolkit.

##### **Example - Precaution**

The design principle of “Precaution” had been removed from the prototype, as most of the participants had not assigned any design process phase. After probing into the deeper reasons of these selections, it became apparent that the concept of precaution was not considered to be addressable from a design perspective. Rather the principle pertains to more far-reaching aspects of a product or service development, touching on topics like data security, human resources and internal education.

##### **Example - Responsibility**

The principle of “Responsibility” had been removed from the toolkit, as it was considered to be a fundamental requirement, not a design principle. During the laddering, it was described as a prerequisite for the general mindset of involved stakeholders, designers and engineers. One workshop participant considered it to be of “introductory character”, as it should be understood as a stringent necessity before even embarking on a project.

##### **Example - Trustworthiness**

The principle of “Trustworthiness” was defunct from the toolkit, as it was considered to be recursive in regards to the general goal of the toolkit to design for trustworthy AI. Despite that some participants had assigned a design process phase to the principle, during the laddering questions it became clear that there was reservation and confusion about this specific

principle. For the sake of keeping the toolkit action and easy to understand, the principle was removed.

#### **4.3.2. EXTERNAL REDUNDANCY MITIGATION**

Another thread in the collaborative activity was the mitigation of redundancies with already available frameworks, norms or guidelines. This section in the research design aimed to keep the design principles as relevant as possible by avoiding the repetition of common and established practices. The participants were prompted, if they would think that a design principle from the toolkit prototype was already addressed, or at least affected, by an existing guideline or recommendation they know from their professional practice. To initiate the thought process, the participants were given examples of such guidelines, like the “Usability Heuristics” defined by Jakob Nielsen (1994) or the ISO 9241-210 norm, which provides requirements and recommendations for human-centered design principles (International Organization for Standardization [ISO], 2019). The conversations that spawned from this prompt resulted in interesting discussions on the relevance and scope of the existing guidelines compared to the design guidelines defined in the toolkit prototype.

##### **Example - Accessibility**

The design principle of “Accessibility” had been removed from the toolkit, as there were multiple sources mentioned which already address this dimension during the design of technological artifacts. One participant stated, that accessibility is specifically listed in the ISO-9241-210 norms as the “extent to which products, systems, services, environments and facilities can be used by people from a population with the widest range of user needs, characteristics and capabilities to achieve identified goals in identified contexts of use (ISO, 2019). Furthermore, as AI and ML technology is also applied on websites, browser applications and other systems on the world wide web, the existing Web Content Accessibility Guidelines (WCAG 2.0, 2008) were also mentioned as existing norms.

##### **Example - Human Agency**

The principle of “Human Agency” was also removed from the toolkit, as it was found that the existing Usability Heuristic #3 - “Control and freedom” does address this principle already. In the toolkit prototype, there were two similar, yet slightly different principles addressing the capabilities of control, interference and power of decision: Human Agency and Human

Oversight. The participants mentioned that the conceptual understanding agency would be sufficiently addressed in the Usability Heuristic, yet the aspect of Human Oversight would not. Therefore the principle was eliminated.

For some of the principles from the toolkit, there were guidelines and directives mentioned and discussed during the workshops, but they were still kept in the toolkit. For these principles, their relevance was balanced against their redundancy, and whether the existing guidelines would still be considered to be sufficient enough in the relation to designing for trustworthy AI products and services.

#### **Example - Reversibility**

One participant mentioned in the discussion, that the principle of “Reversibility” could potentially be understood to fall under the remit of the Usability Heuristic #1: “Visibility of system status”. The heuristic says: “The design should always keep users informed about what is going on, through appropriate feedback within a reasonable amount of time” (Nielsen, 1994). However, it was mentioned that in relation to AI systems, it might not be immediately clear that the description of the Usability Heuristic would also specifically entail the aspect of reversibility. Therefore it was suggested to keep the principle in the toolkit, despite a seeming redundancy, as its relevance would be considered more important.

#### **Example - Transparency**

Also for the principle of “Transparency”, the existing Usability Heuristic #1 (Visibility of system status) was mentioned in multiple workshops. It was stated that for an AI system to be transparent, it needs to provide an accurate representation of the system's status. However, for advanced and complex systems, which include AI and ML applications, the importance of transparency is considered to be of much higher importance. Therefore it was decided to keep the principle in the toolkit.

#### **4.3.4. SEMANTIC CONSOLIDATION**

The final thread of the evaluation of the principles from the toolkit prototype was to consolidate principles where possible, with the intention to reduce semantic overlap within the toolkit. Again, the laddering technique allowed for specific probing into considerations about consequences as well as relevant and underlying values within the participants. This

helped to better comprehend how participants understood particular principles and how they would foresee their application. This was the most sensitive part of the activity, as it was important to not dilute the importance and relevance of a specific principle when merging it with another, similar principle.

**Example - Beneficence, Justification & Non-subversion → Purpose**

In the toolkit prototype, there were multiple principles concerning the creation of added value. The principle of “Beneficence” was described as “complimenting the human experience”; “Justification” as “clear purpose and linked to a benefit” and “Non-subversion” as “respect and improve the life of citizens”. Based on the conversations with the participants emerged the consensus that all three principles address a primal concept, which is to ensure that there is a clear need for the creation of the AI system. Evaluating the value and benefit of a product or service is core to this principle. For the merged principle, the term “Purpose” was chosen

**Example - Consideration of long term effects, impact assessment → Impact assessment**

Another consolidation was done for the two principles of “Consideration of long term effects” as well as “Impact Assessment”. Although there was a difference in the provided descriptions, where “Impact Assessment” would put a specific focus on the documentation of potential consequences, it was agreed that this additional perspective could be added to a merged description. In addition, the specifically mentioned process phases in the description were found to be not specifically relevant, since the assessment and evaluation of potential ramifications should be done along the whole design and development process. The merged principles kept the term “Impact Assessment”.

**Example - Diversity, Inclusion, plurality, solidarity → Inclusion**

The concepts of “Diversity” and “Inclusion” are of fundamental relevance, regardless of the product or service. Yet in the context of AI systems, they become even more important when designing and developing, considering the influence and impact such systems have. During the workshops, it was discussed how to ensure that the included principles would live up to their purpose and not tarnish each other because of a lack of focus. Therefore it was decided to merge the principles “Diversity”, “Inclusion”, “Plurality” and “Solidarity” and to only keep one dedicated principle in the toolkit, providing focus, rather than dilution. The principle was chosen to keep the name “Inclusion”.

### **Example - Explainability, Understandability → Explainability**

“Explainability” is a relevant concept and area of research in itself. However, it also affects the trustworthiness of a system (Doran, Schulz & Besold, 2017). Furthermore, during the workshops, it was identified that “Explainability” induces “Understandability”, which makes the former principles more relevant than the latter. It was decided to drop the “Understandability” principle and only keep the “Explainability” principle in the toolkit.

### **4.3.3. TOOLS AND METHODS SUGGESTIONS**

As mentioned above, the worksheet also included an optional column to mention specific tools and techniques, which the participants would consider to be applicable and useful when working with a particular principle. The goal of this additional question was to harness the expertise of participants being design professionals. The assumption was that by asking for proven methods and tools from the field, the chasm between theoretical knowledge and practical application could be alleviated. However, the outcome of this question was rather negligible. Only one of the participants addressed this question throughout the whole list of principles, providing some suggestions and indications about specific methods as well as research types that could be considered.

### **Example - Dignity**

The participant suggested that to address the principle of “Dignity”, ethnographic research would be particularly useful and necessary. The reasoning behind this was, that despite dignity being a fundamental value across humanity, there are cultural differences which can affect the preservation or loss of dignity. Especially in relation to AI, automated processes and “cold” data outcomes, dignity becomes an even more important concept to be conscious about, across cultures, locations and age groups.

### **4.3.4. EVALUATION**

To address research sub-question 3 *“How does the toolkit enable the design of trustworthy AI systems in the dimensions of “usefulness”, “satisfaction” and “efficiency?”*, the worksheet also included a final task, rating the concept of the toolkit in regards to these dimensions. The toolkit prototype had been evaluated “Very useful” by all participants. The need and domain of application were acknowledged and the development of such a toolkit was welcomed.

However, the current state was considered to be not satisfactory and efficient enough. The critique described the toolkit as containing too many doubling-ups and in need of clearer and distinct descriptions.

#### **4.3.5. TOOL IDENTIFICATION**

To provide the most practical benefit for designers, the toolkit was also supplemented with suggestions about which tools could be used when addressing a particular principle. The tool suggestions are based on two sources: First, the list of existing tools and methods identified during the literature review were consulted. The second source was supposed to be based on suggestions coming from the design professionals during the participatory design workshops. However, as mentioned earlier, the results of the latter source weren't as comprehensive as hoped.

