Aspects to Responsible Artificial Intelligence- Ethics of Artificial Intelligence and Ethical Guidelines in SHAPES Project

MINNA NEVANPERÄ

2021 Laurea

LAU
REA

Laurea-ammattikorkeakoulu

Aspects to Responsible Artificial Intelligence- Ethics of Artificial Intelligence and Ethical Guidelines in SHAPES Project

Minna Nevanperä
Innovative Digital Services of the Future

Master’s thesis
October, 2021

Laurea-ammattikorkeakoulu **Tiivistelmä**

Tulevaisuuden innovatiiviset digitaaliset palvelut

Tradenomi (YAMK) Tietojenkäsittely

Minna Nevanperä

Näkökulmia vastuulliseen tekoälyyn- Tekoälyn etiikka ja eettiset ohjeistukset SHAPES-hankkeessa

Vuosi 2021 Sivumäärä 71+13 liitesivua

Tämän tutkimuksen tarkoituksena on tutkia näkemyksiä ja lähestymistapoja tekoälyn etiikkaan ja löytää olennaisimmat erityispiirteet tekoälyn kehittämisessä SHAPES-projektille. Tavoitteena on tarjota kehittäjille tarvittavia työkaluja ja ohjeita heidän eettiseen päätöksentekoonsa ja toimintaansa sekä herättämään keskustelua kiistanalaisimmista asioista, jotka liittyvät tekoälyn kehittämiseen ja käyttöön. Tämä tutkimus on osa SHAPES-hanketta (Smart and Healthy Aging through People Engaging in Supportive Systems), joka on H2020-innovaatiotoimintaprojekti (sopimusnumero 857159). Hankkeen tavoitteena on rakentaa ratkaisuja, joilla voidaan helpottaa vanhusten asumista kotona, kuten robotit, älyvaatteet, anturiteknologiat.

Tutkimuksen menetelmänä on Alan Hevnerin Design Science Research. Teoreettinen viitekehys on tehty kirjallisuuskatsauksena, joka sisältää olennaisimmat eettiset teoriat, tutkimustietoa tekoälyn etiikasta, kone-etiikasta sekä ihmisoikeuksia koskevat tutkimuksista. Työ sisältää Hevnerin menetelmän mukaisesti myös ympäristön, johon eettiset ohjeet liittyvät. Tämä käsittää ikääntyvät ihmiset ja SHAPES-ekosysteemin. Tässä tutkimuksessa SHAPESin tekoälyn kehittäjille suunnatut eettiset ohjeistukset suunniteltiin tutkimalla jo olemassa olevia eettisiä ohjeistuksia ja vertaamalla niitä SHAPESin erityispiirteisiin. SHAPESin eettiset ohjeet ovat seuraavista teemoista; vastuullisuus, avoimuus ja selitettävyys, monimuotoisuus, osallisuus ja oikeudenmukaisuus, turvallisuus ja yhteiskunnallinen hyvinvointi ja inhimillisyys.

Asiasanat: Tekoälyn etiikka, eettinen ohjeistus, Design Science Research, SHAPES

Laurea University of Applied Sciences **Abstract**
Innovative Digital Services of the Future
Master of Business Administration
Minna Nevaperä

Aspects to Responsible Artificial Intelligence- AI Ethics and Ethical Guidelines in SHAPES Project

Year 2021 Pages 71+13 appendix pages

The purpose of this study is to examine different views and approaches to the ethics of artificial intelligence (AI) and to find the most relevant and puzzling issues in the development of artificial intelligence for the SHAPES project. The object is to help in providing necessary tools and guidelines to developers for their ethical consideration and action, as well as raise discussion of the most controversial matters related to the development and use of artificial intelligence. This study targets the SHAPES project (Smart and Healthy Aging through People Engaging in Supportive Systems), which is the H2020 Innovation Action project (grant agreement No. 857159). The aim of the project is to build solutions that can make it easier for the elderly to live at home, such as robots, smart clothes, sensor technologies.

The method of this study is Alan Hevner's Design Science Research. The theoretical background is conducted in a form of literature overview, which contains the most relevant ethical theories and research on AI ethics, machine ethics and human rights. The study also contains in accordance to Hevner's method the environment in which the ethical guidelines are related to. This includes ageing people and the SHAPES ecosystem. In this study, the ethical guidelines for the SHAPES AI developers were designed by examining existing guidelines and compare them to special features of SHAPES. The SHAPES guidelines include the following themes; accountability, transparency and explainability, diversity, inclusion and fairness, safety and security and societal wellbeing and humanity.

Keywords: AI Ethics, ethical guidelines, Design Science Research, SHAPES

Contents

1 Introduction

The core of artificial intelligence is the prediction of probable future events and making decisions based on the data available. In some cases AI can help decision-makers to make better decisions and in other cases AI comes to decisions by itself without human interference. A major obstacle to wider use of artificial intelligence is not so much the technical side, as it is evolving at a tremendous pace and becoming increasingly applicaple. The bottleneck that we see is adapting the human thinking and behaviour to the new way of working with the machines. Artificial intelligence has great potential to change the world for the better, but it also becomes with the possibility of great destruction. This is the reason why we need carefully to discuss the ethical use of artificial intelligence and set the common guidelines both in development and in use of AI systems.

In this study, the working mechanics or technologies that form artificial intelligence are not studied in detail. The study focuses more on the effects and ethical dilemmas that use of artificial intelligence is bringing into our lives.

This study targets the SHAPES project (Smart and Healthy Aging through People Engaging in Supportive Systems), which is a H2020 Innovation Action project. The aim of the project is to enable new types of operating models and markets through an open ecosystem. The aim is to develop digital solutions for older individuals who are in some way impaired or have illnesses that make their lives difficult. The aim of the project is to build solutions that can make it easier for the ageing people to live at home, such as, robots, wearables, sensor technologies. The purpose of an artificial intelligence-based ecosystem is to collect and analyze information on the needs of older people and to use this information to produce individual solutions to perceived aging-related problems. Technological or social analysis alone is not enough, but we also need to take into account the views of the target group, such as how artificial intelligence systems can effect on good ageing and whether it can replace human care or reduce exclusion or loneliness, for example. Perspectives can also be contradictory. What might be effective and desirable for the society may not be desirable for the individual.

The purpose of this study is to examine different views and approaches to ethics for artificial intelligence and to find the most relevant and puzzling issues for the SHAPES-project. The object is to provide necessary tools and guidelines to developers for their ethical consideration and action as well as raise discussion on the most controversial matters releated to development and use of artificial intelligence.

The method of this study is Alan Hevner's Design Science Research which contains three separate components; Knowledge base that provides theoretical background, Environment that defines people, systems and organizations that are relevant for the project and Design Science which introduces the artifact and design of the output. The decisions was taken that structure of this study does not follow the usual form of the master's thesis, but to follow the composition of the Hevner's Research Design Method components. All the elements of the thesis structure are still present in this study. Theoretical background is conducted in a form of literature overview which contains the most relevant ethical theories and research on AI ethics.

This study is part of the European Commission Horizon funded SHAPES (Smart and Healthy Ageing through People Engaging in Supportive Systems) project. The aim of this study is to produce ethical guidelines for the developers of the AI systems. At the same time the purpose is to examine how European Commission's guidelines for trustworthy AI are relevant for this project and what might be missing from these guidelines that might be substantial. Aim of this study is to bring into project knowledge that is essential for planning, development and implementation of the AI systems. This is obtained by bringing together essential views of common ethical theories, research on machine ethics and guidelines of AI ethics. It is also important to examine legistlative viewpoints that frame the possible solutions and guidelines. I belive it is especially important to examine those ethical concerns that are not regulated by the law and do not have any convinient technical solutions to promote ethical behaviour of AI.

Abbreviations used in the study:

AI	Artificial Intelligence
AI Ethics	Ethics of Artificial Intelligence
SHAPES	Smart and Healthy Aging through People Engaging in Supportive Systems-project
DSR	Design Science Research

2 Research Design of this Study

One of the research problems is to review what the ethical artificial intelligence is and how to promote the development of responsible AI in general. Another research problem of this study is how to promote ethical development and design of the artificial intelligence systems in SHAPES project and how to promote ethical competence of the developers. The aim is to provide information on what kind of discussion there is around ethical issues around artificial intelligence and on what kind of solutions there is to solve these issues.

The research material of this study is twofold, the theoretical literature of the artificial intelligence and ethics and especially ethical guidelines from the different organisations and the webinars on AI ethics. One of the most important data that was analysed is the European Commission High Level Expent Group on Artificial Intelligence's Ethics guidelines for trustworthy AI. This is because many guidelines of the companies or organisations refer to these European Guidelines and also because of the fact that the SHAPES project is Europeanwide. Mika Nieminen has stated in his webinar presentation, there are more than hundred AI ethics guidelines available. The quality of the guidelines vary enormously. The guidelines that were chosen for this study are from the organisations that are large, known and have big impact on users lives. One criteria of choosing these guidelines was that guidelines should be easily available to everyone to analyse. Another material that was used as a material for this study were the AI ethics webinars. Three webinars were attended.

The method of analysis of the research material gathered is data-driven. Soon after beginning of study, it was clear that there was not much theoretical research on AI ethics itself available. Since the purpose of this study was to create concrete ethical guidelines, the methodology of constructive study was the best practice to follow. The methodology of analysis in discussed more deeply in section on Design Science.

3 Writing Articles and Attending a Conference as part of the Study Process

In addition to the study of AI ethics and designing ethical guidelines for the SHAPES project, I participated in a process of writing articles on the subject and attended one international conference to present our work.

The first writing work I participated was as a co-writer on an article “Privacy and data protection in Open Source Intelligence and Big Data Analytics: Case ‘MARISA’”. It was published in Laurea publication Ethics as a resource Examples of RDI projects and educational development. (Rajamäki, Sarlio-Siintola, Alapuranen & Nevanperä, 2020, 23-29.)

The conference paper for the 25th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems was written together with Jaakko Helin and Jyri Rajamäki and it will be published in Elsevier's Procedia Computer Science soon. The name of the article was “Design Science Research and Designing Ethical Guidelines for the SHAPES AI Developers” and I was virtually presenting it in invited session in the conference in question in September 2021. (Nevanperä, Helin & Rajamäki, 2021a.) The article abstract is provided in Appendix 2.

Yet unpublished article “Comparison of European Commission’s Ethical Guidelines for AI to Other Organizational Ethical Guidelines”. was written also together with Jaakko Helin and Jyri Rajamäki about comparison of the variety of ethical guidelines. This article will be presented in the 3rd European Conference on the Impact of Artificial Intelligence and Robotics by Jaakko Helin. (Nevanperä, Helin & Rajamäki, 2021b.)

4 Methodology: Constructive Study and Design Science Research

4.1 Constructive Study Approach

When the aim of the study is to create a concrete artifact, that is for example a plan, a model or a product, the constructive study is a good option. In short, the aim of the constructive study is to build a new kind of substance based on research data. This is a practical problem-solving approach which combines theoretical knowledge on the subject to the new empirical data. The aim is to find a new and theoretically justified solution to a practical problem, which also brings new information. Constructive research is about design, conceptual modeling, implementation, and testing. It is also essential to tie the solution into existing theoretical knowledge. (Ojasalo, Moilanen & Ritalahti, 2015, 65-66; Kasanen, Lukka & Siitonen, 1993.)

When designing AI ethics guidelines or instructions for the SHAPES project, both theory of ethics, especially AI ethics, and the special features of the project need to be taken into consideration. It was not an easy task to decide how to approach the matter methodologically in sufficient extent since content analysis of the European Commission’s guidelines seemed not to be enough. Considering this, the Design Science Research process was adding to the content analysis the design process, review of the environment and the theoretical background. (Hevner et al., 2004; Hevner & Chatterjee, 2010;

4.2 Design Science Research and Hevner’s Design Science Research Cycles

The basis of the Design Science Research (DSR) is in information technology and information system science. The Design Science Research is a methodological paradigm that emphasizes designer’s role as a creator of innovative artifacts and that way designer contributes new knowledge to scientific evidence by that artifact. The artifacts designed are fundamental part of understanding the problem. Technology is seen as means to practical purposes, not as an end itself. Technologies are developed in response to the specific problems or tasks and are developed based on certain practical knowledge and requirements. (Hevner & Chatterjee, 2010, 5.)

Information systems is a discipline that studies how IT interacts with organizations and how it is managed. The IS paradigm draws from two disciplines, behavioural sciences and design sciences. Behavioural sciences are using methodology familiar from natural sciences where the research process starts with the hypothesis and the studying process ends up either prove or disprove the hypothesis. The theory develops in time. On the other hand, design science is a paradigm based on practical problem-solving. End goal is always an artifact which must be built and evaluated. The knowledge is generated by how the artifact can be improved for the certain purpose that it is trying to solve. But the designing process of the artifact is not free from the theory. It will rely on various theories of design process, even though some argue that the theories of design are vague and more based on practical advice than theoretical matters. (Hevner & Chatterjee, 2010, 5-6.)

Information systems are implemented in the organization for some specific purpose, often they are used to improve the effectiveness and effiency of the organization. Not only information systems itself characterize how the purpose is achieved, but also the people, characteristics of the organization, how the work is done etc. influence on the achievement of the purpose. The design science research seeks to find through analysis and design innovations solutions to the problems that are complex. It values ideas, practices and technical capabilities that form the heart of the artifact. (Hevner & Chatterjee, 2010, 10-11.)

Hevner has identified three Design Research Science cycles which are relevant when positioning design project into the wider context as shown in Figure 1. The Relevance cycle brings in the context that the design research process. The cycle is also interested in bringing something back to environment during the designing process. Usually this is achived by bringing in innovative artifacts that improve the environment. The Relevance Cycle not only define research process context, but also gives the criteria for acceptance of the research results and evaluation. (Hevner & Chatterjee, 2010, 16-17.)