#### **Example - Stakeholder participation**

The concept of involving stakeholders during the design and development process of technology, is neither new nor revolutionary. In fact, it should be second nature for designers working with a human-centered mindset. However, for the context of AI, the criteria of who qualifies as being a stakeholder potentially expands, or at least changes from common technology development. The Ethics & Society team at Microsoft has created a card game called "Judgement Call" (Ballard et al., 2019), which is specifically catered for addressing ethical concerns related to AI. During the game, product teams identify stakeholders and write fictional product reviews from those stakeholders' perspectives. These reviews, which are related to ethical principles, spark conversations about the respective ethical concerns (Ballard et al., 2019, p.424). The game is based on concepts from value sensitive design (Friedman & Hendry, 2019) and design fiction (Baumer et al., 2018).

#### **4.5. REFLECTION**

Through the workshops, it was possible to co-design the next iteration of the toolkit in close collaboration with its intended users – design professionals. The collaborative activity resulted in some highly engaged conversations which helped to advance the next iteration of the toolkit. The discussions with the practitioners revealed feedback on a conceptual level, the practical application as well as some research design suggestions for improvements.

Principles had been removed for a lack of relevance as well as redundancy. Furthermore, other principles had been consolidated and merged and their descriptions clarified to provide the most useful and actionable set of items for the toolkit. The list of principles had been reduced to a number of 16 principles identified as applicable for the design of trustworthy AI systems. However, some flaws and friction points were identified during the workshops. It was ascertained that specific allocations into a design process phase, or regarding a potential trust quality wouldn't be considered practical. It was acknowledged that within the research context the question if a principle could be associated with a design process phase as well as a trust quality made sense to validate its general relevance. Yet, when envisioning the toolkit being used in the field, those would be rather seen as constraints, hindering the explorative benefit of the principles. Limiting the scope toward specific phases or trust qualities and could hamper the general use of the toolkit.

#### **Example - Human oversight**

During multiple workshops, the principle “Human oversight” was mapped onto the “Define” phase. Probing on why it was considered to be relevant during this phase, the participant explained that a user should always keep the power of the ultimate decision, and this feature needed to be incorporated from the beginning. The choice of opting out of an automated decision is not an afterthought, which can be just added at the user interface layer, but it needs to be considered already during the architectural planning process, to ensure that the option to contest and object is technically even possible. However, this principle would further still be relevant in later phases of the design process. Imposing a false limitation wouldn't do justice to the relevance of the principle.

Therefore it was decided that the next iteration of the toolkit should not mention any specific design process phases or trust quality assignments. The toolkit would be considered more useful when applicable across all phases, working as a general means and advisor when working on AI products, services and systems.

## 5. DISCUSSION

### 5.1. THE RESULTS ACHIEVED

The toolkit which has been developed in this thesis research project consists of 16 design principles, their respective descriptions as well as suggested tools when working with these principles. The ultimate list of principles is the result of an extensive literature review, during which common themes and guidelines for the development of ethical AI systems have been identified. Based on the most common ethical AI guidelines, the specifically relevant principles for the design towards the value of trust have been elicited. The toolkit has been co-developed and co-designed in collaboration with several design practitioners from the field as well as a subject matter expert in the area of trust in technology. The principles have been validated in multiple participatory workshops and underwent extensive scrutiny during the co-design sessions. The toolkit should help designers working with AI systems, to consider the important and pertinent aspects to define and design trustworthiness into the product or service.

**Table 5:** Toolkit for the design of trustworthy AI [Tool ID in Appendix 1]

PRINCIPLE	DESCRIPTION	TOOL SUGGESTION
Autonomy	AI organisations should ensure that end users are informed, not deceived or manipulated by AI and should be allowed to exercise their autonomy (Ryan & Stahl, 2020, p.14).	<ul style="list-style-type: none"> <li>AI &amp; Ethics Cards [1]</li> </ul>
Consent	The use of personal data must be clearly articulated and agreed upon before its use (UNDG, 2017 in Ryan & Stahl, 2020, p.14).	<ul style="list-style-type: none"> <li>Ethics Kit [27]</li> </ul>
Dignity	AI should be developed and used in a way that respects, serves and protects humans physical and mental integrity, personal and cultural sense of identity, and satisfaction of their essential needs (HLEG-AI, 2019, p. 10, in Ryan & Stahl, 2020, p.15).	<ul style="list-style-type: none"> <li>Ethnographic Research</li> <li>Expanding The Ethical Circle [23]</li> </ul>
Explainability	The translation of technical concepts and decision outputs into intelligible, comprehensible formats suitable for evaluation (Fjeld et al. 2020, p.42).	<ul style="list-style-type: none"> <li>Google PAIR [3]</li> <li>LIME [11]</li> </ul>
Fairness	There should be steps in place to ensure that data being used by AI is not unfair, or contains errors and inaccuracies, that will corrupt the response and decisions taken by the AI (ICO, 2017 in Ryan & Stahl, 2020, p.7).	<ul style="list-style-type: none"> <li>Trustworthy AI Factsheet [14]</li> <li>Moral Value Map [20]</li> <li>UnBias Fairness Toolkit [28]</li> </ul>
Human Oversight	The “ability to opt out of automated decision” principle is defined, as affording individuals the opportunity and choice not to be subject to AI systems where they are implemented (Fjeld et al. 2020, p.54).	<ul style="list-style-type: none"> <li>Google PAIR [3]</li> <li>Assessment List for Trustworthy Artificial Intelligence [6]</li> <li>Trustworthy AI Factsheet [14]</li> </ul>

Impact Assessment	The objectives and expected impact of AI must be assessed, reviewed and documented on an ongoing basis (Algo.Rules, 2019 in Ryan & Stahl, 2020, p.7).	<ul style="list-style-type: none"> <li>• Trustworthy AI Factsheet [14]</li> <li>• Data Ethics Canvas [15]</li> <li>• Layers of Effect [22]</li> <li>• Black Mirror/White Mirror [24]</li> <li>• Consequence scanning [32]</li> <li>• Envisioning Cards [33]</li> </ul>
Inclusion	Attention should be given to under-represented and vulnerable groups and communities, such as those with disabilities, ethnic minorities, children and those in the developing world. Data that is being used should be representative of the target population and should be as inclusive as possible (HLEG-AI, 2019, p. 10, in Ryan & Stahl, 2020, p.8).	<ul style="list-style-type: none"> <li>• Judgement Call [10]</li> <li>• Data Ethics Canvas [15]</li> <li>• Expanding The Ethical Circle [23]</li> </ul>
Non-Bias	Developers should examine unfair biases at every stage of the development process and should eliminate those found (Ryan & Stahl, 2020, p.8).	<ul style="list-style-type: none"> <li>• AI &amp; Ethics Cards [1]</li> <li>• Data Ethics Canvas [15]</li> </ul>
Non-Discrimination	AI should be designed for universal usage and not discriminate against people, or groups of people, based on gender, race, culture, religion, age or ethnicity (Cerna Collectif, 2018 in Ryan & Stahl, 2020, p.8).	<ul style="list-style-type: none"> <li>• Data Ethics Canvas [15]</li> <li>• UnBias Fairness Toolkit [28]</li> </ul>
Privacy	Users should have control and access to data stored about them (The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2019, in Ryan & Stahl, 2020, p.12).	<ul style="list-style-type: none"> <li>• AI &amp; Ethics Cards [1]</li> <li>• Google PAIR [3]</li> <li>• Judgement Call [10]</li> <li>• Ethics Canvas [29]</li> </ul>
Purpose	The purpose for building the system must be clear and linked to a clear benefit —system’s should not be built for the sake of it (Morley et al., 2020, p.2151).	<ul style="list-style-type: none"> <li>• Google PAIR [3]</li> <li>• Data Ethics Canvas [15]</li> <li>• Design for Trust [16]</li> </ul>
Reversibility	It is important to clearly articulate if the outcomes of AI decisions are reversible. The ability to undo the last action or a sequence of actions allows users to undo undesired actions and get back to the ‘good’ stage of their work (Personal Data Protection Commission Singapore, 2019, p. 16; Clark, 2019 in Ryan & Stahl, 2020, p.9).	<ul style="list-style-type: none"> <li>• Trustworthy AI Factsheet [14]</li> </ul>
Showing	It should be clear to the end user that they are interacting with an AI system, rather than a human. Further, where an AI has been employed, the person to whom it was subject should know (EPSRC, 2011, Ryan & Stahl, 2020, p.7).	<ul style="list-style-type: none"> <li>• Ethically Aligned Design [18]</li> <li>• Design with Intent [25]</li> </ul>
Stakeholder participation	To develop systems that are trustworthy and support human flourishing, those who will be affected by the system should be consulted (Morley et al., 2020, p.2151).	<ul style="list-style-type: none"> <li>• Judgement Call [10]</li> <li>• Ethical contract [20]</li> <li>• Envisioning Cards [33]</li> </ul>
Transparency	The principle of “transparency” is the assertion that AI systems should be designed and implemented in such a way that oversight of their operations are possible (Fjeld et al. 2020, p.42).	<ul style="list-style-type: none"> <li>• Judgement Call [10]</li> <li>• Data Ethics Canvas [15]</li> </ul>

This iteration of the toolkit constitutes the final version for this research project, but this should not mean that the discovery is done here. The toolkit needs to be further developed and improved and field-tested to further enhance its usefulness, satisfaction and efficiency. In section 6.4, there are suggestions about potential further research.

## 5.2. LIMITATIONS OF THE STUDY

During the research process, a couple of limitations have been identified. Those limitations have been documented to highlight potential shortcomings and to encourage further research. The first potential limitation can be delineated by the lack of a formalized theoretical framework during the initial phase of principle selection - the second pillar. Although the principles passed through multiple iterations of evaluation by the author, a design expert with multiple years of experience, as well as the subject matter expert on trust in technology, it could be argued that the selection has happened strictly speaking based on personal judgment. It is within the realms of possibility, that a different set of experiences and backgrounds could have yielded different results. It might be worthwhile to consider a similar study with a specific research method. One of the workshop participants suggested an initial affinity diagramming method, to improve the common understanding and overcome synthesizing difficulties (Martin & Hanington, 2012, p.12). Another potential deficiency within the research structure could be found in the selection of the descriptions for the toolkit prototype. Although the descriptions had been chosen to the best of the knowledge and belief of the author, it could have happened that in cases where multiple descriptions were mentioned in the original list, this pre-selection might have influenced the understanding of the principles in the participants. However, this selection was done to avoid an unreasonably inflated toolkit, which would have increased the cognitive load in the participants during the workshop. An issue identified during the participatory design workshops was the varying understanding of the selected design process and its particular phases. Even though the Double Diamond process is a well-known and documented process and an explicit process description had been provided to the participants, some misunderstandings and confusions occurred. This rendered the comparison additionally complex. It might be that another design process model, e.g. the design thinking process or the human-centered design (HCD) process phases defined by IDEO could have been less confusing. However, this also raises the more fundamental question, if a human-centered design process is the best approach when approaching complex systems like AI or ML? Does an HCD approach maybe ignore or neglect other perspectives like environmental sustainability? Could a “More-Than-Human Design” approach, as discussed by Giaccardi & Redström (2020) be the more inclusive maxim? And a final potential limitation concerns the chosen approach of participatory design. Bratteteig & Verne (2018) state that “participatory design can help create ideas for the AI System through diverse methods. However, it only represents the design space for the period of the project, and users

need to have a basic understanding of what AI can do and not do” (Bratteteig & Verne, 2018 in Auernhammer, 2020, p.7). As much as the critique might seem justified, it is in the nature of design to address the “unknown”. Rittel & Webber (1973) describe “design, as a ‘satisficing activity’[...] dealing with messy indeterminate situations and “wicked problems”. [...]before designers can solve a design problem, they need to understand some basics, such as what they are designing, what it should do, and who should use it and in what circumstances” (Rittel & Webber, 1973 in Randall & Rouncefield, 2013, p.2107).

#### **5.4. IDEAS FOR FURTHER RESEARCH**

One of the most apparent suggestions for continuing this research would be to conduct another evaluation on the latest iteration of the toolkit in regards to the dimensions of “usefulness”, “satisfaction” and “efficiency”. For now, the evaluation happened on the prototype, without any usage of the toolkit, but only the crude list of principles and the general concept. The latest iteration has implemented the feedback from the workshops, which would also provide a good opportunity to link the evaluation with a real-world task, in which designers could field-test the toolkit in their own projects. Another opportunity would be to conduct more and slightly variant workshops. As Hair Jr et al. (2016) state, “when the aim of the study is to contribute to scale development, non-significant attributes from initial studies cannot simply be discarded without testing them in different ways and in different contexts to ensure proper construct validity and reliability” (Hair Jr et al. 2016 in Gulati et al, 2019, p.4). As stated above, there are parts of the research design that could be changed to learn how they would affect the outcome of the study, for example, if different pre-selected descriptions for the principles would yield different results in the mapping exercise. Or if potentially in the future the epidemiological situation allows for in-person research, the workshops could be conducted in multi-participant sessions, looking for potential synergies between the participants.

Another direction for continuing this research would be to turn the toolkit into a more actionable and practical format to be suitable for field implementation. One possible manifestation would be in the format of physical cards. Cards are considered playful and creative in their usage and that they can be applied in collaborative sessions (Lucero et al., 2016). Pärnpuu (2020, p.18-19) describes several implementations where design cards have proven their usefulness. They help to spark discussion and support reflection by being tangible

and accessible. Lucero et al. (2016, p.94) describe them as “tangible idea containers”. An exemplary attempt has been added to this thesis and can be found in the appendix (see Appendix 3). However an already identified drawback of the card format is its lack of digital interactivity. As the toolkit contains suggested tools and methods with the corresponding hyperlinks, maybe a digital format like a website would be even more practical.

## 5.5. CONCLUSION

This research project began with a gargantuan task, hidden in the main research question of *“How to help designers design for trustworthy AI?”*. This objective demanded to first understand what designing for trustworthiness means in the context of artificial intelligence. In a recent half-humoristic, half-cynical MIT Technology Review article, Karen Hao described trustworthiness as *“trustworthy (adj) - An assessment of an AI system that can be manufactured with enough coordinated publicity”*. Although this description aims to be excessive and political, it holds a grain of truth in relation to this research: How can design manufacture the assessment of trust in an AI system? And how can design help to apply existing ethical guidelines into the practical development of AI systems? The outcome of this research thesis, an iterated toolkit to enable the design of trustworthy AI, is the result of the three sub research questions SRQ1: *“What current design practices exist to support the translation of existing trustworthy AI guidelines into practice?”*, SRQ2: *“How to create a toolkit that will enable designers to design for the formation of trustworthy AI systems?”* and SRQ3: *“How does the toolkit enable the design of trustworthy AI systems in the dimensions of “usefulness”, “satisfaction” and “efficiency?”*. In an extensive literature review it has been recognized that the current culprit in designing for trustworthy AI, isn't a lack of theoretical suggestions or recommendations. It's rather even the sheer volume of available ethical guidelines which can result in paralysis and obstruct their practical application. It was identified that a practice-oriented compilation containing only trust-relevant principles with some helpful tool suggestions could be a potential remedy. In collaboration with design professionals as well as experts in the field of trust in technology, relevant design principles have been identified, compiled and iterated on. Furthermore, a list of useful methods and tools has been composed to undergird the theoretical framework with hands-on suggestions to apply in practice. This toolkit does not aim to be the answer to all the questions and concerns related to the progress of AI. But it provides a small advancement in the field of designing for trustworthy AI, as it addresses the aforementioned lack of knowledge, skills and roles in the

field of design for beneficial technologies. And it's not alone in this endeavor. Others devote themselves to the same goal, like the Danish Design Center in Copenhagen, which is working on an "Ethics Compass" for AI systems (Dansk Design Center, 2021). But providing tools and theoretical frameworks constitutes only one angle to address the situation. Governmental regulation is another one, which is necessary to ensure beneficial and flourishing progress. During the development of this research thesis, the European Commission proposed new rules and actions for excellence and trust in Artificial Intelligence. The proposition plans to regulate and even ban malicious and dangerous applications. But again, this doesn't absolve designers and developers of advanced technology from their responsibility. If anything, it makes sensitive design just even more important.

## 6. REFERENCES

---

- AI Ethics Impact Group. (2020). *From Principles to Practice - An interdisciplinary framework to operationalise AI ethics*. VDE Association for Electrical Electronic & Information Technologies e.V., Bertelsmann Stiftung, 1–56.
- Algo.Rules. (2019). *Rules for the Design of Algorithmic Systems*. Bertelsmann Stiftung.  
<https://algorules.org/en/home>
- Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). *Guidelines for Human-AI Interaction*. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 1–13.  
<https://doi.org/10.1145/3290605.3300233>
- Amnesty International/Access Now. (2018). *The Toronto declaration: protecting the rights to equality and non-discrimination in machine learning systems*.  
[www.accessnow.org/cms/assets/uploads/2018/08/The-Toronto-Declaration\\_ENG\\_08-2018.pdf](http://www.accessnow.org/cms/assets/uploads/2018/08/The-Toronto-Declaration_ENG_08-2018.pdf)
- Andras, P., Esterle, L., Guckert, M., Han, T. A., Lewis, P. R., Milanovic, K., Wells, S. (2018). *Trusting Intelligent Machines: Deepening Trust Within Socio-Technical Systems*. 37(4), 76-83
- Antonov, A., & Kerikmäe, T. (2020). *Trustworthy AI as a Future Driver for Competitiveness and Social Change in the EU*. The EU in the 21st Century, 135–154.  
[https://doi.org/10.1007/978-3-030-38399-2\\_9](https://doi.org/10.1007/978-3-030-38399-2_9)
- Archer, L. B. (1965). *Systematic Method for Designers*. London. The Design Council.
- Arksey, H. & O'Malley, L. (2005). *Scoping studies: towards a methodological framework*. International Journal of Social Research Methodology, 8 (1). pp. 19-32.
- Arnold, M., Piorkowski, D., Reimer, D., Richards, J., Tsay, J., Varshney, K. R., Bellamy, R. K. E., Hind, M., Houde, S., Mehta, S., Mojsilovic, A., Nair, R., Ramamurthy, K. N., & Olteanu, A. (2019). *FactSheets: Increasing trust in AI services through supplier's declarations of conformity*. IBM Journal of Research and Development, 63(4/5), 6:1–6:13. <https://doi.org/10.1147/jrd.2019.2942288>
- Auernhammer, J. (2020). *Human-centered AI: The role of Human-centered Design Research in the development of AI*. DRS2020: Synergy, 1–19.  
<https://doi.org/10.21606/drs.2020.282>