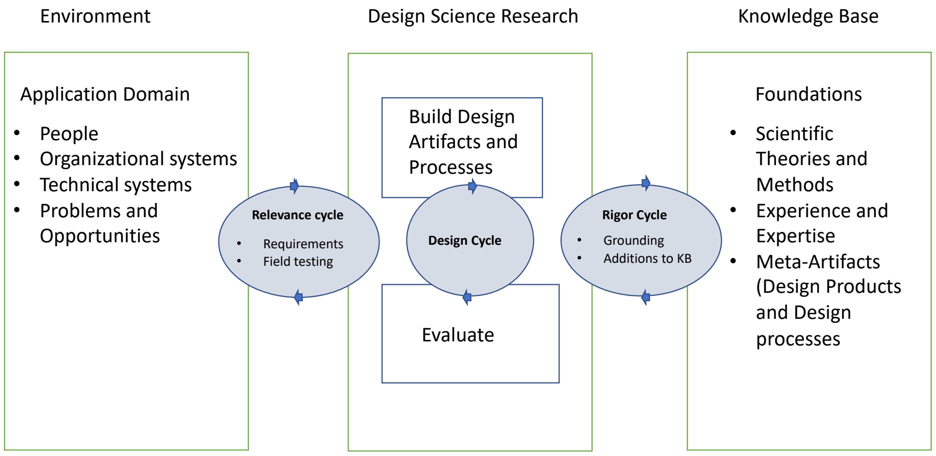


Figure 1. Hevner's Design Science Research (Hevner & Chatterjee, 2010)

Rigor Cycle brings in the knowledge base and theorethical background into the designing process. Researcher needs to make sure that existing theoretical knowledge is taken into consideration in design process to ensure that the designs produced are research contributions, not only routine designs based on known design artifacts and processes. (Hevner & Chatterjee, 2010, 16-18.)

The design cycle is to iterate between the design activities and evaluation of the artifact and the theorethical background and processes of the research. This can be seen as a heart of the design research process. Even though the design cycle draw from the other two cycles, it is important to understand that it is not dependent on the other cycles. (Hevner & Chatterjee, 2010, 16-18.)

In Information Systems research, the artifacts are divided into categories. The artifacts are defined as constructs (vocabulary and symbols), models (abstractions and representations), methods (algorithms and and practices), instantiations (implemented and prototyped systems). These are seen as concrete descriptions for researchers and practioners to understand and address the problems. (Hevner et al., 2004, 77.) The role of the design science has been described to be aiming to describe effective development processes and system solution for specific user requirements. In SHAPES project, the construction of artifact could be described to be ethical guidelines (method) in SHAPES ecosystem (instantiation) for the ageing and other stakeholders.

It has been also argued that the theories of IS can be devided into five classes: theory for analysis, theory for explaining, theory for predicting, theory for explaining and theory for design and action. Theories of design and action are separated for the other classes of IS theories by its nature of practical approach. It is focused on "how to do something" instead of increasing theorethical knowledge. (Gregor & Jones, 2007, 313.) However, Hevner's design science theory is not only keen on this practical side of the design theory, it is also important to link the design of the artifact to the prior knowledge and theoretical background. Hevner's objective is also that the artifact created and the design process gives knowledge back to the knowledge base and theory.

Hevner's Design Science Research is a typical constructive research approach. The constructive research approach is characterized by that it is focused on on real-life problems that are solved by creating a construction (artifact, model, plan, instruction etc.) that is tested in real environment. Costructive approach links the research and design work closely to the existing theoretical knowledge and reflects it back to the theoretical background. (Hevner & Chatterjee, 2010; Hevner et al., 2004.)

.

4.3 Design research theory and AI ethics in SHAPES

In this paper the Hevner's theory of design science research is used as a methodological background for the constructing ethical instructions for the technical developers of the SHAPES project. Designing ethical instructions is considered as Hevner's theory's examination of the artifact. This means that the aim for this research is to find useful methods for designing ethical guidelines for AI projects and to find the best way to give this kind of guidance to developers of the AI systems. It is important to see that DSR consideres that the knowledge and understanding of the design problem and its solution are acquired when building the artifact. According to DSR outputs can be constructs, models, methods or instantiations (Hevner et al. 2007, 77). In this case the artifact can be seen as method since the purpose of this study is to create a practice for ethical guidance. SHAPES ecosystem is an AI solution that collects and analyses data and information from the various sources including the applications provided by the cooperation partners.

The structure of this study is based on the Hevner's model of Design Science Research Cycles. Firstly, the Knowledge Base is introduced, then the Environment and finally the Design Science artifact. Figure 2 will highlight this structure and bring into light the Design Science Research framework of the SHAPES project.

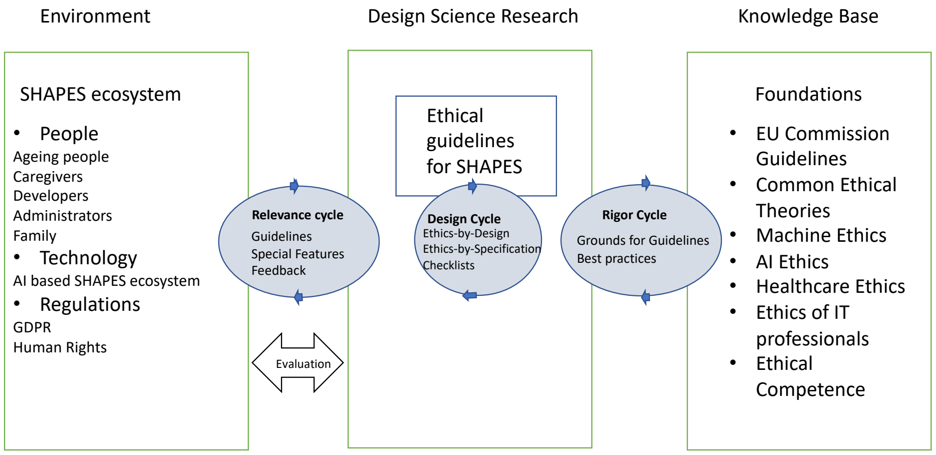


Figure 2. SHAPES project Design Science Research

5 Knowledge Base

5.1 Common ethical theories

Ethics is defined as a rational and systematic analysis of conduct that might do benefit or harm for others. Because the ethics is based on reasoning, people need to explain why they hold the opinion they have. This means that we are able to evaluate and compare ethical evaluations. (Quinn, 2015, 82-83.)

Ethics as a formal study is not a new thing. Study of ethics dates back to Greek philosopher Socrates. Socrates did not leave behind anything written, but his student Plato used his ethical reasoning in his writings. More recently there has been more ethical theories proposed. Some of them will be now examined briefly here. (Quinn, 2015, 82-83.)

Ethics can be roughly be devided into three subfields. Meta-ethics studies the meaning of ethical concepts and existence of ethical thought, normative ethics studies practical means of ethically correct action and morals. Applied ethics is concerned of actions of the moral agent in specific situation. AI ethics is mostly considered as a subfield of applied ethics. (University of Helsinki, 2020.)

Ethical thought can be also devided into three categories according to the time frame that their effects are considered. Immediate effects are things like security, data protection or transparency, intermediate concerns include the use cases like can AI systems be used in military and what kinds of effects the use of AI systems have in health care and education. Long-term ethical concerns are things like what kind of effects implementation of AI systems has in the society and the whole world. (University of Helsinki, 2020.)

5.1.1 Virtue as a framework for ethics

Virtue ethics is possibly the most important development in moral philosophy in late twentieth century (Hursthouse, 2000). Virtue ethics can be traced back to ancient Greece, studies of Aristotle. In his book Nicomachean Aristotle states that the path to true happiness is through the life of virtues. According to Aristotle there is two kinds of virtues, intellectual and moral. Intellectual virtues are associated with reasoning and truth and moral virtues are habits and virtuous actions. Theories of ethics are usually concentrating on moral virtues Virtual ethics is also concentrating on agent, a person who is performing the moral action. A good person does the right thing for moral reasons. (Quinn, 2015, 117.) Rosalind Hursthouse describes virtue ethics as addressing the question "What kind of person should I be?" when the question "what action should I take?" is less relevant. (Hursthouse, 2000, 25.) However, there is discussion on the relationship of virtue and action. For instance, Christine Swanton has studied this relation (Swanton, 2003).

Michael J. Quinn states in his study advandages and disadvantages of virtue ethics. In many situations it is more valuable to concentrate on virtues than on obligations, rights and concequences. This also means that the morality in virtue theory is more personal than in other theories. It recognizes the important role of emotions when people are making moral decisions. Virtue ethics also recognises that the moral decision-making skills develop over time and make the theory more flexible. The moral dilemmas are considered in their context and right action can be different in different situation. There are also some arguements against virtue theory. We do not live in world that is homogenous. The perspectives of what characteristics can be seen as virtues vary. This means that we cannot agree how a virtuous person would do in particular situation. Virtue ethics is concentrating on the actions of the individual and cannot be used as a guideline of the government policy as such since the actions taken are always a decision of the group. (Quinn, 2015, 120-121.)

When regarding virtue as a basis of ethical examination, focus is on good character rather than on rights, duties and consequences. The virtue theory states that the purpose of life is to practise good character in such way that the well-being of the community is maximized. Organizations can achieve this goal by demonstrating virtue internally or on the markets in general. (Neubert and Montanez, 2019, 197.)

5.1.2 Virtues in AI ethics

One might ask that what are the common virtues? Neubert and Montanez stated that the common virtues that are relevant in development of AI are prudence, temperance, justice, courage, faith, hope and love. They also give definitions on each. For example, faith is defined by trust and trusting that the others act in that way that they do not harm intentionally. According to Neubert and Montanez there is evidence that those organizations that use virtues as ethical guideline for designing artificial intelligence can attract and retaining developers of AI. The virtues also affect positively to the reputation of the organization among AI users. (Neubert and Montanez, 2019, 198.) This is not very far from the guidelines that European Commission has been giving for the AI and the effects that are hoped to be achieved with the guidelines. In their study Neubert and Montanez give a virtue-based framework on AI ethics, similar to European Commission guidelines. It also includes the list of questions that developers and deployers of AI should ask when assessing ethics.

According to Neubert and Montanez the virtue behind the prevention of harm is prudence. When developing AI systems that might have long-term effects, the implications to all stakeholders should be considered through the specific decision-making process to be able to tackle the harmful effects. The organization takes action to foresee the dangers and effects that their actions might have. (Neubert and Montanez, 2019, 201.)

Neubert and Montanez approach the same issue in virtue point of view. The virtues of justice and temperance contribute to fairness when considering AI. They suggest that justice should be a measure in all AI development by design to prevent bias. Justice also demands the organization to take responsibility of its own actions whether unforeseen or accidental. Neubert and Montanez give a real-life example of virtue of temperance in use in a context of artificial intelligence. The training of AI might be easy and inexpensive if using interactions with random humans instead of building the AI system via well-defined tasks and training observation. Temperance as virtue is seen here as a safety measure for not to pursue immediate profit, but having reliable safety measures in place. (Neubert and Montanez, 2019, 200-201.)

I believe it is easier to consider the principle of fairness as fairness than as the virtues of justice and temperance. However, it must be said that the theory of virtues is clearly visible behind also the European Commissions guidelines. It is more logical to consider fairness as a basis of developing AI than virtues.

The virtue of love is theological virtue. It seems to be rather out of place when discussing organizations. However, in this context the virtue of love can be considered as another way of saying valuating human life and well-being. In the context of artificial intelligence, the virtue can be for example consideration of what field or how the AI system is used. e.g. military use. (Neubert and Montanez, 2019, 201.)

Neubert and Montanez also consider virtue of hope in a context of artificial intelligence. According to them the developers of AI should consider the thought that does the application that is developed promote hope? If the application is increasing well-being of humans it creates hope that the mankind can overcome the obstacles we have. (Neubert and Montanez, 2019, 201.)

5.1.3 Relativism

Relativism is a common ethical theory that states that there is no universal moral norms, but different individuals or groups might have completely different views on moral issues and both can be right. There are two kinds of relativism. According to subjective relativism each individual decides of its own moral grounds. In short, this means that the ethical debates are disagreeable and pointless when both sides are right according to relativism. Subjective relatism has been critized that it allows people to make decisions by any means that fit to their current state of mind. They may choose their point of view on grounds other than logic and reason. (Quinn, 2015, 84-85.)

Cultural relativism is an ethical theory that states that the meaning of right and wrong are culturally binded to the society they are produced. They also can vary in different times. Cultural relativism also takes into account that different social contexts demand different ethical guidelines. This means that the theory makes possible the idea of change. There is also some criticism against cultural relativism. For example, if societies have different views on moral issues now, it does not mean that it should be the case always. It does not offer any framework that would allow cultural reconciliation in moral conflicts. Even though there are moral practices in the culture, that does not mean that the practices are acceptable. It should not be reasonable to assume that all moral practices are equally legitimate. (Quinn, 2015, 88-89.)

In the current literature the relativism has not been used as a base of the ethical analysis of AI. However, I believe that the especially cultural relativism gives some grounds to concider that all cultures do not think alike on moral and ethical matters. When the features of the culture are different, we might see different solutions on the ethical matters on AI. Considering SHAPES the Eastern or Southern European views on elderly might have a very different perspective from the Nordics. That might effect on what kind of ethical issues are considered the most relevant.

5.1.4 Kantianism

Kantianism is an ethical theory that is named after German philosopher Immanuel Kant. Kant’s theory is based on the belive that the people’s actions are guided by moral laws and that the moral laws are universal. Kant’s theory can be seen as an opposite to cultural relativism theory. Kant states that only thing that is truly good without qualification is good will. Kant asks what makes a moral rule appropriate? His anwer is theoretical structure called Categorical Imperative. To evaluate moral rule, we must universalize it. (Quinn, 2015, 90-91).

5.1.5 Utilitarianism

Utilitarianism is a moral theory which is based on the works of Jeremy Bentham and John Stuart Mill. In short, utilitarianism states that the actions of the moral agent should always aim for the happiness and pleasure and oppose harm and unhappiness. It also applies to society as a whole. In utilitarianism, the decision is right when it promotes happiness for the greatest number of people in the society even though it might produce unhappiness for some. (Quinn, 2015.)

However, the goal of utilitarianism is not easy to achieve. How do we know that the decision made now is a best possible in the future? There is only right or wrong, no betweens according to utilitarianism. Especially, when discussing on artificial intelligence, the risk of causing harm might not depend on only one action, but there might be several actions that affect the result. (University of Helsinki, 2020.)

5.1.6 Social Contract theory

The social contract theory is an ethical theory based on works of the philosopher Thomas Hobbes, who emphasizes that moral rules are the rules that are necessary to civilized society to work. According to Hobbes the social contract is formed when the citizens living in certain society agree on two things; that moral rules are needed to govern relations between the citizens and that there are government capable of enforcing these moral rules. Jean-Jacques Rousseau continued the theoretical reflection of the social contract theory. According to him The critical problem facing the society is to find the balance between guaranteeing everybody’s safety and property and that all the citizens should still remain free. This is achieved by the community to define the rules to its members and to obligate each member of the society to obey these rules. The concept of rights and duties is also closely related to social contract theory. These have close correspondence to each other. This means that if you have a certain right, it obligates the other members of the society to provide it to you. It creates a duty. (Quinn, 2015.)

5.2 The European Commission High Level Expent Group on Artificial Intelligence’s Ethics guidelines for trustworthy AI

In April 2019 European Commission High Level Expert Group on Artificial Intelligence published Ethics Guidelines for Trustworthy AI. The guidelines have partly different perspective on AI ethics that has been seen on the literature. The background is not so much on the ethics theory but on human rights. The group’s aim was to create the ethical framework to reliable and trustworthy artificial intelligence. These guidelines concern all the stakeholders and interest groups that are involved i.e. end users, developers, decision makers etc.

On 19.2.2020 European Commission published a white paper on Artificial Intelligence- A European approach to excellence and trust. In this paper they promoted an ecosystem of trust which should give European citizens confidence to deploy AI applications and to give the organizations and companies the legal certainty to develop AI systems. (European Commission, 2020, 3.)

The guidelines have three elements that should be met throughout the lifecycle of the AI system. First of all, it is required to be legitimate and it has to follow the law and regulations altogether. But on its' report the expert group did not cover legal framework of AI systems. They argue that the legal aspects are more depended on the field that AI is utilized than of AI in general and they see more need for guidelines for two other components. European Commission's guidelines are based on the presumption that AI solutions that are developed are lawful. In addition to primary legislation like fundamental human rights there is also secondary legislation to consider. In case of SHAPES this is for example heath care legislation or legislation specific to elderly. (European Commission, 2019, 1-7.) There are also legal aspects that are to be met also in general level such as GDPR or accessibility.

Secondly, trustworthy AI has to be ethical. It must be ensured that ethics guidelines and values are followed from planning through development to all phases of using the AI system. What does this exactly mean? This will be examined in detail later since this is the core of the ethical guidelines by European Commission. (European Commission, 2019, 1-8.)

Thirdly, systems using AI are required to be technically and socially reliable and trustworthy. The expert group uses the word robust to describe the component. They must not generate harm intentionally or unintentionally. Ideally all these three components are aligned when AI system is estimated, but it should be noticed that in practice there might be contradiction between the components. (European Commission, 2019, 1.)

5.2.1 Ethical framework

European Commission's vision of ethical and safe artificial intelligence is based on increase of public and private investments on AI development to create circumstances that promote deployment of AI, prepare for socio-economical change and ensure ethical and legal framework to be able to strengthen European values. AI is one of the new technologies that will change the society and promote the sustainability and equality and restrain climate change. European Commission requires AI systems to be human-centric and they must promote common good for all mankind. By using the ethical framework, the Commission also desires European manufacturers and providers of AI systems to gain competitive advantage. This requires that the gains of AI are maximized and the harms are minimized and that the public will consider the provider's AI technology to be trustworthy. The commission's aim is to promote increase in use of AI systems by creating trust towards the technical development and deployment of new technologies. The commission's aspiration is to obtain this by trustworthiness. Since AI is a technology particularly global, the vision of this framework is also to work as an example of how the ethical guidelines and procedures can be designed and brought to practical level.

This framework acknowledges that the development and utilization of AI is rather limited to the context. The application that makes suggestions of what movie to see does not constitute the same ethical issues than an application that makes decisions about your health. European Commission's group suggests that this framework might not be enough or it is too high level when considering applications that might decide about complicated issues that might involve ethical issues that are contradictory. (European Commission, 2019, 6.)

5.2.2 Fundamental human rights as groundwork for AI ethics

Respect for human rights provides a good groundwork for employment of ethics for AI as they emphasize basic democratic and ethical principles and values. The EU Treaties and EU Charter describe the rights by reference to dignity, freedoms, equality, solidarity and the rights of citizens and justice. The foundation for all of these can been seen as human-centric view which is based on human dignity. Human dignity is based on the presumption that every human being has an absolute value that should not be diminished, suppressed or endangered by another human being or any technology like AI. When developing AI systems, the human beings should be seen as active moral subjects instead of seeing them as objects of the action AI performs. (European Commission, 2019, 10-12.) The human-centric AI is required to be developed in the way that it is aligned with society and community it affects. It has to be based on the cultural and ethical values that prescribe standards of right and wrong in terms of rights, obligations and fairness.

There are four ethical principles that EU Commission requires all AI systems to have. They are all based on the EU Charter and they are based on the requirement that all AI systems should improve individual and collective well-being without causing any harm.

1. The principle of respect of human autonomy

The AI systems should be developed in such way that they ensure the freedom and self-determination of the individual in all occasions. The systems should not subordinate, manipulate, mislead, constrain or herd humans, but instead they should be designed to empower the good in people and complement the cultural, cognitive and social skills (European Commission, 2019, 13-14). When developing AI systems there should also be a good understanding of user groups involved and the development should take their special features into consideration.

2. The principle of prevention of harm

The AI systems should not harm or deteriorate the harm. This harm might be the safety and health of individuals, including loss of life or damage to property or loss of privacy, limitations to the right of freedom of expression, human dignity or discrimination. The physical and mental integrity of humans should always be protected. The AI systems are required to be technically reliable and safe for all users. Particular attention should be on vulnerable user groups and their special features. These users should be included into development, deployment and use of AI systems. Special attention should be paid to the situations where the AI system might result in harmful asymmetries of power or information. The systems should also be secured from the malicious use and they should be designed, developed and used in sustainably and environmentally friendly fashion. (European Commission 2019, 14.)

3. The principle of fairness

Fairness as a concept is not explicit. The fairness in the development and deployment of AI systems means that both advantages and costs will be equally distributed and that the ensuring that the decisions AI makes are not prejudiced, discriminative or biased. At best AI systems can promote social fairness and build new prospects for equal access to education, services, products and technology. Utilization of AI systems should never mislead the users or stakeholders or impair their freedom of choice. (European Commission, 2019, 14.)

4. The principle of explicability

One essential demand for the public to consider AI trustworthy is explicability. People need to understand the processes and decision-making mechanisms behind the AI resolutions. Without this knowledge it is difficult to contest the decision. As AI systems are remarkably complicated setups, it is not always possible to explain thoroughly why the model used is giving this exact resolution. These "black box" algorithms need special attention. The other measures such as traceability, auditability and transparency should be in place to ensure the explicability in these cases. Also, the system must otherwise respect fundamental human rights. It is important to take into account the context of use since the incorrect or inaccurate information might be in some cases fatal. (European Commission, 2019, 14-15.)

5.2.3 European Commission's seven requirements for trustworthy AI

In European Commission's framework introduces seven requirements for trustworthy artificial intelligence. It is stated that all seven requirements are equally important and they support each other through the whole life cycle of AI system. The requirements should not only be met by developers of AI, but also by people responsible of deployment and end users. The different stakeholders have different responsibilities on these requirements. The developers need to apply and implement these requirements on their designing processes, whereas the role of the deployers is to make sure that the systems that they offer are meeting the requirements at all times. The end users and society as a whole must be aware of the requirements so that they are able to request and monitor the implementation of these requirements. (European Commission, 2019, 15-16.)

The seven requirements presented are human agency and oversight, technical robustness and safety, privacy and data protection, transparency, diversity, non-discrimination and fairness, societal and environmental well-being and accountability.

1.Human agency and oversight

Fundamental rights.AI systems should at all times foster human rights and support human autonomy and authority. When developing and using AI systems fundamental rights can be either enabled or hampered. When there is a risk of human rights to be violated or even risk of harming fundamental rights to be fulfilled, the impact assessment should be undertaken. Evaluation of the risks and how they could be reduced or can the risks be otherwise justified in order to respect rights of the others. Mechanisms that enable external feedback from any violations against human rights. (European Commission, 2019, 18.)

Human agency. The requirement of human agency has two distinct standing points. Firstly, the users should receive all the necessary information and tools so that they are able to understand and interact with AI systems. The system should enable user to evaluate its actions and if necessary, contest or challenge its decisions. A goal for AI system should be to promote individuals towards better decision-making by distributing knowledge and information. Secondly, user autonomy must always be respected. The individual user must have a choice not to be an object of automated decision-making if it has a significant effect or legal consequences on the user's life. (European Commission, 2019, 18.)

Human oversight. Human oversight is significant measure to ensure that AI system does not undermine human autonomy or cause other harmful effects. Oversight is executed by governance mechanisms that embrace human authority. The human-in-the-loop approach means capability of human intervention in every decision cycle of the system. However, this might not be possible or necessary in most AI systems. Human-on-the-loop approach refers to capability of human action in the design cycle of AI and monitoring the operations of AI system. Human-in-command approach enables human to have oversight on AI systems overall activity including economic, social, legal and ethical impacts. This also means ability to decide when and how the system is used. The common rule is that the less there is human oversight over the system, the more testing and stricter governance is required. (European Commission, 2019, 18-19.) Christian Huyck et al. suggest potential solutions for this dilemma of giving artificial intelligence enough of autonomy to work efficiently, but also taking human into decision-making to avoid unethical and harmful decisions. Firstly, keeping human in the loop and involving humans as part of the process to verify the decisions of the AI system. Secondly, they suggest getting gradually human out of the loop. At first the AI system just observes and builds internal models of behavior of the participants of the ecosystem. Gradually the system starts to give suggestions to the occupants and in the end the actions can be delegated to the AI system. (Huyck et al, 2015 28.) This second approach is rather radical. Huyck et al. were studying AI aided medicine management system monitoring medicine intake at home and they considered that need to consult with humans defeats the purpose of artificial intelligence in this case. However, this approach does not mean that the humans should not monitor AI at all. I believe this is could be ideal solution for the systems that only need human-in-command approach, but there are systems that need more supervision by humans.

2.Technical robustness and safety

One of the key elements of creating trustworthy AI is the technical robustness and safety. Without technical reliability and safe-to-use and secure solutions the principle of prevention of harm will not be achieved. Technical reliability requires that the AI systems will be developed avoiding known risks. It is also required that the AI systems are reliable when in use in a way that they behave as expected and planned. They need to be able to minimize unexpected and unintentional harm and to prevent unacceptable harm. The systems should in all occasions secure physical and mental integrity of humans. (European Commission, 2019, 19.) In their study H.J. Toh et al. on telegeriatrics, concluded that one of the main problems in virtual geriatric care was the delays and connection problems of the technology and problems with audio and video quality. However, the participants adapted well to the new technology. (Toh et al, 2015, 99.) I believe that the this is one of the core issues of the AI system development. However, it is not always considered to be an ethical issue in the same sense that for example the European Commission is considering it. In the literature and the case studies it is mostly presented, but not as a moral or ethical issue, but a matter of legal requirements.

Cybersecurity and resilience to attacks. When considering AI systems, there are possibility of new kinds of attacks that are specific only for artificial intelligence. Artificial intelligence uses technologies like shape and pattern recognition in its decision-making. The attackers might focus their attacks specifically against the process specific to AI. As GDPR states, all software systems should be protected in such way that the vulnerabilities cannot be exploited by malicious parties. The developers must be aware that the attacks can be targeted on data (data poisoning), on model (model leakage), on the underlying infrastructure, both hardware and software. It should be acknowledged how the attacks may harm the AI system in question and what kind of impacts the attack might have on the decisions the AI system makes. For example, the system might change its behavior or seize to operate. (European Commission, 2019, 19.) It should not only to consider how to protect the systems against attacks, but also how the system should react or operate when attack happens.

Fallback plan and general safety. All AI systems should have safeguards that enable fallback plans when problems occur. European Commission suggests that this could be managed in two ways. Either the system switches from a statistical to a rule-based procedure or it requires human interaction to be able to continue operations. The level of safety procedures should be based on the risk that the system failure can cause and the application area. (European Commission, 2019, 18-19.) Considering SHAPES which includes both vulnerable target group and confidential and intimate data the fallback plan for AI system should always require human action and verification before continuing operations.

A lack of clear safety guidance of using AI might lead not only to the risks to the individuals, but also uncertainty for the companies and authorities. AI systems might not fall under any current legislation. This might lead to situation where the individual who has suffered harm by AI system might not obtain any compensation. Furthermore, the person who has suffered harm might not have access to the information and evidence that should be essential to build a case in court. (European Commission, 2020, 12.)

Accuracy. The AI system must be able to make accurate predictions, recommendations and decisions based on available data and models. The well-formed and explicit development and evaluation process can contribute to better understanding of unintended risks and diminish risk of inaccurate predictions. When the occasional inaccurate decisions cannot be fully avoided, the AI system should be able to give probability of these errors. When the system is affecting human life, the accuracy should be high-levelled. (European Commission, 2019, 19.)

Reliability and reproducibility. It is important that the results of AI system decision-making are reproducible. This means that everyone with the same knowledge and data should be able to have the same results in the same circumstances. Replication files can give guidance of the process of how to reproduce and test the behaviors. Reliable AI interacts properly with a range of inputs and in different situations. (European Commission, 2019, 20.)

One of the European Commission's requirements is technical robustness and safety. This seems quite oblivious requirement for any IT system. It is one of the key elements to solve before we can even consider other requirements of trustworthiness of the AI. If people do not trust the system to be technically robust and safe to use, they simply do not use the system even though it might bring other benefits to their lives. The important question here is that how should the AI system behave when something goes wrong? Should it have a safety system which shuts it down or should it just give alert to someone?

3.Privacy protection and data governance

Privacy protection. Privacy protection should be fundamental assumption for all AI systems. It must be ensured throughout the lifecycle of the AI system. This must cover all the data user provides for the system and the data created over the course of their interaction. The information gathered of the users, must be handled in such way that it does not cause any harm to the user or the data cannot be used to discriminate or to be used unlawfully in any way. (European Commission, 2019, 20.)

Quality and integrity of data. European Commission highlights that improving access to data is crucial. Without data the development of AI and other digital applications is not possible (European Commission, 2020, 8). Also, the quality of data used is very important for the AI system decision-making. When collecting the data, it might include socially constructed biases, inaccuracies, errors and mistakes. This must be taken into account before using it to train AI. The processes and data must be tested and documented in all stages such as planning, training, testing and deployment. This must also apply to the AI that is created by the third party but acquired elsewhere (European Commission, 2019, 20).