- Ballard, S., Chappell, K. M., & Kennedy, K. (2019). *Judgment Call the Game*. Proceedings of the 2019 on Designing Interactive Systems Conference, 421–433.  
<https://doi.org/10.1145/3322276.3323697>
- Baumer, E. P., Berrill, T., Botwinick, S. C., Gonzales, J. L., Ho, K., Kundrik, A., Kwon, L., LaRowe, T., Nguyen, C. P., Ramirez, F., Schaedler, P., Ulrich, W., Wallace, A., Wan, Y., & Weinfeld, B. (2018). *What Would You Do?* Proceedings of the 2018 ACM Conference on Supporting Groupwork, 244–256.  
<https://doi.org/10.1145/3148330.3149405>
- Bietti, E. (2020). *From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy*. FAT\* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 210–219. <https://doi.org/10.1145/3351095.3372860>
- Bitkina, O. V., Jeong, H., Lee, B. C., Park, J., Park, J., & Kim, H. K. (2020). *Perceived trust in artificial intelligence technologies: A preliminary study*. Human Factors and Ergonomics in Manufacturing & Service Industries, 30(4), 282–290.  
<https://doi.org/10.1002/hfm.20839>
- Blackler, A., Swann, L., Chamorro-Koc, M., Mohotti, W. A., Balasubramaniam, T., & Nayak, R. (2021). *Can We Define Design? Analyzing Twenty Years of Debate on a Large Email Discussion List*. She Ji: The Journal of Design, Economics, and Innovation, 7(1), 41–70. <https://doi.org/10.1016/j.sheji.2020.11.004>
- Boden, M., A. (2016). *Ai: Its Nature and Future*. Oxford University Press, Incorporated. ProQuest Ebook Central,  
<http://ebookcentral.proquest.com/lib/cut-ebooks/detail.action?docID=4545415>.
- Bostrom, N. (2014). *The ethics of artificial intelligence*. In E. Yudkowsky (Ed.), *The Cambridge Handbook of Artificial Intelligence* (pp. 316–334). Machine Intelligence Research Institute.  
[https://www.cambridge.org/core/product/identifier/CBO9781139046855A027/type/book\\_part](https://www.cambridge.org/core/product/identifier/CBO9781139046855A027/type/book_part)
- Bratteteig, T., & Verne, G. (2018). *Does AI make PD obsolete?: exploring challenges from artificial intelligence to participatory design*. Proceedings of the 15th Participatory Design Conference: Short Papers, Situated Actions, Workshops and Tutorial - Volume 2, Hasselt and Genk, Belgium.
- Buxton, B. (2007). *Sketching User Experiences*. Morgan Kaufmann. Elsevier.
- Caldwell, B., Cooper, M., Reid, L. G., & Vanderheiden, G. (ed.). (2008). *Web Content Accessibility Guidelines (WCAG) 2.0*. Web Content Accessibility Guidelines Working Group World Wide Web Consortium (W3C).

- Cerna Collectif. (2018). *Research Ethics in Machine Learning*. [Research Report] CERNA; ALLISTENE. pp.51. <https://hal.archives-ouvertes.fr/hal-01724307>
- Clarke, R. (2019, April 15). *Roger Clarke's 'Principles for AI'*. <http://www.rogerclarke.com/EC/GAIP>
- Coeckelbergh, M. (2020). *AI Ethics*. Cambridge, MA: MIT Press.
- Colquhoun, H. L., Levac, D., O'Brien, K. K., Straus, S., Tricco, A. C., Perrier, L., Kastner, M., & Moher, D. (2014). *Scoping reviews: time for clarity in definition, methods, and reporting*. *Journal of Clinical Epidemiology*, 67(12), 1291–1294. <https://doi.org/10.1016/j.jclinepi.2014.03.013>
- Davis, F. D. (1989). *Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology*. *MIS Quarterly*, 13(3), 319–340. <https://doi.org/10.2307/249008>
- Design Council. (2005). *The Design Process*. <http://www.designcouncil.org.uk/designprocess>
- Design Council. (2007). *Eleven lessons: managing design in eleven global brands. A study of the design process*. London. [https://www.designcouncil.org.uk/sites/default/files/asset/document/ElevenLessons\\_Design\\_Council%20%282%29.pdf](https://www.designcouncil.org.uk/sites/default/files/asset/document/ElevenLessons_Design_Council%20%282%29.pdf)
- Det Digitale Etikkompas: Hvordan utvikler vi etisk teknologi?* (2021, March 4). Dansk Design Center. <https://danskdesigncenter.dk/da/etikkompas>
- Deutsche Telekom AG. (2018, May 11). *Guidelines for Artificial Intelligence*. <https://www.telekom.com/en/company/digital-responsibility/details/artificial-intelligence-ai-guideline-524366>
- Dignum, V. (2018). *Ethics in artificial intelligence: introduction to the special issue*. *Ethics Inf Technol* 20,1–3 . <https://doi.org/10.1007/s10676-018-9450-z>
- Doran, D., Schulz, S., & Besold, T. R. (2017). *What does explainable ai really mean? A new conceptualization of perspectives*. 1–8. <https://arxiv.org/abs/1710.00794>
- Ehsan, U. & Riedl, M. (2019). *On Design and Evaluation of Human-centered Explainable AI systems*.
- Engineering and Physical Sciences Research Council (EPSRC). (2011). *Principles of robotics - EPSRC*. <https://epsrc.ukri.org/research/ourportfolio/themes/engineering/activities/principlesofrobotics/>

- Europe fit for the Digital Age: Commission proposes new rules and actions for excellence and trust in Artificial Intelligence.* (2021, April 21). European Commission - European Commission. [https://ec.europa.eu/commission/presscorner/detail/en/IP\\_21\\_1682](https://ec.europa.eu/commission/presscorner/detail/en/IP_21_1682)
- European Parliament. (2017). *Report with recommendations to the commission on civil law rules on robotics.* Committee on Legal Affairs. [https://www.europarl.europa.eu/doceo/document/A-8-2017-0005\\_EN.html](https://www.europarl.europa.eu/doceo/document/A-8-2017-0005_EN.html)
- Ferrario, A. & Loi, M. & Viganò, E. (2019). *In AI We Trust Incrementally: a Multi-layer Model of Trust to Analyze Human-Artificial Intelligence Interactions.* Philosophy & Technology, 33. <https://doi.org/10.1007/s13347-019-00378-3>.
- Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI.* SSRN Electronic Journal, 1–68. <https://doi.org/10.2139/ssrn.3518482>
- Fleetwood, A., Unsworth, G., & Tobia, T. (2019, April). *Product Development with Consequence Scanning.* TechTransformed. <https://www.tech-transformed.com/product-development/>
- Floridi, L. (2019). *Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical.* Philosophy & Technology, 32(2), 185–193. <https://doi.org/10.1007/s13347-019-00354-x>
- Floridi, L., Cows, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). *AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations.* Minds and Machines, 28(4), 689–707. <https://doi.org/10.1007/s11023-018-9482-5>
- Floridi, L., Cows, J., King, T. C., & Taddeo, M. (2020). *How to Design AI for Social Good: Seven Essential Factors.* Science and Engineering Ethics, 26(3), 1771–1796. <https://doi.org/10.1007/s11948-020-00213-5>
- Friedman, B., & Hendry, D. G. (2019). *Value Sensitive Design: Shaping Technology with Moral Imagination* (The MIT Press) (Illustrated ed.). The MIT Press.
- Future of Life Institute. (2018, April 11). *Asilomar AI Principles.* <https://futureoflife.org/ai-principles/>
- Gambelin, O. (2020). *Brave: what it means to be an AI Ethicist.* AI and Ethics. <https://doi.org/10.1007/s43681-020-00020-5>