Access to data. It is necessary to create a policy for by whom and under what circumstances the data can be accessed and for what purpose the data is used. Only duly qualified personnel should be able to access the individual’s data. (European Commission, 2019, 20.). It is important to promote the responsible data management to be able to build trust and ensure that the data remains re-usable (European Commission, 2020, 8).

Technical robustness and safety and privacy and data protection are not specific only for the AI systems and solutions, but to all digital and IT systems. When the data used or stored is highly personal and harmful in the hands of the wrong people, the safety measures and features must be on the highest level.

Especially in SHAPES project the high level of data protection and privacy must be applied since the data contains personal information and health data. Also, the technical capabilities of the planned users must be taken into consideration.

4.Transparency

The requirement of transparency contains three main issues, traceability, explainability and communication.

Traceability. To able traceability and transparency on AI system decision-making all data collection, data labelling and algorithms used should be documented carefully. This is essential feature when the AI system decision-making is faulty or questioned. Traceability enables to look back and find the cause of the faulty decision and helps to prevent future mistakes.

Explainability. Explainability means ability to explain the technical processes of AI systems and how the decision-making process works. Technical explainablity means that the decisions made by artificial intelligence must be understandable and traceable by human beings. (European Commission, 2019, 21.) This does not mean that all the people who are users of AI system must understand all technical details or all features of algorithms, but it means that these details must be understandable for some. However, when there is significant impact on people's lives, it is necessary one to be able to demand explanation of AI system decision-making process in such way that is timely and adapted to the level of expertise of the stakeholder. In addition, it should be reported how the use of AI system effects the decision-making process of the organization. Also, there should be reports on the AI system development and deployment processes. This ensures the transparency of the business models. (European Commission, 2019, 21.)

When considering SHAPES, there should be considerations how to make AI systems explainable for the aging. The ageing is not homogenous group considering technical ability. It is fair to suppose that at the moment the age group of 75+ is not mostly highly skilled on the technical matters and their capabilities might be limited due to high age and health issues. In the future, the situation might be different since the ageing are more used to work with technology. AI systems should be designed that way that they rather compensate the disabilities of the ageing than complicate using technology.

Communication. Humans have the right to know that they are interacting with artificial intelligence therefore the AI system should never represent themselves as human. This means that the system must be easily identified as an AI system and there should be possibility to choose to interact with a human so that the fundamental rights can be ensured. In addition, the capacity and capability of AI system should be communicated as well as the limitations and level of accuracy. (European Commission, 2019, 21.)

5.Diversity, non-discrimination and fairness

The principle of fairness is closely combined with diversity and non-discrimination. It means that to be able to create trustworthy artificial intelligence, the inclusion and diversity must be ensured throughout the lifecycle of the AI system. In practice this will include three main requirements, avoidance of unfair bias, accessibility and universal design and stakeholder participation. (European Commission, 2019, 21-22.)

Avoidance of unfair bias. Identifiable and discriminatory bias should be removed in data collection phase when recognized and possible to remove. Unfair bias might also be created in the development of an AI system. Counteraction should be taken to avoid this kind of bias. This is possible to achieve by hiring from diverse backgrounds, cultures and fields of study to ensure diversity of opinions. Likewise, the oversight processes should be in place to analyze the AI system decision-making, purposes, limitations and requirements so that they are developed and used in a transparent manner. (European Commission, 2019, 21.)

Accessibility and universal design. AI systems should be designed user-centric by default. It should be ensured that AI systems are developed in such way that it enables all kinds of users to have access and possibility to use products and services regardless of age, gender, abilities or characteristics. The major attention should be assigned to ensure the accessibility to those individuals with disabilities. The equal access and active participation can be ensured by avoiding one-size-fits-all approach and by using Universal Design principles when developing AI systems. (European Commission, 2019, 22.) It has been indicated that involving older people in the development of ICT and finding ways to build a bridge between the ageing and younger people to enable younger people to assist elderly as users of ICT might increase the technical competencies of the ageing (Gilhooly et al., 2009, 68.)

In SHAPES the AI ecosystem is addressed to ageing, health care professional, caretakers and governance. This means that the variation in user abilities is great. Therefore, the SHAPES ecosystem should be designed in such way that it enables different user groups to fully participate and use services and products on their own capacity level. In the case of ageing the reduced capabilities should be taken into consideration. For example, hearing, eye-sight or fine hand movements might be limited and complicates using technology if there are not special setups for better accessibility. Designing age-based technology should involve the focus groups from the beginning of the design process. If the design only involves younger developers, users, marketing people etc. the abilities and demands of older individuals are not heard. The designing of the products and services for the ageing should be based on user-friendliness. (Gassmann and Keupp, 2009, 89.)

Stakeholder participation. To be able to create trustworthy artificial intelligence it is recommended to include stakeholders that might be affected, directly or indirectly, by the AI system. Furthermore, long term mechanisms to increase participation should be in place. (European Commission, 2019, 22.)

6.Societal and environmental well-being

According to the principle of fairness and avoiding harm the whole society and other sentient beings and environment should be taken into account as stakeholders. Development of artificial intelligence should promote sustainability and responsibility. It should aim to resolve global concerns and ideally to be used to benefit all mankind, also the future generations. (European Commission, 2019, 22.)

Sustainable and environmentally friendly AI. Artificial intelligence might be solution to some of the globally most pressing societal problems, but there are some environmental concerns included to AI. The whole process of developing, deploying and using AI should be examined on the environmental point of view. The use of resources and energy are main concerns related to artificial intelligence. Measures that ensure environmental friendliness should be encouraged. (European Commission, 2019, 22.)

Social impacts. A contiguous exposure to AI systems might change our perception of social agency and have an impact on our relationships and attachments. Although AI systems might enhance social skills there might be just an opposite effect. The social impacts include affects on human beings mental and physical life and these impacts have to be carefully monitored. (European Commission, 2019, 22.)

SHAPES is a project that aims to improve the possibilities for ageing to live longer at home by offering AI based solutions. In this project there must be careful considerations what might be the social impacts of bringing technology to the lives of the elderly. Does it increase the opportunities to social communication or is there a chance that new technology will make the ageing population lonelier and have less face-to-face social contacts?

Society and democracy. In addition to the impacts of AI on individuals also the impacts on society as a whole should be examined. These impacts may include institutions, democracy and society at large. Especially important is to consider how the AI systems may affect political decision-making and electoral contexts. (European Commission, 2019, 23.)

7.Accountability

Accountability is closely linked to principle of fairness. It means that all AI system processes and results that AI gives must be accountable and responsible (European Commission, 2019, 23). It is important to have mechanisms that ensure in all phases of the AI lifecycle there is a responsibility stated if not personal level, but at least organizational level.

Auditability. Auditability means that the algorithms, information, data and development process are open to evaluation. This does not mean that the all business models and intellectual property related to AI should be openly available. However, it should be available for internal and external auditors to evaluate. Possibility to independent evaluation will increase the trustworthiness of the AI and it should be always available for audition in cases when there is fundamental rights or safety-related applications involved. (European Commission, 2019, 23.)

Minimisation and reporting of negative impacts. Identifying, assessing, documenting and minimizing negative impacts and harm related to artificial intelligence is fundamental to those who are directly or indirectly affected. Protection must be available to those who report legitimate concerns related to AI system. Using impact assessment tools during the development and design process, deployment and use of AI are helpful to attack negative impact. (European Commission, 2019, 23.)

Compromises and redress. There might be situations when there is a tension between the requirements stated above. They may lead to compromises. In the situation when there is a demand for compromise, it should be made in rational and methodological manner. When the ethically appropriate decision cannot be produced, the development, deployment or use of AI system must be deployed until the ethically acceptable way is in place. In addition, when the trade-off is made the decision-maker must reason, report and documendate the decision. There should be mechanisms in place in order to redress the unjust impacts. Knowing that redress is possible when things go wrong creates trustworthiness. (European Commission, 2019, 23-24.)

European Commission decided to give additional requirements for high risk solutions in its white paper 2020. The additions include separate instructions for training data, keeping of records and data and specific requirements for remote biometrics devices.

Training data. Training data sets are required to be broad and need to cover all relevant scenarios needed to avoid harm and danger. The data sets should be sufficiently representative. In addition, the privacy matters should be already taken into account when training AI. If the conformity assessment shows that the AI system does not meet the requirements for example with the training data used, the identified errors and shortcomings should be correctified for example by re-training the program. (European Commission, 2020, 19-23.)

Keeping of records and data. In high-risk scenarios European Commission recommends keeping records in the relation to the programming of the algorithms and the data used to train the high-risk applications. This also includes how the data set used was selected. In special cases also keeping of data is recommended. With these measures developers can increase traceability and decision-making of the AI system can be verified. (European Commission, 2020, 19-20.)

Specific requirements for remote biometric identification. The gathering and using biometric data such as facial recognition carries specific risks to fundamental rights violation. GDPR gives in principle the specific limits of using data that identifies natural person. Such processing can take place only in specified circumstances. (European Commission, 2020, 21-22.)

5.2.4 Technical and non-technical methods to achieve trustworthy AI

European Commission gives methodological advice how the requirements introduced can be achieved. The methods implemented should be constantly evaluated, reported and justified. European Commission's aim is to provide a list of methods that might help to implement trustworthy and ethical artificial intelligence. However, Commission declared in its white paper released on 19.2.2020 that the same kind of regulation than for electrical devices might be needed regarding artificial intelligence. (European Commission, 2020, 9-10.) That would mean mandatory technical testing and regulations.

1.Technical methods

Architecture. The requirements introduced earlier should be somehow turned into technical procedures that the AI system should follow. This can be obtained with rules that the system should always follow, in other words “a white list”. There should also be “a black list” which provides constrictions on behaviors or states of the system that it should never exceed. (European Commission, 2019, 24.)

Another method is so called “sense-plan-act”-cycle which is easier to apply for the AI systems with the learning capacities. The architecture of AI system must be adapted to the all phases of the cycle. Firstly, “sense”-phase the system must recognize all the elements in the environment that are necessary to ensure that the requirements are followed. Secondly, at the “plan”-step, the system should only consider plans that are levelled with the requirements. In third, at the “act”-stage, the system actions should be restricted only to the behaviors that embrace and realize the requirements. (European Commission, 2019, 25.)

Ethics by design. Methods that embrace values-by-design provide explicit links between the abstract principles that the system must obey and specific implementation decisions that the system makes. In values-by-design-method the essential feature is that the compliance of norms and ethics is implemented by design. In other words, the developers are considering the impacts and norms throughout the design process to be able to avoid the harms the AI system might otherwise carry. There should be fail-safe shutdown mechanism in place for the situations that are impossible to cope otherwise. (European Commission, 2019, 25.)

Explanation methods. To the AI system to be trustworthy, it must be explainable. This might be difficult in the case of the neural networks and AI with learning capacities. At the moment the field of Explainable AI is young, but important. It should be used as much as possible. (European Commission, 2019, 25.)

Testing and validating. Testing and validation of the AI system should take place in all stages of the AI system lifecycle. They should include all system components data, pre-trained models, environments and behavior of the system as a whole. It must be ensured that the development process results are consistent with the input and the decision process can be validated. The metrics for the categories that are tested should be developed. These should include the adversarial testing by the trusted and diverse groups that try to deliberately break the system to find the vulnerabilities, errors and weaknesses. (European Commission, 2019, 25.)

2.Other methods

Regulation. There is already regulation that supports the trustworthy AI. For example, product safety legislation and GDPR.

Code of conduct. Organizations and stakeholders can establish a Code of Conduct and adopt sustainability and responsibility as one of their Key Performance Indicators. The organizations developing, employing and using AI can document their aims and procedures towards fundamental rights, transparency and avoiding harm to create atmosphere of trust. (European Commission, 2019, 26.)

Standardization and certifications. Standards for design, manufacturing and business practices are already in place, but development and use of artificial intelligence does not yet have standards. European Commission promotes some sort of certification or standard for trustworthy AI. It could include for example technical safety, security and explicability. (European Commission, 2019, 26.)

Accountability via governance frameworks. The ethical dimensions of AI development, deployment and use can be included when the organizations will have both internal and external governance frameworks. There can be an ethical advisor or advisory board that can give advice and provide oversight on ethical issues that may arise. Sharing best practices is also recommendable. (European Commission, 2019, 27.)

Education and awareness. Basic AI literacy, awareness and education should be promoted in a society as a whole. This should include awareness of the possible impacts of artificial intelligence in human lives and to the society as a whole (European Commission, 2019, 27). The Commission suggests that the ethical guidelines produced for the technical developers, could be used also as a resource of the training institutions (European Commission, 2020, 6).

Participation, social dialogue and diversity. Artificial intelligence may bring great benefits. It is essential to make these benefits available for all. This demands for open discussion and participation of all stakeholders, social partners and the public. Organizations can promote discussion by creating stakeholder panels which include experts on various fields like legal advisors, technical experts, ethical advisors and so on. Ideally these panels also promote diversity by including people from different cultural background, gender, education and ability. (European Commission, 2019, 27.)

5.2.5 The List

Finally, the European Commission introduces a checklist for the AI solution developers and providers to follow throughout the development, implementation and use of AI system. The complete list is in the Appendix 1. The purpose of the list is to provide a simple checklist of the matters that should be covered and considered when developing AI. Advantage of this list is that is a good reminder and very thorough. On the downside, the list is quite heavy tool for everyday work. It might be relevant to consider having a shortlist to measure every day ethics instead of working based on the list as a whole.

5.2.6 Discussion on European Commission's AI guidance

As we can see the European Commission's expert group has discussed very throroughly about issues that are involved when trying to develop responsible AI. Some of the problem areas addressed are not in fact related to AI ethics. Many issues are mainly technical and involved with safety and security. Those issues need to be discussed when AI systems are developed in responsible way. In discussions on AI ethics there are four issues that keep appearing; transparency, fairness, accountability and explainability.

5.3 Examples of policies for ethics of artificial intelligence

In addition to European Commission's guidelines to AI ethics, also some companies policies to AI ethics were taken into consideration. Of course, the problem to this approach is that many companies do not share their guidelines publicly or their public guidelines are on very general level. However, it is important to see what are the ethical objectives that business life sees the most relevant. However, it should be noted, that some research believe that ethical guidelines are also limiting when discussing on AI ethics. This might address the discussion only to matters that are stated in the guidelines and the issues outside the guidelines or new issues emerging will be underestimated or left outside from the discussion.