- Glikson, E. & Woolley, A. (2020). *Human trust in artificial intelligence: Review of empirical research*. *Academy of Management Annals* (in press). The Academy of Management Annals.
- Giaccardi, E., & Redström, J. (2020). *Technology and More-Than-Human Design*. *Design Issues*, 36(4), 33–44. [https://doi.org/10.1162/desi\\_a\\_00612](https://doi.org/10.1162/desi_a_00612)
- Gulati, S., Sousa, S., & Lamas, D. (2017). *Modelling Trust: An Empirical Assessment*. *Human-Computer Interaction – INTERACT 2017*, 40–61. [https://doi.org/10.1007/978-3-319-68059-0\\_3](https://doi.org/10.1007/978-3-319-68059-0_3)
- Gulati, S., Sousa, S. & Lamas, D. (2018). *Modelling Trust in Human-Like Technologies*. In *Proceedings of the 9th Indian Conference on Human Computer Interaction*, 1–10. ACM.
- Gulati, S., Sousa, S., & Lamas, D. (2019). *Design, development and evaluation of a human-computer trust scale*. *Behaviour & Information Technology*, 38(10), 1004–1015. <https://doi.org/10.1080/0144929x.2019.1656779>
- Guszcza, J., Lee, M. A., Ammanath, B., & Kuder, D. (2020, January). *Human values in the loop: Design principles for ethical AI*. (Deloitte Insights) (No. 26). Deloitte Development LLC. [https://www2.deloitte.com/content/dam/insights/us/articles/6452\\_human-values-in-the-loop/DI\\_DR26-Human-values-in-the-loop.pdf](https://www2.deloitte.com/content/dam/insights/us/articles/6452_human-values-in-the-loop/DI_DR26-Human-values-in-the-loop.pdf)
- Gutman, J. (1982). *A Means-End Chain Model Based on Consumer Categorization Processes*. *Journal of Marketing*, 46(2), 60–72. <https://doi.org/10.1177/002224298204600207>
- Hagendorff, T. (2020). *The Ethics of AI Ethics: An Evaluation of Guidelines*. *Minds and Machines* 30. pp. 99–120
- Hagerty, A., & Rubinov, I. (2019). *Global AI Ethics: A Review of the Social Impacts and Ethical Implications of Artificial Intelligence*. 1–27. <https://arxiv.org/abs/1907.07892>
- Hair Jr, J. F., Hult, G. T. M., Ringle, C. & Sarstedt, M. (2016). *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM)*. Los Angeles: Sage Publications.
- Hao, K. (2021, April 13). *Big Tech's guide to talking about AI ethics*. MIT Technology Review. <https://www.technologyreview.com/2021/04/13/1022568/big-tech-ai-ethics-guide>
- Helsinki.fi. (2020). *The Ethics of AI online course urges us to consider what technology should be used for*. *University of Helsinki*. Retrieved: 27 Nov. 2020. <https://www.helsinki.fi/en/news/data-science-news/the-ethics-of-ai-online-course-urges-us-to-consider-what-technology-should-be-used-for>

- Hesketh, P. (2019, August 7). *Ethics Kit in 2019 - Ethics Kit*. Medium.  
<https://medium.com/ethics-kit/ethics-kit-in-2019-ba1bf483663>
- High-Level Expert Group on Artificial Intelligence (AI-HLEG). (2020). *Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment*. European Commission.  
<https://ec.europa.eu/digital-single-market/en/news/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment>
- High-Level Expert Group on Artificial Intelligence (AI-HLEG). (2019). *Ethics Guidelines for Trustworthy AI*. Brussels: European Commission.  
<https://ec.europa.eu/futurium/en/ai-alliance-consultation/>
- Hoffman, R. R., Mueller, S. T., Klein, G. & Litman, J. (2018). *Metrics for Explainable AI: Challenges and Prospects*. cite arxiv:1812.04608.
- ICO - Information Commissioner's Office. (2017). *Big data, artificial intelligence, machine learning and data protection*.  
<https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf>
- IDEO.org. (2015). *The Field Guide to Human-Centered Design* (1st ed.). IDEO.org / Design Kit.
- Institute for Business Ethics. (2018, January). *Business Ethics and Artificial Intelligence*. Business Ethics Briefing (No. 58).  
<https://www.ibe.org.uk/uploads/assets/5f167681-e05f-4fae-ae1bef7699625a0d/ibebriefing58businessethicsandartificialintelligence.pdf>
- International Organization for Standardization. (2019). *Ergonomics of human-system interaction — Part 210: Human-centred design for interactive systems (ISO Standard No. 9241-210)*. Retrieved from  
<https://www.iso.org/obp/ui/#iso:std:iso:9241:-210:ed-2:v1:en>
- Jobin, A., Ienca, M. & Vayena, E. (2019). *Artificial Intelligence: the global landscape of ethics guidelines*. ETH Zürich: Health Ethics & Policy Lab.
- Krafft, T. D., Zweig, K. A., & König, P. D. (2020). *How to regulate algorithmic decision-making: A framework of regulatory requirements for different applications*. Regulation & Governance, 1–18. <https://doi.org/10.1111/rego.12369>
- Lane, G., Angus, A., & Murdoch, A. (2018). *UnBias Fairness Toolkit (Version 1)*. Zenodo.  
<http://doi.org/10.5281/zenodo.2667808>

- Lankton, N., McKnight, D. H., & Tripp, J. (2015). *Technology, Humanness, and Trust: Rethinking Trust in Technology*. Journal of the Association for Information Systems, 16(10), 880–918. <https://doi.org/10.17705/1jais.00411>
- Lawson, B. (1990). *How Designers Think: The Design Process Demystified*. (2nd ed.). Butterworth-Heinemann.
- Leijnen, S., Aldewereld, H., van Belkom, R., Bijvank, R., & Ossewaarde, R. (2020). *An Agile Framework for Trustworthy AI*. 1–4. [https://www.academia.edu/43680746/An\\_Agile\\_Framework\\_for\\_Trustworthy\\_AI](https://www.academia.edu/43680746/An_Agile_Framework_for_Trustworthy_AI)
- Leong, T. W., & Iversen, O. S. (2015). *Values-led Participatory Design as a Pursuit of Meaningful Alternatives*. Proceedings of the Annual Meeting of the Australian Special Interest Group for Computer Human Interaction, 314–323. <https://doi.org/10.1145/2838739.2838784>
- Leslie, D. (2019). *Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector*. The Alan Turing Institute. <https://doi.org/10.5281/zenodo.3240529>
- Lewis, D., Hogan, L., Filip, D., & Wall, P. J. (2020). *Global Challenges in the Standardization of Ethics for Trustworthy AI*. Journal of ICT Standardization, 123–150. <https://doi.org/10.13052/jicts2245-800x.823>
- Li, X. & Hess, T. & Valacich, J. (2008). *Why do we trust new technology? A study of initial trust formation with organizational information systems*. J. Strategic Inf. Sys.. 17. 39-71. <http://doi.org/10.1016/j.jsis.2008.01.001>
- Liao, Q. V., Gruen, D., & Miller, S. (2020). *Questioning the AI: Informing Design Practices for Explainable AI User Experiences*. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 1–15. <https://doi.org/10.1145/3313831.3376590>
- Löwgren, J. & Stolterman, E. (2004). *Thoughtful Interaction Design: A Design Perspective on Information Technology*. 10.7551/mitpress/6814.001.0001.
- Leong, T. & Iversen, O. (2015). *Values-led Participatory Design as a Pursuit of Meaningful Alternatives*. 10.1145/2838739.2838784.
- Lucero, A., Dalsgaard, P., Halskov, K., & Buur, J. (2016). *Designing with Cards*. Collaboration in Creative Design, 75–95. [https://doi.org/10.1007/978-3-319-29155-0\\_5](https://doi.org/10.1007/978-3-319-29155-0_5)

- Manders-Huits, N., & Zimmer, M. (2009). *Values and Pragmatic Action: The Challenges of Introducing Ethical Intelligence in Technical Design Communities*. *International Review of Information Ethics*, 10(02), 37–44.  
<http://fiz1.fh-potsdam.de/volltext/ijie/09123.pdf>
- Martin, B. & Hanington, B. M. (2012). *Universal methods of design: 100 ways to research complex problems, develop innovative ideas, and design effective solutions*. Beverly, MA: Rockport Publishers.
- Martín-Martín, A., Thelwall, M., Orduna-Malea, E., & Delgado López-Cózar, E. (2020). *Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: a multidisciplinary comparison of coverage via citations*. *Scientometrics*, 126(1), 871–906. <https://doi.org/10.1007/s11192-020-03690-4>
- Mayer, R. C., Davis, J. H., & Schoorman, D. F. (1995). *An integrative model of organizational trust*. *The Academy of Management Review*, 20(3): 709–734.
- McKnight, D., Carter, M., Thatcher, J. & Clay, P. (2011). *Trust in a specific technology: An investigation of its components and measures*. *ACM Trans. Management Inf. Syst.*. 2. 12.
- McNamara, A., Smith, J., & Murphy-Hill, E. (2018). *Does ACM's code of ethics change ethical decision making in software development?* Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, 729–733.  
<https://doi.org/10.1145/3236024.3264833>
- Mertens, D. M. (2010). *Research and evaluation in education and psychology: Integrating diversity with quantitative, qualitative, and mixed methods*. (3rd ed.). Thousand Oaks, CA: Sage.
- Microsoft. (2018a, November). *Responsible AI principles from Microsoft*.  
<https://www.microsoft.com/en-us/ai/responsible-ai>
- Microsoft. (2018b, November). *Responsible bots: 10 guidelines for developers of conversational AI*.  
[https://www.microsoft.com/en-us/research/uploads/prod/2018/11/Bot\\_Guidelines\\_Nov\\_2018.pdf](https://www.microsoft.com/en-us/research/uploads/prod/2018/11/Bot_Guidelines_Nov_2018.pdf)
- MIT Sloan Office of Media Relations. (2020, November 19). *'Human-Centered AI': How can the technology industry fight bias in machines and people?* MIT Sloan School of Management.  
<https://mitsloan.mit.edu/experts/human-centered-ai-how-can-technology-industry-fight-bias-machines-and-people>

- Mittelstadt, B. (2019a). *AI Ethics – Too Principled to Fail?* SSRN Electronic Journal, 1–15. <https://doi.org/10.2139/ssrn.3391293>
- Mittelstadt, B. (2019b). *Principles alone cannot guarantee ethical AI*. *Nature Machine Intelligence*, 1(11), 501–507. <https://doi.org/10.1038/s42256-019-0114-4>
- Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). *Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement*. *BMJ*, 339(jul21 1), b2535. <https://doi.org/10.1136/bmj.b2535>
- Montreal AI Ethics Institute (MAIEI), Gupta, A., Ganapini, M., Butalid, R., Fancy, M., Royer, A., Khurana, R., Akif, M., Heath, V., Wright, C., Khan, F. A., Sweidan, M., & Galinkin, E. (2021, January). *The State of AI Ethics, January 2021*. Montreal AI Ethics Institute (MAIEI). <https://montrealetics.ai/wp-content/uploads/2021/01/State-of-AI-Ethics-Report-January-2021.pdf>
- Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2019). *From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices*. *Science and Engineering Ethics*, 26(4), 2141–2168. <https://doi.org/10.1007/s11948-019-00165-5>
- Mucha, H., Mevißen, D., Robert, S., Jacobi, R., Meyer, K., Heusler, W., & Arztmann, D. (2020). *Co-Design Futures for AI and Space: A Workbook Sprint*. Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems, 1–8. <https://doi.org/10.1145/3334480.3375203>
- Nickel, P. J. (2015). *Design for the Value of Trust*. *Handbook of Ethics, Values, and Technological Design*, 551–567. [https://doi.org/10.1007/978-94-007-6970-0\\_21](https://doi.org/10.1007/978-94-007-6970-0_21)
- Nielsen, J. (1994). *Enhancing the explanatory power of usability heuristics*. *Proc. ACM CHI'94 Conf. (Boston, MA, April 24-28)*, 152-158.
- NITI Aayog. (2018, June). *National Strategy for Artificial Intelligence*. [https://www.academia.edu/43115903/NationalStrategy\\_for\\_AI\\_Discussion\\_Paper](https://www.academia.edu/43115903/NationalStrategy_for_AI_Discussion_Paper)
- Oppenheim, A. N. (1992). *Questionnaire design, interviewing, and attitude measurement*. New York, NY: Printer Publishers.
- Pärnpuu, M. (2020). *Designing for values: value elicitation toolkit (Thesis)*. Tallinn University. <https://www.etis.ee/Portal/Mentorships/Display/fl774a45-1cd1-4712-853a-f7a10d3fd315>

- Personal Data Protection Commission Singapore. (2019, January). *A Proposed Model Artificial Intelligence Governance Framework*.  
<https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE4u7Mv>
- Pooley, L. (Director) & Metcalfe, M. (Producer). (2020). *We Need to Talk About A.I* [Motion picture]. GFC Films.
- Raftopoulos, M. (2015). *Playful Card-Based Tools for Gamification Design*. Proceedings of the Annual Meeting of the Australian Special Interest Group for Computer Human Interaction, 109–113. <https://doi.org/10.1145/2838739.2838797>
- Randall, D. & Rouncefield, M. (2013). *Ethnography* In "The Encyclopedia of Human-Computer Interaction, 3rd Ed.". In M. Soegaard & R. F. (e. Dam (eds.).
- Reijers, W., Lewis, D., Levacher, K., Calvo, A., Burburan, A., & Mohri, F. (2017). The Ethics Canvas. The Ethics Canvas. <https://www.ethicscanvas.org/>
- Rességuier, A., & Rodrigues, R. (2020). *AI ethics should not remain toothless! A call to bring back the teeth of ethics*. *Big Data & Society*, 7(2), 1–5.  
<https://doi.org/10.1177/2053951720942541>
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). 'Why Should I Trust You?' Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. <https://doi.org/10.1145/2939672.2939778>
- Riegelsberger, J., Sasse, M. A., & McCarthy, J. D. (2005). *The mechanics of trust: A framework for research and design*. *International Journal of Human-Computer Studies*, 62(3), 381–422. <https://doi.org/10.1016/j.ijhcs.2005.01.001>
- Rittel, H. W. & Webber, M. M. (1973). *Dilemmas in a general theory of planning*. In *Policy Sciences*, 4 (2) pp. 155-169
- Rossi, F. (2018). *Building Trust in Artificial Intelligence*. *Journal of International Affairs*, 72, 127. 127-133.
- Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). *Not so different after all: A cross-discipline view of trust*. *Acad. Manag. Rev.* 23, 3, 393–404.
- Ryan, M. (2020). *In AI We Trust: Ethics, Artificial Intelligence, and Reliability*. *Science and Engineering Ethics*, 26(5), 2749–2767. <https://doi.org/10.1007/s11948-020-00228-y>.
- Ryan, M., & Stahl, B. C. (2020). *Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications*. *Journal of Information, Communication and Ethics in Society*, of, 1–26.  
<https://doi.org/10.1108/jices-12-2019-0138>

- Sanders, E. B. N., Brandt, E., & Binder, T. (2010). *A framework for organizing the tools and techniques of participatory design*. In Proceedings of the 11th biennial participatory design conference (pp. 195-198).
- Sanders, E. B. N., & Stappers, P. J. (2008). *Co-creation and the new landscapes of design*. *CoDesign*, 4(1), 5–18. <https://doi.org/10.1080/15710880701875068>
- Sanders, E. B. N., & Stappers, P. J. (2016). *Convivial Toolbox: Generative Research for the Front End of Design*. BIS Publishers, Amsterdam.
- Sankaran, S., Zhang, C., Gutierrez Lopez, M., & Väänänen, K. (2020). *Respecting Human Autonomy through Human-Centered AI*. Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society, 10–12. <https://doi.org/10.1145/3419249.3420098>
- Schmidt, P., Biessmann, F., & Teubner, T. (2020). *Transparency and trust in artificial intelligence systems*. *Journal of Decision Systems*, 29(4), 260–278. <https://doi.org/10.1080/12460125.2020.1819094>
- Sharp, H., Rogers, Y., & Preece, J. (2007). *Interaction Design: Beyond Human-Computer Interaction* (2nd ed.). Wiley.
- Shneiderman, B. (2020a). *Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-centered AI Systems*. *ACM Trans. Interact. Intell. Syst.* 10, 4, Article 26 (October 2020), pp. 1-31. <https://doi.org/10.1145/3419764>.
- Shneiderman, B. (2020b). *Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy*. *International Journal of Human-Computer Interaction*, 36(6), 495–504. <https://doi.org/10.1080/10447318.2020.1741118>
- Shneiderman, B. (2020c). *Human-Centered Artificial Intelligence: Three Fresh Ideas*. *AIS Transactions on Human-Computer Interaction*, 109–124. <https://doi.org/10.17705/1thci.00131>
- Smart Dubai. (2019). *AI Principles & Ethics*. <https://www.smartdubai.ae/initiatives/ai-principles-ethics>
- Smith, C. J. (2019). *Designing Trustworthy AI: A Human-Machine Teaming Framework to Guide Development*. -, 1–6. <http://arxiv.org/abs/1910.03515>
- Söllner, M., Pavlou, P. A., & Leimeister, J. M. (2013). *Understanding Trust in IT Artifacts A New Conceptual Approach*. *SSRN Electronic Journal*, 1–7. <https://doi.org/10.2139/ssrn.2475382>

- Sony. (2018, September). *Sony Group AI Ethics Guidelines*.  
[https://www.sony.com/en/SonyInfo/csr\\_report/humanrights/AI\\_Engagement\\_within\\_Sony\\_Group.pdf](https://www.sony.com/en/SonyInfo/csr_report/humanrights/AI_Engagement_within_Sony_Group.pdf)
- Sousa, S., Lamas, D., & Dias, P. (2014). *A Model for Human-Computer Trust. Learning and Collaboration Technologies*. Designing and Developing Novel Learning Experiences, 128–137. [https://doi.org/10.1007/978-3-319-07482-5\\_13](https://doi.org/10.1007/978-3-319-07482-5_13)
- SRI International. (2020). *Design for Trust*. <https://dft.sri.com/>
- Stickdorn, M., & Schneider, J. (2012). *This is Service Design Thinking: Basics, Tools, Cases* (1st ed.). BIS Publishers Amsterdam.
- Stone, P., Brooks, R., Brynjolfsson, E., Calo, R., Etzioni, O., Hager, G., Hirschberg, J., Kalyanakrishnan, S., Kamar, E., Kraus, S., Leyton-Brown, K., Parkes, D., Press, W., Press, A., Shah, J., Tambe, M., & Teller, A. (2016). *Artificial Intelligence and Life in 2030*. Stanford University, Stanford, CA. <http://ai100.stanford.edu/2016-report>
- Stouten, E. (2019). *Exploring How Developers Could Include the European Commission's Ethics Guidelines to Strive Toward Trustworthy AI (Thesis)*. Leiden University. <https://mediatechnology.leiden.edu/images/uploads/docs/stouten-eu-guidelines-for-trustworthy-ai-thesis.pdf>
- Straus, J. (2020). *Artificial Intelligence – Challenges and Chances for Europe*. European Review, 29(1), 142–158. <https://doi.org/10.1017/s1062798720001106>
- Sutrop, M. (2019). *Should We Trust Artificial Intelligence?* TRAMES, XXIII(4), 499–522. [https://kirj.ee/public/trames\\_pdf/2019/issue\\_4/Trames-4-2019-499-522.pdf](https://kirj.ee/public/trames_pdf/2019/issue_4/Trames-4-2019-499-522.pdf)
- The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. (2019). *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, First Edition*. IEEE. <https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html>
- The Open Data Institute. (2019, July 3). *The Data Ethics Canvas – The ODI*. <https://theodi.org/article/data-ethics-canvas/>
- Thiebes, S., Lins, S. & Sunyaev, A. (2020). *Trustworthy artificial intelligence*. Electron Markets . <https://doi.org/10.1007/s12525-020-00441-4>.
- Toward human-centric A.I.* (2016). Berkeley Engineering. <https://engineering.berkeley.edu/news/2016/09/toward-human-centric-a-i/>
- Tripp, J., McKnight, H. and Lankton, N. K. (2011). *Degrees of Humanness in Technology: What Type of Trust Matters?*. AMCIS 2011 Proceedings - All Submissions. 149.