5.3.1 IBM's Everyday Ethics for Artificial Intelligence

IBM has been one of the most open operators sharing their ethical guidelines. They have published their guidelines in a guidebook Everyday ethics for artificial intelligence. The structure of this guidebook is quite similar to European Commission's guidelines. Firstly, they give five areas of ethical focus: Accountability, value alignment, explainability, fairness and user data rights. These can be seen as a selection of European Commission's most important guidelines that are more straightforward to take into action. Throughout the guidebook IBM gives use cases and examples, how these ethical values can be executed and they often give recommended action to take and like European Commission, IBM uses questions for the team to promote ethical discussion. In general, IBM has taken their guidelines well to practical level. (IBM, 2019.)

5.3.2 Google's Artificial Intelligence at Google: Our Principles

Another big player in a field of AI is Google. Google has published their common level principles for ethical AI. It emphasis is on general level guidelines. Google's first principle is "Be socially beneficial". This principle is missing on IBM's guidelines, but it is one of the most important guidelines in European Commission's paper. The reason why it is missing from IBM's guidebook is probably because it is more practical level than Google's guidelines. This principle is not easy to take on the practical level since many innovations can be used both good and bad. Google has taken this into account in their principles. They have created a checklist that they use in evaluation of the AI application. These are 1) primary purpose and use, 2) natura and uniqueness, 3) scale and 4) nature of Google's involvement. Google's approach to ethical principles also differ traditional approaches since it gives a guideline to AI applications that Google does not pursue. This includes applications that cause harm (compare to European Commission's) and weaponary. Google also has raised scientific excellence as one of their principles. This means that Google will promote sharing AI knowledge by educational materials, best practices and research. This is aligned with European Commission's goal to share AI knowledge openly. Otherwise Google's AI principles are very much alike with IBM's and other companies published guidelines including accountability, unfair bias etc.(Google, 2020.)

5.3.3 IEEE's Ethically Alligned Design

The Institute of Electrical and Electronics Engineers (IEEE) has published ethical guidelines 2019. This guidebook is referred in many company policies considering AI among European Commission's guidelines. The approach to ethics is very similar to European Commission's, but somehow more complex. The basis (pillars) for their recommendations is in human rights, self-determination of data agency and technical dependability. The same way than in European Commission's paper, also IEEE has set general principles to ethically sustainable design. The IEEE's principles are equivalent to European Commissions seven requirements, even though IEEE has eight principles instead of seven. The IEEE's principles are human rights, well-being, data agency, effectiveness, transparency, accountability, awareness of misuse and competence. As we can see the principles are named differently eqvivalent to European Commission's, but the content is mostly the same. Only notable difference is that awareness of misuse is separate principle (IEEE, 2019; European Commission, 2019).

The content of European Commission's requirements for AI and IEEE's principles are almost identical. Also how these requirements and principles are mapped to upper level ethical basis is similar and has the same values included. However, how the priciples or requirements are considered to introduced to action are different. European Commission has introduced an assessment list which has listed a number of questions that AI system should be assessed to be trustworthy. IEEE has introduced issues and recommendations to take ethical guidelines into action. IEEE's approach gives broader recommendations than European Commission's list. It's recommendations are much more on the same level than European Commission's requirements. European Commission's list takes the practical side a step further by giving simple yes and no-questions to developers of AI to consider (IEEE, 2019; European Commission, 2019).

Thilo Hagerdorff in 2020 article evaluated 22 different ethical guidelines for AI. He found that in 80 per cent of the guidelines handle privacy, fairness and accountability as minimal requirements of responsible AI system. He also noted that these matters in addition to robustness and explainability are more easily solved as technical matters than the social issues that might be arising from the development of AI system. Hagerdorff also found that the company codes of ethics were the most minimalistic which was also verified in my review. He also found that the ethical guidelines usually do not commit to larger societal interests. I believe this is partly because the societal issues and wider effects that AI has on the society are hard to write on the form of the simple guidelines. (Hagerdorff, 2020.)

McNamara et al. reviewed how the ethical guidelines effect the work of the software engineers. They concluded that the ethical guidelines given had almost zero effect on the practices of the professionals. Unfortunately, the study did not examine the reasons why there was no effect. (McNamara et al, 2018). It has been also discussed that checklists like the one European Commission has provided, might instruct the development to focus only to the matters that are on the checklist, not the problems that might be completely new or that was not added on the checklist at hand.

5.4 Machine ethics

Machine ethics is a field of research which instead of developing ethics for humans that use the machines, is developing ethics for machines.The definition of machine ethics is not always clear. In contrast to computer ethics, machine ethics is considered to include machine behaviour toward human beings and other machines. (Omari & Mohammadian, 2016, 231.) The theoretical idea is to give machines ethical principles or procedure that they follow when they encounter an ethical dilemma. The machine itself will function ethically responsible manner via this built-in procedure. This view of thinking has one relevant outcome. The machines are able to work autonomously without human intervention. Machine ethics recognizes also the alternative in which the machines are seen as tools that are used by human beings. The ethics in this case is ethics of human beigns and how they should use the machines in ethical way. (Anderson & Leigh Anderson (ed.), 2011, 1-2.)

James Moor has discussed in his study of four ways of bringing value discussion into machines. Firstly, machines like computers can be seen as normative agents that are not necessarily seen as ethical since they are created for the specific technological purpose in mind. In this role they are assessed based on their performance on this specified purpose. On ethical point of view they are seen as they would not have an ethical impact, but they still have value which can be economical, aesthetic or practical. (Moor, 2011, 13-14.)

Secondly, the machines can be seen as ethical impact agents that not only perform tasks, but also their performance has an ethical impact on to the society. Moor recognizes that neither of these two ways introduced are actually not bringing ethics into the machines, but the next two are truly ascribing values into machines and are on the core of the machine ethics.

Implicit ethical agents are the machines that are programmed to support ethical behaviour. Often this is done by avoidance of the unethical decision making. The designers of the machines are following the ethical principles that are guiding the buiding of the machine and its features. When the machines act as a implicit ethical agents they are designed not to cheat and to protect lives, but they are still not making the ethical decisions per se. Moor argues that the action of the machines is based on their virtues, which brings us back to more general framework of common ethical theories.Human beings learn virtues by habit, but with machines virtues are more purpose related and limited since the machines do not have priactical wisdom. (Moor, 2011, 14-16.) This approach raises immediate question, who's values are machines to follow since the virtues of the machines are still given from the outside? It means that the ethical guidelines of developers must be transparent and open to critique.

The fourth way of bringing ethics into machines that Moor studied is machines as explicit ethical agents. Are the machines able to a similar approach to ethics than human beings? Even though this approach seems unrealistic, in the literature there are some attempts to advance this matter. Jeroen van den Hoven and Gert-Jan Lokhorst (2012) have studied the relation between logical reasoning and bringing ethics into machines. Deontic logic states the permissions and obligations, epistemic logic states beliefs and knowledge and action logic states the action. Together these logics might be the solution that describes the ethical decision-making process with sufficient precision to be able to teach the machine to make ethical judgements. Machines as explicit ethical agents suggest that the machines like AI would be able to make ethical judgements and justify them. At the moment it is not clear how this might be achieved. In his study Moor provides an good example for problems that might arise. In disaster situation the decision-making of the human beings might not be as systematic as machines since there are emotions and stress involved, but are people ready to give the machine the power to decide on life and death even though it might save more lives than human decision?

This leads us to consider the possibilities of machines ever to become full ethical agents. An average adult is considered to be full ethical agent with free will, consciousness and intentionality. Many argue that machines cannot have these features and therefore it is impossible to them to become full ethical agents. Therefore it is more important to focus on researching machines as implicit and explicit ethical agents. This is important since it makes us think ethics and why it is so important to teach the machines to act ethically. (Moor, 2011, 17-20.)

This brings us to consider the matter if the ethics is computable problem at all. On what grounds the machine is acting ethically? This brings us back to common ethical theories. If the human beings have several theories of ethics and considerations how we should act when acting ethically, how should we choose upon which theory the machines or AI should make its decisions? Omari & Mohammadian argue in their study that AI should use not just one, but several theories as basis of its decision-making.

6 Environment

6.1 People

Hevner's method for Design Science Research includes Environment as one of the pilars for the design process of the artifact. This includes the examination of the special features of the people involved.

6.1.1 Ageing citizens

Europe is facing a major demographic transformation. When the birth rates are continuously declining and the lifespan of the European citizens is increasing, Europe needs to find new solutions to be able to cope the massive change of the age structure. At the moment we have four working citizens per aging, by year 2050 we only have two. This means that we need to create a new kind of solutions to be able to offer good and valuable life for the aging population and to maintain the functional society. As Cabera and Malanowski state on their introduction of Information and Communication Technologies for Active Ageing, this transformation can be also seen as an opportunity to create an innovative market and society that has new products, services and structures that are addressed for the ageing. (Cabrera & Malanowski, 2009, 1.)

According to Alan Walker there has been a close relationship between older citizens and welfare state since the World War II. This has mostly defined the discourses that the elderly has been discussed. Even though there has been variation throughout Europe how the society and welfare system have been constructed, the perception of elderly has been very similar. The ageing people have been seen as passive recipients of pensions, health care and other welfare state policies. The retirement meant also social and political exclusion. The ageing people were excluded from main sources of political and social channels of representation since they were not part of the economic system. (Walker, 2009, 35-37.)

From 1990s onwards there has been a new discourse of perceiving elderly arising. The key element of this active aging discourse has been the rise of the individualistic consumerism. The ageing people have more funds to spend and they are in better health than the generations before. This has promoted an emergence of the new market for the products and services directed for the ageing population. The new policy discourse “active ageing” has been promoted in EU in recent decades. One of the first milestones was the promote the employment of the older workers. In 2001 the EU released Employment directive that created more ageing friendly employment policies and concept of active ageing which aims to keep the older workers in working life longer. At the same time the pension regimes were made more flexible so that the employment rate of the older individuals could be increased. The main goal for the changes in employment of the ageing is to create more sustainable pension system. (Walker, 2009, 40-41.)

EU has also underlined the need for the ICT solutions to be able to promote active ageing. For example, the Ministerial Declaration from the Conference ICT for Inclusive Society in Riga addressed more broader view for the active ageing. The most important view is the right to participate. The aim is to create via ICT more active participation of the aging to the societal and economic life and to promote self-expression and social contacts of the elderly. This means increased quality of life and the sense of safety, security and autonomy while the aging can continue to live in the familiar environment of home. (Walker, 2009, 41-42.)

Walker articulates also a few arguments that are to be considered when the active ageing is considered. The EUs vision of active ageing is quite concerned with employment and productivity. How about other meaningful activity that ageing citizen can have within the family, their own community or in the society? The approach focusing on productivity might produce the top-down policy action instead of ageing to develop their own forms of activities. Secondly, Walker argues that the aging should be seen as a life-long process that all age groups should be seen as active members of the society. The close eye on should be maintained to keep the solidarity between generations. Thirdly, when focus is on the younger old, the older elderly might be forgotten and they will not get the attention to their special demands they need. European wide action plan should consider also the fact that the cultural differences exist in the ways of participation especially between southern and northern Europe (Walker, 2009, 45-46). United Nations and the World Health Organization have enhanced the active ageing as a goal. This means the process of opportunities for health, participation and security in ageing societies. (Hausknecht et al, 2015, 198.) I might add that there are also great differences between individuals of willingness to participate and being active. The activity should not be the norm, but a choice.

6.1.2 Ageing citizens and technology

Technology we have today is developed based on the needs of the young people with full capabilities. Rapid change in technology has forgotten the needs of the ageing population, therefore they have no access to all functions and services. The need for development of the technology that is based on the needs and existing capabilities of the ageing population is essential.

Gerontechnology is one possibility to bring technology into the life of the ageing. Gerontechnology is defined to have five ways of promoting development of ageing-friendly technology. Firstly, technology should aim to prevent accidents and delay loss of capabilities due to ageing. Secondly, gerontechnology should utilize existing capabilities and knowledge of the ageing. In third, technology should be developed and designed in that way that is compensates disabilities and can give answers to challenges the ageing population might have in their everyday life. Fourthly, the technology should help caretakers and health care professionals to support active ageing and in fifth, gerontechnology should promote active research of the ageing. (Kaakinen & Törmä, 1999, 6-7.)

When considering the ageing of the future, the challenges might not lie in technical capabilities of the ageing since most of them have used ICT at work. The challenges are more of the loss of other capabilities like eye-sight, movement or memory. Often the barriers for the ageing population to live at home are releated to ordinary everyday issues like house maintance or maintaining daily routines at home.

When considering implementing technology for the ageing population, one must be aware that the deployment of technology itself should not be the main intention, but developing good and working concepts of services and policies that aim to improve the life of ageing. Deployment of technology is not only building a user relationship between the ageing and the technology, but bringing the individual to be part of the larger system of different parties, technical solutions and objectives. Technology should be coordinated with social network which enables social and technical support. Technical solutions should be part of the larger service ecosystem that aim to increase well-being not only for ageing, but also for caretakers. Likewise, the technology should rather facilitate and make the work of caretakers easier than make it more complex and take more time. (Viirkorpi, 2017, 45-46.)

6.1.3 Health Care Professionals

SHAPES ecosystem is not only used by elderly, but also by health care professionals. The data gathered must be analysed and decisions for example of medical treatment must be made by health care professionals. Fairly, we could say that even the best AI system cannot treat medical conditions alone. There must be professionals working besides of the AI. AI system is considered to work best when combined with a human being. Both can use their strenghts to make better decisions.

Technology will change the work of the health care professionals. There are technological concepts like robotics, information systems, sensors and medical devices that will have significant effect on health care. Robots might work side by side with humans, information systems will advice diagnostics, sensors and medical device provide data and different kinds of remote health care systems will make caring less place dependent.

6.2 Technology, the SHAPES Ecosystem

Living at home if possible is considered the best option for elderly. As cababilities of the ageing are often weakened, the need for assistance, health care and medical services at home are essential to able ageing population to continue to live at home.

SHAPES (Smart and Healthy Aging Promoting Empowering Systems) is a European wide scheme that aims to create an open ecosystem that enables different kinds of digital solutions for supporting independent living for aging individuals who are facing reduced functionality or capabilities. The aim for SHAPES is also to build an ecosystem or a platform that not only has smart digital solutions for aging but also integrates these solutions in order to collect and analyse health, environmental and lifestyle information to be able to identity special needs for the elderly and provide personalised solutions. (https://shapes2020.eu/.)

One of the key goals of SHAPES is to build European ecosystem that is attractive to health care industry and policy-makers and builds a market for deployment of innovative digital health and care solutions and services supporting and extending healthy and independent living of the aging population in Europe.(https://shapes2020.eu/.)

There are 36 partners in 15 different countries and the project is funded for four years. SHAPES implements a co-creation methodology between social sciences, technological development, piloting and deployment activities. The technology created should be based on user needs and requirements. SHAPES also adopts an ethics-based approach taking the protection of the human rights of the aging population as central focus point. (https://shapes2020.eu/.)

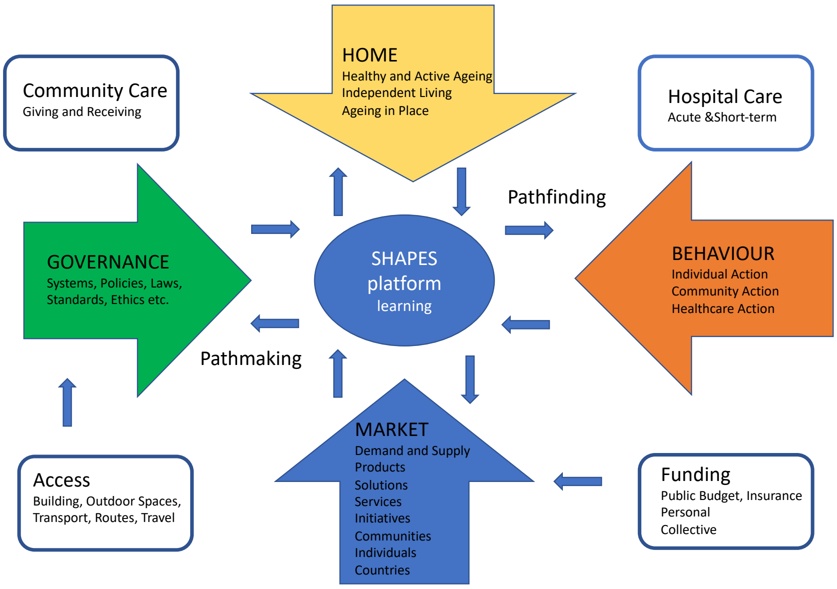


Figure 3. SHAPES platform (https://shapes2020.eu/)

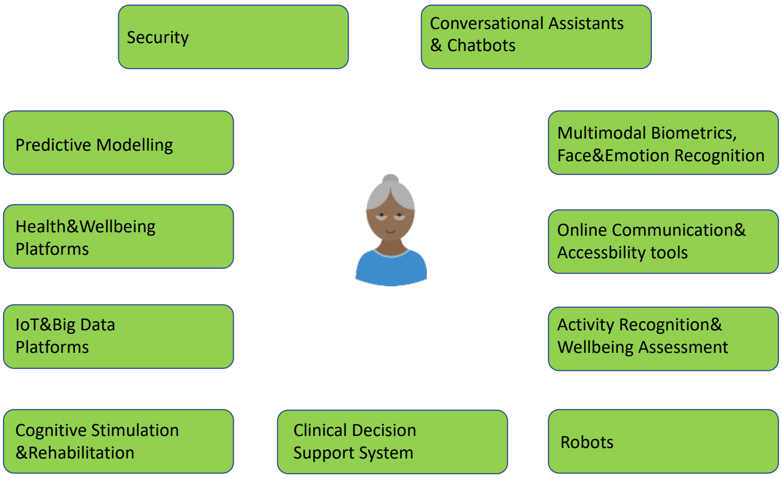

Security
Conversational Assistants & Chatbots
Predictive Modelling
Multimodal Biometrics, Face&Emotion Recognition
Health&Wellbeing Platforms
Online Communication& Accessbility tools
IoT&Big Data Platforms
Activity Recognition& Wellbeing Assessment
Cognitive Stimulation &Rehabilitation
Clinical Decision Support System
Robots

Figure 4. SHAPES applications (https://shapes2020.eu/)

6.3 Legistlative regulation

6.3.1 GDPR

The General Data Protection Regulation is a legal framework which purpose is to set guidelines for data collection and processing personal data in European Union. Personal data is all information that relates the individual and that individual can be directly identified with. This data includes names, social security numbers, e-mail addresses and data like biometric data, ethnicity, gender or political background. GDPR obligate the record keeper to ask specific consent for the individual to process the personal data in their system and the subject must have right to withdraw this consent at any point of time. The aim of the GDPR is to give individuals more control over their personal data. The rights that GDPR gives include The right to be informed, the right of access (how the data is used), the right to be forgotten and data deletion and the right to deny the use of their own data.

GDPR also obligate the record keeper to process data securely and protect the data with appropriate technical methods. (GDPR.EU, 2020).

6.3.2 Human rights

United Nations' Universal Declaration of Human Rights is a basis for all human rights work. It devided into 30 articles which have an important role of promoting respect of these rights and freedoms in all societies and nations. Human rights are in place to protect all people from political, social, legal and other kind of abuse. The human rights include Civil and political rights (the right to life, liberty, and property, freedom of expression, pursuit of happiness, and equality before the law), social, cultural and economic rights (the right to participate in science and culture, the right to work, and the right to education). The human rights are the basis for the AI ethics. When for example economic matters and human rights are in conflict, the human rights should be prioritized. But also the different human rights might be in conflict with each other. The discussion might be if the right to life is more important than the right to privacy. (University of Helsinki, 2020, United Nations, 2020). The human rights that are the closely discussed when discussing artificial intelligence are privacy, inclusion and security.

6.3.3 Health care regulation

The overall legistlation concerning health care is not regulated in pan-European level. European Union has common EU Health policy. The role of the European commission is to complement national policies, to propose legistlation, provide financial support and facilitate the exchange of the best practices between EU countries and health experts. (European Union, 2021.)

The principles behind the national legistlation are the same in all European countries even though the exact contents of the regulation may vary. For example, the purpose of the Finnish Health Care Act is, among other things to promote and maintain the health, well-being, ability to work and function and social security of the population, to reduce health inequalities between population groups and to implement equal access, quality and patient safety of the services needed by the population. (THL.fi, 2021).

One European Union regulation that is concerning use of AI in health care is the legistlation and regulation of medical devices. It aims to ensure the smooth functioning of the internal market and protect the health of the patients and users of these devices. This act gives boundary conditions among other things to safety, traceability and transparency. (European Union, 2021.)

It is important to see that the health care regulation gives also frames to use of artificial intelligence. The patient safety and protecting patients medical history are regulated by the law and are valid also when use of artificial intelligence in health care decision-making become more common.

7 Special features of AI ethics in SHAPES

SHAPES is a project that aims to increase wellbeing of the ageing population with technology. This aim brings along special issues to consider. First of all, the issues around AI ethics can be devided into three categories; issues related to data, issues related to use and safety and societal issues. The most important issue on AI ethics related to the certain solution is that the open discussion on problematic issues is allowed. Another aspiration is to consentrate the discussion on the matters that have the major impact. For example it is more relevant to concentrate the ethical discussion on the issues of healh care than on which film Netflix is recommending even though both might be using AI and we might face ethical issues in both cases.

7.1 AI ethics in health care and use of health data

SHAPES is a health care project. The ecosystem planned will use health data gathered from its users. When using this kind of data we will face special issues. For example if we want to use the collected data in such way that person can be recognized in any way, there must be content from that person for each use that the data is planned to be used.

Health data can be collected from various sources, wearables like sports watches or heart rate monitors, implants or other medical devices, health records. Health data is fractured and it cannot be easily combined, shared or analysed because its nature. This is where the AI technologies might be useful when they can analyse enourmous amounts of data from various sources and help professionals like doctors to see correclations and connections that might be otherwise hidden. (Neittaanmäki et al, 2019, 92.) The other relevant field of use of AI technology in health care is intervention and prevention. This is strongly related to the data that is received from wearables and other devices. Kaasalainen, Ruohonen & Neittaanmäki concluded in final report of the University of Jyväskylä project on AI and health care that if there is only data from users in certain application, it is not bringing enough information and data benefit the society or health care professionals as a whole. Wider ecosystems and portals are needed to secure cooperation between professionals, patients and society as a whole. (Kaasalainen et al, 2019.) SHAPES-project is aim is to create such an ecosystem on European level.

In their study Neittaanmäki et al. identified the gains and advantages when using AI-based systems in health care. These are the better understanding of individual's health and better understanding of national health situation, identifying those at bigger risk and who are not receiveing enough treatment, better managing of the costs, better support for clinical decision-making and research, recommendations for intervention and overall early detection and identifying of health issues and better coordination of treatment. When combining user interface for patients to artificial intelligence, the advantages for individual are easier sharing information, better collaboration between patient and health care professionals, empowerment of self-management of health and overall well-being. (Neittaanmäki et al, 2019, 93-94.)

There is a lot of studies of how AI systems can be used in medicine, but most of these studies are not covering ethical matters to the extent. As Anna Seppänen in her lecture on AI ethics stated, the more important the matter and the more it effects on the human lives, the more there should be ethical discussion. Using AI in medicine and health care have a enormous effect on individuals, so we need to weigh the ethical matters carefully. If we can save human lives by using AI, what could be the viewpoints to prevent the use? Is it losing human control or is it losing our ability to check if the resolution that AI is correct? For example if the AI can find the cancer better and earlier than a doctor, is it unethical not to use it, even if we do not fully understand how AI does it? Or if the AI system makes an error or wrong decision. What grounds we as humans decide to use or not to use AI? Is it that the AI can make same amount of mistakes than human counterpart or do we demand that AI system cannot make any mistakes?

Deloitte Belgium has created a good review on AI in healthcare. They summarize that artificial intelligence is currently developed in eight fields of healthcare: wearables, imaging, laboratory applications, physiological monitoring, real world data, virtual health assistance, robotics and personal applications. In their study they reviewed impacts on healthcare in three measures, time saved, lives saved and financial resources saved. This approach is the same than on European Union recommended socioeconomical impact assessment (Deloitte Belgium, 2020). Downside of this approach is that it does not take into account the problems underlie by using these applications or data gathering. Ethical issues remain undiscussed. However, this approach is good at making the impacts of the artificial intelligence in healthcare visible and gives a good basis for the ethical discussion.

7.2 Use cases of AI in health care

When discussing AI ethics in health care, it is easier to approach the issues using examples. In this chapter there is discussion on use cases where the artificial intelligence is already in use in health care and has a lot of potential to change the structures of health care. There is also some discussion on the ethical issues that needs to be considered.

7.2.1 Prevention and intervention

Prevention and intervention is one of the most interesting use case of the AI solutions in health care. On one hand it is a effective way of health care cost management and on the other it is a good way for further better health for citizens by preventing future illness. The use of artificial intelligence has a great potential to shifiting the focus of the healthcare from treatment of illnesses to preventing them. According to Deloitte’s study preventive measures that could be achieved just with use of wearables combined with artificial intelligence might save EU-widely up to 313 000 lives, circa 50 million euros and about 300 million hours saved as working time of the healthcare professionals. This has a huge effect in societies if we can make this kind of change. (Deloitte Belgium, 2020.)

The ethical issue that might arise from this kind of use of the data from the wearables or other kind of physiological monitoring is the privacy. For example, from the data of wearables AI could be able to find the factors that are risk for depression. We could be able to indentify the individuals that are at greater risk based on this data and inform them beforehand. We could make intervention with preventive measures. However, GDPR restricts the use when the individuals have not specifically given their permission to use the data on this purpose. The person with these risks might not want to know. How about if we had this kind of warning system and this highly sensitive data was leaked to malicious party or if the data of this kind would be criteria if you are able to get a health insurance or a job? Is the monitoring the risk factors one form of surveillance per se?

7.2.2 Virtual medicine and healthcare

The distribution of medical resources is often geographically uneven. People living in remote areas are more likely to say that they do not have the same possibilities to have health care than the people living in cities. Also the costs of health care services are unevenly devided. Remote medicine and virtual health care might bring some relief to this availability problem.

The health care platforms are integrated to users' mobile devices to gather relevant data and the systems based on artificial intelligence can bring the relevant information to the experts to analyse or AI can be used as an analytic tool to give health advice without any contact with the healh care professionals. For example, IBM Watson Health is able to analyze unstructured data and answer the questions based on this information. (Neittaanmäki et al., 2019, 154-157.) Applications like this can be used as "virtual doctor" which analyze data based on the patients symptoms and give preliminary diagnostics or make recommendation if the appointment to the doctor is needed or if the patient could manage with the symptoms at home.

Advantages of the applications like mentioned above are easy to see. Availability discussed earlier is one of them. People living in remote areas might receive some kind of response to their medical issue more easily and they might not have to travel long way to have medical attention. The applications are also saving the resources. The health care professionals can concentrate on the cases that need immediate medical attention. European Union has also considered this matter of access to health care services, since they have stated this to be special sectoral challenge of public health care in Europe. They have a special concern also that the pricing of the medical devices should be regulated so that all EU citizens could afford these new health care inventions. (European Union, 2020.)

7.2.3 Diagnostics

Diagnostics, especially imaging is already now benefiting from the use of artificial intelligence. AI applications can be trained on large sets of medical images to detect anomalies, for example cancer. The benefits of the use of AI in diagnostics is its accurancy, speed and tirelessness. This might save lives since the diagnostics is more accurate and the results can be delivered faster to the patient and medical actions can be taken accordingly. (Deloitte Belgium, 2020.)

The ethical issue that might arise from both virtual health care and diagnostics is the possibility of error or faulse diagnosis. The data used might not be accurate or the artificial intelligence might make faulse decision. It is important that we make sure that the data accuracy is tracked and decisions are monitored by humans. It is also important to discuss if the decisions that artificial intelligence makes should be more accurate than the ones that humans make. Also the issues of privacy and safety remain like in all health care data use.