- Troy, T. (2014). *Re: What are the implications of using even or odd Likert scales for a research survey?*. Retrieved from:  
[https://www.researchgate.net/post/What\\_are\\_the\\_implications\\_of\\_using\\_even\\_or\\_odd\\_Likert\\_scales\\_for\\_a\\_research\\_survey/52dbf427d5a3f2f1218b458a/citation/download](https://www.researchgate.net/post/What_are_the_implications_of_using_even_or_odd_Likert_scales_for_a_research_survey/52dbf427d5a3f2f1218b458a/citation/download)
- Uga, B. (2019). *Towards Trustworthy AI : A proposed set of design guidelines for understandable, trustworthy and actionable AI (Thesis)*. Uppsala University.  
<http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-385392>
- Umbrello, S., & De Bellis, A. F. (2018). *A Value-Sensitive Design Approach to Intelligent Agents*. *Artificial Intelligence Safety and Security*, 395–409.  
<https://doi.org/10.1201/9781351251389-26>
- United Nations Development Group (UNDG). (2017, November). *Data Privacy, Ethics And Protection Guidance Note On Big Data For Achievement Of The 2030 Agenda*.  
[https://unsdg.un.org/sites/default/files/UNDG\\_BigData\\_final\\_web.pdf](https://unsdg.un.org/sites/default/files/UNDG_BigData_final_web.pdf)
- Unity Technologies. (2018, November 28). *Introducing Unity's Guiding Principles for Ethical AI*. Unity Technologies Blog.  
<https://blogs.unity3d.com/2018/11/28/introducing-unitys-guiding-principles-for-ethical-ai/>
- University of Montreal. (2017). *Montreal declaration for a responsible development of artificial Intelligence*. [www.montrealdeclaration-responsibleai.com/the-declaration](http://www.montrealdeclaration-responsibleai.com/the-declaration)
- Van den Hoven, J., Vermaas, P. E., & van de Poel, I. (2015). *Design for Values: An Introduction*. *Handbook of Ethics, Values, and Technological Design*, 1–7.  
[https://doi.org/10.1007/978-94-007-6970-0\\_40](https://doi.org/10.1007/978-94-007-6970-0_40)
- Van de Poel, I. (2020). *Embedding Values in Artificial Intelligence (AI) Systems*. *Minds and Machines*, 30(3), 385–409. <https://doi.org/10.1007/s11023-020-09537-4>
- Van Wynsberghe, A., & Robbins, S. (2013). *Ethicist as Designer: a pragmatic approach to ethics in the lab*. *Science and engineering ethics*, 20(4), 947-961.  
<https://doi.org/10.1007/s11948-013-9498-4>.
- Vanden Abeele, V., Zaman, B., & De Grooff, D. (2011). *User eXperience Laddering with preschoolers: unveiling attributes and benefits of cuddly toy interfaces*. *Personal and Ubiquitous Computing*, 16(4), 451–465. <https://doi.org/10.1007/s00779-011-0408-y>
- Vermaas, P. E., Tan, Y.-H., van den Hoven, J., Burgemeestre, B., & Hulstijn, J. (2010). *Designing for Trust: A Case of Value-Sensitive Design*. *Knowledge, Technology & Policy*, 23(3–4), 491–505. <https://doi.org/10.1007/s12130-010-9130-8>

- Wickramasinghe, C. S., Marino, D. L., Grandio, J., & Manic, M. (2020). *Trustworthy AI Development Guidelines for Human System Interaction*. 2020 13th International Conference on Human System Interaction (HSI), 130–136. <https://doi.org/10.1109/hsi49210.2020.9142644>
- Wolff, S., Auernhammer, J., Schockenhoff, F., Angerer, C., & Wittmann, M. (2020). *Mobility Box: A Design Research Methodology To Examine People's Needs In Relation To Autonomous Vehicle Designs And Mobility Business Model*. Proceedings of the Design Society: DESIGN Conference, 1, 1185–1194. <https://doi.org/10.1017/dsd.2020.285>
- Xu, W. (2019). *Toward human-centered AI*. Interactions, 26(4), 42–46. <https://doi.org/10.1145/3328485>
- Zimmerman, J., Forlizzi, J., & Evenson, S. (2007). *Research through design as a method for interaction design research in HCI*. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 493–502. <https://doi.org/10.1145/1240624.1240704>.

## 7. APPENDICES

### 7.1. APPENDIX 1: EXISTING TOOLS

**Table 6:** Identified existing tools, methods and frameworks

#	TOOL	ORIGINATOR	AREA	TYPE
1	<a href="#">AI &amp; Ethics: Collaborative Activities for Designers</a>	IDEO	AI	Cards
2	<a href="#">Google AI Principles</a>	Google	AI	Principles
3	<a href="#">People and AI Research (PAIR)</a>	Google	AI	Instructions, Worksheets & Tools
4	<a href="#">Trusting AI</a>	IBM	AI	Toolkits & Factsheets
5	<a href="#">Achieving Trustworthy AI</a>	KPMG	AI	Conceptual Model
6	<a href="#">Assessment List for Trustworthy Artificial Intelligence (ALTAD)</a>	European Commission	AI	Checklist
7	<a href="#">AI4People's Ethical Framework</a>	Floridi et al. (2019)	AI	Framework
8	<a href="#">Humans in AI Trello Board</a>	Butnaru et al. (2018)	AI	Resource collection
9	<a href="#">Risk Matrix</a>	Kraft & Zweig (2019)	AI	Framework
10	<a href="#">Judgement Call</a>	Ballard et al. (2019)	AI	Cards
11	<a href="#">LIME</a>	Ribeiro et al. (2016)	AI	Explanation generator
12	<a href="#">Principles for Accountable Algorithms and a Social Impact Statement for Algorithms</a>	FAT/ML	AI	Principles Guiding questions
13	<a href="#">Wizard of Oz</a>	Wolff et al. (2020)	AI	Design research method
14	<a href="#">Trustworthy AI Factsheet</a>	Smith (2019)	AI	Checklist
15	<a href="#">Data Ethics Canvas</a>	Open Data Institute (2019)	Data Ethics	Worksheet
16	<a href="#">Design for Trust</a>	SRI International (2020)	AI	Instructions Worksheets

17	<a href="#">Ethical OS</a>	Institute for the Future	Ethical Design	Toolkit
18	<a href="#">Ethically Aligned Design</a>	IEEE	Ethical Design	Guidelines
19	<a href="#">TTC Labs Toolkit</a>	TTC Lab	Ethical Design	Toolset
20	<a href="#">Ethics for Designers</a>	Jet Gipson	Ethical Design	Tool collection
21	<a href="#">IDEO Method Cards</a>	IDEO	Ethical Design	Cards
22	<a href="#">Design Ethically</a>	Zhou (2021)	Ethical Design	Framework & Toolkit
23	<a href="#">Ethical Toolkit for Engineering/Design Practice</a>	Vallor et al. (2018)	Ethical Design	Toolkit
24	<a href="#">Black Mirror, Light Mirror</a>	Fieseler (2018)	Ethical Design	Speculative design method
25	<a href="#">Design with Intent</a>	Lockton (2010)	Ethical Design	Cards Worksheets
26	<a href="#">Toolkit: Design For Trust</a>	Catalyst Fund	Ethical Design	Toolkit
27	<a href="#">Ethics Kit</a>	Hesketh (2019)	Ethical Design	Toolkit
28	<a href="#">UnBias Fairness Toolkit</a>	Lane et al. (2018)	Ethical Design	Toolkit
29	<a href="#">Ethics Canvas</a>	Reijers et al. (2017)	Ethical Design	Canvas
30	<a href="#">Privacy by design: essential for organisational accountability and strong business practices.</a>	Cavoukin et al. (2012)	Privacy	Impact assessment tool
31	<a href="#">A systematic methodology for privacy impact assessments: A design science approach</a>	Oetzel & Spiekermann (2014)	Privacy	Impact assessment tool
32	<a href="#">Consequence scanning</a>	doteveryone (2019)	Value Alignment	Agile ceremony
33	<a href="#">Value-sensitive design toolkit</a>	Friedman et al. (2019)	Value-sensitive design	Framework & Tools
34	<a href="#">Ai System Ethics Self-assessment Tool</a>	Smart Dubai (2019)	AI	Online self-assessment