8 Design Science

8.1 The Design process

The purpose of this study was to create ethical guidelines for the development and use of artificial intelligence. When I began the process of designing ethical guidelines for the SHAPES-project lifecyle, I was certain that I could provide guidelines that could be used as a guidebook of designing responsible and ethical AI. This would mean technical and human-centered solutions how to solve ethical issues concerning artificial intelligence. The field of artificial intelligence has not established practices on ethics comparing to for example biomedical ethics or environmental ethics. The discussion on this matter is now extremely lively. When I began to explore the ethical guidelines concerning artificial intelligence the most guidelines available were from commercial operators, but for example also Finnish Tax Administration has published their own guidelines for AI. But now there are more than one hundred listings of AI ethical guidelines internationally. The variation in emphasis is large from pratical listings to more theoretical ones. (Nieminen, 20.11.2020.) The guidelines that are publicly available usually are not extensive, but give some insights that the operator has had some thought of ethical matters related to ethics of AI. In many occasions it has been said that the ethical guidelines are a sign that the organization has has some thought of ethical matters of artificial intelligence. It is also said that specific ethical requirements or guidelines might instruct the development and use of AI only to consider the ethical concequences of the matters that are mentioned on the guidelines or are required to be examined.

The further my studies on ethical issues and theories went and the more AI ethics seminars I attended, the harder it became to give conclusive guidelines or instructions on ethical issues that concern the development and use of artificial intelligence. Hevner's Design Science reseach as my guideline I went through a theorethical background of common ethics and artificial intelligence and find the issues that are the most relevant concerning artificial intelligence.

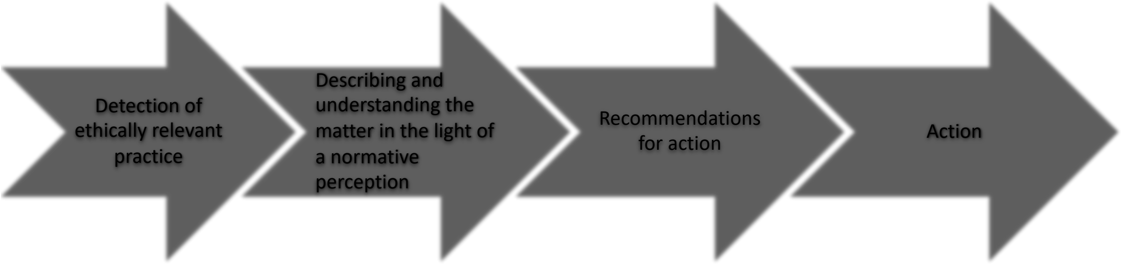
My study of the ethics of artificial intelligence and purpose of the creating guidelines for SHAPES was defined by the Hevner's Design Science Research. My first aim was to gather relevant theoretical background of the common ethics and deepen that knowledge towards machine ethics and ethical viewpoints of AI ethics. Soon I noticed that there were not so much literature on AI ethics since the discussion on the issue is quite recent. I also had a dilemma that what should be included when discussing AI ethics. When studying the field I noticed that one viewpoint always lead to another viewpoint that seemed relevant to the topic at hand. I made the decision that in this study I will only briefly cover technical robustness or data protection since they are of course relevant to trustworthiness of AI and might also have some ethical notions, but they are requirements of any technical solution that effects human lives, not specific to artificial intelligence.

Big part of the theoretical knowledge base is the European Commission's ethical guidelines for artificial intelligence. This is because of the nature of the project as pan-European project that is also EU funded. Since SHAPES is EU-funded project, the European Commission's Ethical Guidelines for AI have a great relevance and I needed to concider if the ethical guidelines for this project could be based on the work European Commission has already done. To compare I also studied ethical guidelines for some companies and organization to get a picture how the ethical issues are approached in commercial environment.

My first version of the ethical instructions or guidelines for this project was quite different from the final version. At first it was quite specific list of ethical matters related to AI and how to try to solve them. Hevner's theory as my guideline I went through many iteration rounds between my guideline suggestion, knowledge base and the environment. In this study the artifact is the ethical instructions or guidelines to be provided to the promote ethical behaviour during the lifecycle of the AI system. I especially had in mind the development process. Because of the nature of the project Hevner's theory's Rigor Cycle became in this study more relevant than the Relevance Cycle. The Design Cycle evaluation was created more in relevance to theoretical consideration than actual field testing or having feedback from the environment or people that the ethical guidelines will be used. This can be seen as a deficiency of this study and it can be topic for further research. After a few interation rounds, I decided that my artifact should be two-fold. There should be ethical guidelines to give practical guidelines and the other part should be some kind of framework for promoting ethical thinking and competence.

After a few iteration rounds I felt that the only guideline I could with certainty to give is the importance of promoting ethical discussion in all phases of the AI system lifecycle. This includes raising ethical awareness. When developers, users and all the people involved are aware that they are possibly facing ethical issues, they are more careful to think, discuss and resolve the issues that emerge. Mika Nieminen in his lecture on responsible technology points out the importance of the bringing guidenlines into practice. He presented four features that are relevant when concidering innovation and promoting responsible technology, forecasting, inclusion, self-reflectiveness and willingness to make necessary changes. These features are also important to bringing guidelines into everyday action. These features are strongly related to transparency and promoting discussion. (Nieminen, 20.11.2020.)

As I attended AI ethics seminars, I noticed that often the same themes were raised by lectures and attendees. One of the most interesting ones was by Anna Seppänen wh. She brought up steps that ethics discussion and evaluation should take. At the first step there is a perception that we might face an ethical issue on for example in development of the system, in second phase there is description, discussion and understanding the matter, the third step is to create recommendations for ethical action for example guidelines and the last stage is bringing the ethical recommendations into action. According to Seppänen, the ethical evaluation is a constant process and hard work also after this process is finished, since the ethical matters are not unsettled and they are constantly under discussion. I also noted the guidance of my own organization that has a similar approach to ethical concerns called Pause-Consider-Action, in which when ethical concern is faced the correct action is to pause to think and concider the correct course of action before acting out.


Detection of ethically relevant practice
Describing and understanding the matter in the light of a normative perception
Recommendations for action
Action

Steps of Ethical Evaluation

Figure 4. The steps of Ethical Evaluation (Seppänen, 29.10.2020)

8.2 Ethics by Design, Values in Design and Ethics for Design(ers)

Ethics by design or values by design is one of the key values when considering designing ethical and trustworthy AI. This approach has a goal to ensure that ethical matters are taken into consideration from the earliest stages of the project and are taken into account throughout the whole development and design process.

8.2.1 Values in design

It is said, that the technical infrastuctures and technology often reveal human values mostly because of the tensions, failures and counterproductivity. Values in design approach as Nussbaum et al. it defines, tries to create discipline that includes values into socio-technical designing process from the beginning. According to VID approach the difference in definition of ethics and values is necessary. Ethics is seen as a set of prescriptions whereas values are seen as action. This is why the values in design approach is discussing about values, not ethics. Values in design approach values cooperation, co-creation and coordination as methods of creating more human friendly technology. It emphasizes the importance of stakeholder participation. (Knobel & Broker, 2011, 26-28.) There are also other disciplines that value the same approach designing technology. For example, Freeman and Nissenbaum call it value-sensitive design, Acre calls it critical technical practice and Sengers et al. calls it reflective design. All of these have a same goal, bring values into development process from the beginning, but they also emphasize the role of the designer as a carrier of the values and the fact that the values of the designer affect the way that technology is imagined, how it handles data and what kind of decisions are made. (Shilton, 2012, 375.)

Katie Shilton studied ethnographically how the designers in CENS laboratory took ethical matters into consideration when designing sensor technology. Shilton noticed that the designers were often aware of the ethical issues of the system they developed might face, but the ethical matters were seen outside of the technical staffs expertise and as a job for someone else to do. In many cases the value-sensitive design process was seen unattractive since it was much slower. Values were not seen as important as functional system and values were often forgotten when they competed with the system offiency and practicality. However, when the designers prototyped the application, they were fully aware of the ethical issues they had, since they saw them in practice. (Shilton, 2012, 377-379.)

8.2.2 Ethics by design

According to Virginie Dignum et al (2018) Ethics by design in AI is concerned with methods, algorithms and tools that are needed to ensure that autonomous agents with capability to reason take the path of ethical decisions and that their behavior stays within the moral boundaries that are given to the system. This means that the AI system should behave in a way that it is beneficial to people and safe to use. AI ethics can be divided in two, regulation (legistlation and standards) and design (system itself). Ethics by design is considering the latter. The most important question according to ethics by design approach is that are we able to and how to build AI systems that have ethically-aware agents? The key issue is to articulate that Artificial intelligence requires researchers and designers to be able to translate human values and ethical considerations into technical requirements. Actually, the ethics by design approach is very near to values in design approach when discussing that designers must take mental shift towards thinking values before performance of the system. Ethics by design approach requires that the question about AI system reasoning should be the priority over performance. (Dignum et al, 2018, 1-2.)

According to Dignum et al. three ethical issues are particularly concerning to AI systems, accountability, responsibility and transparency. We can see that the European Commission is especially concerned with the same issues. Dignum states that these three values are important to discuss when trying to ensure the societal good for everyone. How the AI system is seen to-follow these ethical principles, depends upon what kind of reasoning is seen to be possible for AI system. If we believe that the AI system is not capable of ethical reasoning, it means that we should always have human supervision. That also means that the supervisor should have sufficient knowledge and means to do this job. This approach is called human-in-the-loop. Another approach to ethical reasoning of the system is that the environment itself has been designed that way that the deviation is impossible and the moral decision making of the system is not needed. Ethics by Design considers AI system as an ethical agent itself. These agents are known as artificial moral agents. That means that the AI system is able to include moral reasoning into its deliberation and decision-making and explain its behavior in terms of moral concepts. This approach requires complex decision-making algorithms based on deontic logics. The system design needs very explicit and complex design based on reinforcement learning to be able to act as a moral decision-maker. (Dignum et al., 2018, 1-3.)

8.3 Ethical Assessment

Ethical Assessment of emerging technologies is important since considering the good and the bad that technology, processes and devices can bring into society. Many questions of this kind of new technology cannot be fully answered since its nature, future use and its social effects are still unknown. In R&D phase the devices and technologies are not yet present, the ethical assessment is also in a way speculative. This means that the ethical assessment is to make recommendations for R&D practices and increase the likelihood that the development processes produce more ethical devices and solutions. (Brey, 2012, 2.)

Brey considers two possible approaches for ethical assessment. *Generic approach* pays attention to the generic features of the technology in question. In other words, even though the special characteristics of the application or technology are not yet known, the general ethical issues that are related to the technology can still be identified and discussed already in early stages of the development process. The second approach is to directly speculate future devices, applications and social impacts. *Forecasting approach* requires the ethicists to be educated with forecasting or future studies to be able to predict ethical issues that may occur in the future. Another method to proceed in forecasting approach is technology assessment. It is a field of study that consideres the impacts of the new technology on society, industry and environment. Its purpose is also to prevent harm and promote development of the new technologies to more desired directions. The technology assessment is made on the basis of the known or potential applications that the technology has. According to Brey both approaches have some disadvanges. Generic approach might not reach all relevant ethical issues that there might occur. However, the ethical assessments produced by forecasting approach might be more speculative and in some extent incorrect. (Brey, 2012, 2-3.)

8.4 The structure of ethical decision making

To promote ethical thinking in development process of artificial intelligence solutions the theories and methods that are already available suggest the following approaches. There are ethical guidelines which offer the broad discourse on which matters are relevant to consider when discussing developing AI systems. On this level the discussion is not yet on practical level. The second level is what I decided to call ethics by design even though all of the measures are not available in AI system developed itself. At this level the discussion is more practical. How the guidelines we have discussed on the previous level are produced as technical solutions by for example restrictions and specification. I have also added promoting ethical competence on this level since it is relevant to designers and developers to have competence to think the ways of take the ethical guidelines into consideration and to take them into practice. The third level is the checklists and discussion which offer the reminders of ethical thinking in different phases of the lifecycle of the development process and beyond. For example the checklist of the European Commission include the whole lifecycle of the AI system. The European Commission's checklist is quite heavy tool to use since it is quite long and precise. But as Commission suggests, it can be and should be used in different levels of the organization.

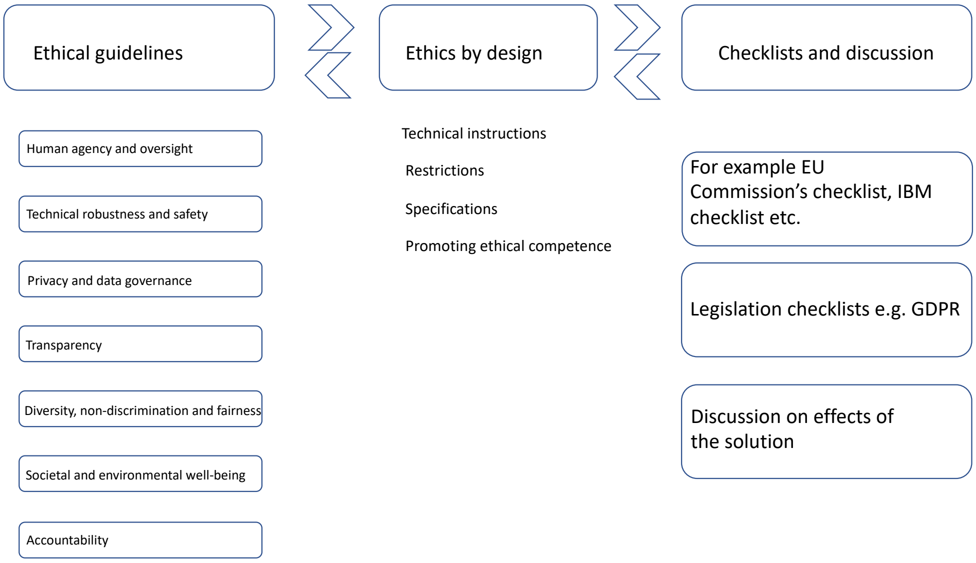

Ethical guidelines
Ethics by design
Checklists and discussion
Human agency and oversight
Technical robustness and safety
Privacy and data governance
Transparency
Diversity, non-discrimination and fairness
Societal and environmental well-being
Accountability
Technical instructions
Restrictions
Specifications
Promoting ethical competence
For example EU Commission's checklist, IBM checklist etc.
Legislation checklists e.g. GDPR
Discussion on effects of the solution

Figure 5. The structure of Ethical Guidelines process

8.5 Ethical guidelines for SHAPES

Ethics of artificial intelligence is a very broad issue. First of all the concept of what is considered to be artificial intelligence is not always easy to define. Secondly, when considering the effects of the AI, it is fast changing technology that has new solutions and use cases every day. The effects are broad and include technical matters as well as societal effects. The hype around AI is enormous and the expectations vary from AI to be the technology that saves the world to the technology that destroys the world. The pontential is huge, but at the moment we are somewhere in between. When the field of study is broad also the ethical issues to be considered vary. Every requirement on the European Commission's list has quite many issues to tackle and in this study I cannot deal with them all. Since the privacy and data governance and technical robustness and safety are more of the technical requirements than under ethical discussion, I decided that I am not concentrating on them in this study. Teemu Birkstedt and Matti Mäntymäki in their Responsible AI seminar stated that the most important issues to explore when considering responsible AI are fairness, accountability, transparency and explainability. Mika Nieminen in his lecture on responsible technology submitted that the most important things in bringing ethical guidelines into action are transparency and discussion. (Birkstedt& Mäntymäki, 28.10.2020, Nieminen, 20.11.2020). University of Helsinki introduced open education course on AI ethics in 2020. In this course they have valued five principles of ethics that are relevant in regarding to AI. These are, the principle of beneficence/ non-maleficence, accountability, transparency, fairness and respect of human rights. As we can see, the principles that are recurring in all discussion on AI ethics are fairness, transparency and accountability. We can also see the difference when the discussion is on responsible AI and when it is on AI ethics. AI ethics seems to bring the human rights and use of AI in good into spotlight while when discussion is on responsible AI the target is more focused on the issues that might be technically resolved. However, it can be stated that intention for good and human rights are really broad matters to discuss and are not always clear. For example if we try to prevent sickness by using artificial intelligence, our intention is to promote good, but at the same time we might have issues with for example with privacy. All the ethical guidelines designed can be found in table form in Appendix 3.