## 7.2. APPENDIX 2: COLLECTIONS OF ETHICAL AI PRINCIPLES

**Table 7:** Collections of ethical AI principles

<b>RYAN &amp; STAHL (2020)</b>	<b>FJELD ET AL. (2020)</b>	<b>MORLEY (2019)</b>
Explicability	Verifiability & Replicability	Traceability
Interpretability		Interpretability
Consistency	Predictability	
Equality	Equality	
Accessibility	Right to rectification	
Redress		Redress
Non-maleficence		
Security	Security	Resilience to attack and security
Benefits		
Well-being	Human Values and Human Flourishing	Social impact
Social Good	Leveraged to benefit society	Society and democracy
Freedom		
Self-determination		
	Evaluation and Auditing requirement	Auditability
Transparency	Transparency	
Explainability	Explainability	Explainability
Understandability		
Communication / Autonomy		
Showing	Notification when interacting with an AI	
	Notification when AI makes decision about individual	
Fairness	Fairness	
Inclusion	Inclusiveness in Design / Inclusiveness in Impact	Accessibility and universal design
Access And Distribution	Access to Technology	
Non-discrimination	Non-discrimination and the prevention of bias	
Non-bias	<i>Non-discrimination and the prevention of bias</i>	Avoidance of unfair bias
Diversity		
Plurality		
Reversibility		
Harm	Impact Assessment	
Precaution		
Non-subversion		
Responsibility	Responsible Design	
Privacy	Privacy	Privacy and Data Protection

Beneficence		
Common Good		Justification
Consent	Consent	
	Ability to Opt out of Automated Decisions / Human Review of Automated Decisions	Human Oversight
Trustworthiness		
Dignity		Protection of fundamental rights
Solidarity		
	Multi-stakeholder Collaboration	Stakeholder participation
	Human Control of Technology	Human Agency
	Consideration of Long Term Effects	Minimisation & reporting of negative impacts
Disclosure	Regular Reporting Requirement	
Justice		
Equity		
	Right to Erasure	
	Right to information	
Remedy	Remedy for Automated Decision	
Challenge		
	Ability to Appeal	
Safety	Safety and Reliability	Fallback plan and general safety
	Safety and Reliability	Reliability and Reproducibility
	Security by Design	
Protection	Creation of a Monitoring Body	
Prevention		
Integrity		
Accountability	Accountability	
Liability	Liability and Legal Responsibility	
Acting With Integrity	Scientific integrity	
	Privacy by design	
Personal Or Private Information	Control over data use	
Peace		
Choice	Ability to Restrict Processing	
Liberty		
Empowerment		
Sustainability		Sustainable & environmentally friendly AI
Environment (Nature)	Environmental Responsibility	<i>Sustainable &amp; environmentally friendly AI</i>
Energy		

Resources (Energy)		
Social Security		
Cohesion		
	Accuracy	Accuracy
	Representative and High Quality Data	Quality and integrity of the data
		Trade-Offs
	Recommendation for Data Protection Laws	
	Open government procurement	
	Recommendation for new regulation	
	Open Source Data and Algorithms	

### 7.3. APPENDIX 3: EXEMPLARY CARD FORMAT OF THE TOOLKIT

**Table 8:** Exemplary card format of the toolkit for the design of trustworthy AI

<p>Autonomy</p> <hr/> <p><b>DESCRIPTION</b></p> <p>AI organisations should ensure that end users are informed, not deceived or manipulated by AI and should be allowed to exercise their autonomy.</p> <p><b>TOOLS &amp; METHODS</b></p> <ul style="list-style-type: none"> <li>• AI &amp; Ethics Cards [1]</li> </ul> <hr/> <p>Autonomy</p>	<p>Consent</p> <hr/> <p><b>DESCRIPTION</b></p> <p>The use of personal data must be clearly articulated and agreed upon before its use.</p> <p><b>TOOLS &amp; METHODS</b></p> <ul style="list-style-type: none"> <li>• Ethics Kit [27]</li> </ul> <hr/> <p>Consent</p>
<p>Dignity</p> <hr/> <p><b>DESCRIPTION</b></p> <p>AI should be developed and used in a way that respects, serves and protects humans physical and mental integrity, personal and cultural sense of identity, and satisfaction of their essential needs.</p> <p><b>TOOLS &amp; METHODS</b></p> <ul style="list-style-type: none"> <li>• Ethnographic Research</li> <li>• Expanding The Ethical Circle [23]</li> </ul> <hr/> <p>Dignity</p>	<p>Explainability</p> <hr/> <p><b>DESCRIPTION</b></p> <p>The translation of technical concepts and decision outputs into intelligible, comprehensible formats suitable for evaluation.</p> <p><b>TOOLS &amp; METHODS</b></p> <ul style="list-style-type: none"> <li>• Google PAIR [3]</li> <li>• LIME [11]</li> </ul> <hr/> <p>Explainability</p>

## Fairness

---

### DESCRIPTION

There should be steps in place to ensure that data being used by AI is not unfair, or contains errors and inaccuracies, that will corrupt the response and decisions taken by the AI.

### TOOLS & METHODS

- Trustworthy AI Factsheet [14]
- Moral Value Map [20]
- UnBias Fairness Toolkit [28]

---

Fairness

## Human Oversight

---

### DESCRIPTION

The “ability to opt out of automated decision” principle is defined, as affording individuals the opportunity and choice not to be subject to AI systems where they are implemented.

### TOOLS & METHODS

- Google PAIR [3]
- Assessment List for Trustworthy Artificial Intelligence [6]
- Trustworthy AI Factsheet [14]

---

Human Oversight

## Impact Assessment

---

### DESCRIPTION

The objectives and expected impact of AI must be assessed, reviewed and documented on an ongoing basis

### TOOLS & METHODS

- Trustworthy AI Factsheet [14]
- Data Ethics Canvas [15]
- Layers of Effect [22]
- Black Mirror / White Mirror [24]
- Consequence scanning [32]
- Envisioning Cards [33]

---

Impact Assessment

## Inclusion

---

### DESCRIPTION

Attention should be given to under-represented and vulnerable groups and communities, such as those with disabilities, ethnic minorities, children and those in the developing world. Data that is being used should be representative of the target population and should be as inclusive as possible

### TOOLS & METHODS

- Judgement Call [10]
- Data Ethics Canvas [15]
- Expanding The Ethical Circle [23]

---

Inclusion

## Non-Bias

---

### DESCRIPTION

Developers should examine unfair biases at every stage of the development process and should eliminate those found

### TOOLS & METHODS

- AI & Ethics Cards [1]
- Data Ethics Canvas [15]

---

Non-Bias

## Non-Discrimination

---

### DESCRIPTION

Developers should examine unfair biases at every stage of the development process and should eliminate those found

### TOOLS & METHODS

- Data Ethics Canvas [15]
- UnBias Fairness Toolkit [28]

---

Non-Discrimination

## Privacy

---

### DESCRIPTION

Users should have control and access to data stored about them.

### TOOLS & METHODS

- AI & Ethics Cards [1]
- Google PAIR [3]
- Judgement Call [10]
- Ethics Canvas [29]

---

Privacy

## Purpose

---

### DESCRIPTION

The purpose for building the system must be clear and linked to a clear benefit —system's should not be built for the sake of it.

### TOOLS & METHODS

- Google PAIR [3]
- Data Ethics Canvas [15]
- Design for Trust [16]

---

Purpose

## Reversibility

---

### DESCRIPTION

It is important to clearly articulate if the outcomes of AI decisions are reversible. The ability to undo the last action or a sequence of actions allows users to undo undesired actions and get back to the 'good' stage of their work.

### TOOLS & METHODS

- Trustworthy AI Factsheet [14]

---

Reversibility

## Showing

---

### DESCRIPTION

It should be clear to the end user that they are interacting with an AI system, rather than a human. Further, where an AI has been employed, the person to whom it was subject should know

### TOOLS & METHODS

- Ethically Aligned Design [18]
- Design with Intent [25]

---

Showing

## Transparency

---

### DESCRIPTION

The principle of "transparency" is the assertion that AI systems should be designed and implemented in such a way that oversight of their operations are possible

### TOOLS & METHODS

- Judgement Call [10]
- Data Ethics Canvas [15]

---

Transparency

## Transparency

---

### DESCRIPTION

The principle of "transparency" is the assertion that AI systems should be designed and implemented in such a way that oversight of their operations are possible

### TOOLS & METHODS

- Judgement Call [10]
- Data Ethics Canvas [15]

---

Transparency