8.5.1 Ethical guidelines for SHAPES project regarding accountability

Even though machine ethics considers artificial intelligence at least partially as a moral agent, in the real life we cannot at least not yet hold artificial intelligence and it's algorithms accountable for it's decisions. Especially in health care project like SHAPES the decisions that AI makes must be a subject of human evaluation. Accountablity is strongly linked to the concept of responsibility. As we earlier discussed in machine ethics, philosophically moral responsibility requires that the moral agent is concious of its own actions. This means that the agent should be able to evaluate and predict the consequences of its actions. This means that AI at its current state cannot be held responsible of its own actions in moral sense. It is safe to say that human being must be accountable for all decisions that artificial intelligence makes. However, it is not always easy to set the criteria how the responsibility should be individualised. For example, we cannot name a person responsible on all effects of AI on society level. Although when we discuss on SHAPES project we must be ready to at least on the project level to have a specific responsible person.

One of the European Commission's requirements is human oversight. In most guidelines this is linked to accountability. Human-in-the loop or human-on-the-loop approaches are mostly discussed, but also ethics-by-design can be seen as a resolution into the dilemma. Ethics-by design does not remove human oversight, but it gives boundaries to AI to make decisions without human oversight. When the decision is not within these given values, AI system cannot make the decision automatically on its own.

The guidelines for SHAPES project:

1) Name responsible person for all stages of AI development and use. Make it easy to contact when AI decisions are unexpectable.
2) AI system should be developed with ethics by design-approach to provide ethical boundaries that within the AI system can make decisions.
3) Decide when the human-in-the-loop and human-on-the-loop-approaches are necessary. Keep in mind that information and decision-making in SHAPES includes health data.
4) Make it easy to overtake AI when it seems to generate unfamiliar decisions.

8.5.2 Ethical guidelines for SHAPES project regarding transparency and explainability

Transparency itself is not an ethical issue. It is just a matter of making operations and decision-making of the AI system visible. What can be seen as an ethical issue is that how much information the developers and operators of AI system are to give to the stakeholders and users. As users of the AI system, do we need to know that we are engaging with artificial intelligence and do we need to know exactly how the systems works? What is the sufficient information? Transparency links to human rights. The user has right to know on what grounds the decision is made and how the information is used in decision-making. For example, does the decision-making of the AI system violate privacy or right for self-determination if it tracks the movements of the ageing person at home when the purpose of this is to prevent insuries? The user has to be aware what information AI system is using and how it is used in decision-making. If this information is not available, the decision-making of the system is not easy to be argued or disagreed.

We also need to discuss on what level of understanding of the algorithmic decision-making is sufficient. This is a matter of explainability, which is mentioned in almost all ethical guidelines. It might be because it is under discussion that what level of understanding of the AI system is sufficient that we can use the AI system. Transparency can be devided into three components, simulatability (the understanding of how the model works), decomposability (understanding of the algorithmic components) and algorithmic transparency (visibility of algorithms). (University of Helsinki, 2020.) The discussion is on do the users need to have understanding or visibility on all three levels of the AI system or what level of understanding the developers and those who maintain the AI system must have? Can AI system be a "black box", the system so complex that no one has full understanding on what basis the decisions are made?

There has been discussions on making algorithms of the AI systems openly available. However, this will not solve the problem of comprehensiveness since the most users cannot understand the algorithms used in their coded form.

SHAPES-project needs to take into account the subject group, the ageing. The ageing people of the day do not always have the necessary IT-skills to understand algorithmic decision-making. This is why it is crucially important to make the AI system used transparent to all people, so that the also the health care professionals or family of the elderly can review the decision-making process.

The guidelines for SHAPES project:

1) Design the AI processes reviewable and give the user the information that the decision is based upon.
2) Enable the user to access an explanation of why the AI system behaved as it did.
3) Design models for users and use visualization when communicating the algorithmic decision-making to stakeholders. Have the ageing people in mind.
4) Inform openly ways of using AI and be open for discussion of the working methods and development of AI.
5) Keep records of the development process and the decision-making.
6) Offer possibilities to users and other stockholders to give feedback.
7) Minimize the risks that might appear from transparency for example security risks.
8) AI's decision-making process should be explainable. The technology behind AI must be understandable for some, but how it comes to conclusions must be understandable to the most.
9) The user must have a choice whether he/she acts with a machine or a human being.
10) Make it easy to decide and interact.
11) Guide through the process.
12) The user must be always aware when and how he/she is interacting with AI.

8.5.3 Ethical guidelines for SHAPES project regarding Diversity, Inclusion and Fairness

Diversity, inclusion and fairness are ethical values that are not always easy to find solutions to. Regarding artificial intelligence, it can marginalise the people who are already in vulnerable position even more when they cannot access the resources that are needed to be able to benefit AI. On the other hand, artificial intelligence has great potential to promote well-being for these groups if they are able to access the resources. In SHAPES the ageing are especially vulnerable group since not only they have difficulties and loss of abilities that ageing brings, but also that their ability to understand and use technology might be limited.

If the ageing have less competences to use digital technology and computers, the women in general are less likely to have these competences, the ageing women are at even greater risk to be excluded from technological evolution. One concern in AI development is also the gender devide. Only 12 per cent of the machine learning researchers are women. When the AI systems are developed and designed by men, there might be a risk for the gender-bias. Also when men are more likely to adopt and use new technologies the solutions are more likely to be designed for them. When designing for elderly, it is crucial that the development process includes older people from both genders. (University of Helsinki, 2020.)

SHAPES project is pan-European project. Even though all the countries involved share common western value base, the cultural differences between regions are still noticeable. If the cultural differences go unnoticed in the development of AI systems, we might be faceing societal problems that we were unaware of. Being aware that cultural differences exist and being sensitive to the fact that the bias might arise from upon the differences, is in the heart of the development of AI systems and specifically important factor to track when the AI system is in use.

One of the most important ethical issue to solve is fairness and discrimination. First of all it is important to define discrimination. Usual definition for discrimination is that people belonging into certain group are treated differently from the people not belonging to this group. (University of Helsinki, 2020.) It should be noted that this definition does not exclude possibility to positive discrimination. Another definition of discrimination is diffent treatment of the group by perceived membership that is causing social harm to this group (University of Helsinki, 2020). This definition sees discrimination always harmful. Discussing discrimination brings us to discuss the concept of equality. Equality by definition means that everyone is treated the same way with no exceptions. But this usually means that the most vulnerable groups do not have the same possibilities than the groups with more power and resources. When discussing AI ethics it is especially important to notice this. Equality might not always the best solution to promote equity. It is important to discuss how the fairness and equity is achieved without risking equality.

When discussing fairness, one of the issue arising is bias. Chavalarias and Ioannidis find in their study 235 biases that could affect research. The same kinds of biases can occur similarily when training algorithms if the data-sets used are imbalanced. Bias itself can be either positive or negative, but in the case of ethics the more problematic is the case of negative bias since the positive bias might even be at least in some cases what is hoped for. Bias can be particularly problematic when it is not visible or when people are not aware of its existance. This is called implicit bias. For example if the medical doctor has a different views of racial groups, it might effect physician's decisions of the treatment. Bias is especially troubling if it is affecting engineering and design. It is worth noticing that there is already action to counteract this bias in design. Universal design is a good attempt to develop things with that way that it is accessible for everyone. (Howard & Borenstein, 2018, 1521-1523.)

Algorithms of AI are not immune to our society's biases. As they find patterns in the datasets and their learning is based on data, they are not free from our stereotypes and other discriminatory behaviour when we use the data that our society produces. For example, facial recognition has been struggling to recognize non-caucasian faces throughout its history. After ten years of developing AI systems, it still has trouble of recognizing Asian-origin faces or the faces of the ageing women.

How should we prevent AI systems to become prejudiced? The most effective way is to pay attention to the composition of the expert group that is developing and teaching AI. The more diverse the group is the more there are different views and recongnition. The selection of datasets is crucially important. How we find a secure dataset that does not promote the same human prejudice and how do we even define the desired dataset and outcome?

There are a few measures that are attempts to tackle the problem of bias, but they all have also some disadvantages. For example anticlassification means removing the data that is causing issues from the dataset e.g. gender or race. In some cases this approach might cause more issues. For example if we remove gender and race from the health data, we might lose the information that is relevant for the decision-making for those groups.

1) Keep in mind the norms, values, experiences and gains of the user group and value the positive in them.
2) Be aware of the cultural differences, gender and age and be sensitive not to bias upon them. Use "design for all" if possible.
3) Review carefully the data used and use ongoing research and diverse data collection to minimize algorithmic bias.
4) If the bias is detected, investigate carefully to understand from where the bias is originated and how it can be removed.
5) Collect feedback from the users.
6) Create and use check lists which promote diversity.
7) Review and create possibilities to empower the groups that are more likely to be excluded from the resources.

8.5.4 Ethical guidelines for SHAPES project regarding Safety and Security

Safety and security are not ethical matters as such, but when considering them against human rights they might raise some ethical discussion. When the goal is to protect individuals from the social, emotional and physical harm the safety and security are the norm. Requirement of safety in designing artificial intelligence systems is an obligation, not an option. The system must not only be safe and secure when it is working as expected, but it must be also safe and secure for users when something unconventional happens.

One area of safety discussion is that how can we produce safety with artificial intelligence. For example in health care, the issue could be creating safe environment for the ageing people. This could be surveillance systems at home that detect falling or safe routes for walking for people who suffer dementia.

Safety and security can be also be seen as technical safety and security. It is strongly linked into privacy and data protection, that are not discussed further here though.

In SHAPES project also the user group, the ageing people, must be taken into account. They might have less experience of using IT-systems and have less ability to understand the security and safety measures needed. This means that the system should be created that way that way that it needs as little as possible user activity to be safe and secure to use.

1) Safety and security must be taken into account in all phases of development and tested thoroughly before release. Provide documentation.
2) User group, the ageing, must be taken into account.
3) Promote easy access and conventions familiar to the ageing population.
4) Promote one user interface whenever possible.
5) Protect individuals from social, physical and emotional harm.
6) Create security measures also for the situations that unconventional action or malfunction happens.
7) Discuss openly the limits of the robustness, security and safety.
8) Set boundaries on which the AI system can work independently.
9) Discuss on what grounds AI can be used in health care. For example, must AI make more reliable decisions than human being or should we allow some false decisions?

8.5.5 Ethical guidelines for SHAPES project regarding Societal well-being and humanity

One of the essential part of the SHAPES project is to find solutions that help ageing people to continue to live at home. At least in Northern Europe the loneliness of the elderly is an immence problem. For many older individual the health care personnel might be the only human contact they have. When developing AI based solutions for their benefit, it is important that we do not replace the human contacts with technology. The aim should be that AI solutions created assist and support health care professionals so that their time with the elderly can be used more on human to human contact that on for example administrative tasks. AI solutions should also not only be developed for surveillance or gathering the health data on ageing people, the focus should be on wellbeing of the individual as a whole. For example solutions should promote maintaining physical and social activity and promote ageing people to remain active also outside their homes.

1) Design and develop the AI systems in that way that human contact for the elderly will remain or increase.
2) Aim for diminishing loneliness. Promote social contacts between generations and peer groups.
3) Promote prevention in all forms.
4) Promote solutions that maintain functional capacity or increase activity
5) Development should aim to promote or maintain the relationships with nature and environment. Solutions should promote outdoor activites.
6) Development should aim for common good and benefit humanity

8.6 Ethical Competence

It was earlier discussed about the McNamara et al.'s study where they found no effect on practises when the ethical guidelines were introduced to software developers. In the light of this I decided that it was not enough to this project to have only ethical guidelines, but more was needed. The one approach I found is the concept of ethical competence.

When considering the development of the SHAPES project and impacts it has, it can be noted that very different perspectives have to be taken into account when the impacts are considered. A mere technological or social analysis is not enough, but we also have to take into account the views of the target group, such as what is considered as good ageing and whether technology can replace human care or reduce exclusion or loneliness, for example. Perspectives can also be contradictory, what could be effective and desirable for society may not be desirable for the individual. Assessing ethical implications is not straightforward, as ethical values themselves are already diverse and different arguments lead to different factors being included in ethical values. In ethical problems, there is often not just one right answer, but often the solutions are incomplete. The study of the ethics of artificial intelligence is not well-established and has many features that influence ethical evaluation, even if they are not really matters of ethics. An example of this is information security.

I think the subject is interesting to study, as it can be assumed that technical developers already inherently have some kind of code of ethics, for example with regard to respect for human rights. It is important to measure whether the guidelines can guide developers to make better ethical decisions, or do the guidelines, for example, influence developers to take into account only the ethical aspects included in the guidelines and to ignore other emerging ethical issues? We earlier discussed Katie Shilton’s ethnological research of technicians that were aware of the ethical issues, but thought that they are the job for someone else. Often ethical guidelines see AI ethics as a technical issue and usually the problem-solving is attempting to find a technical solution to ethical issues than trying to solve the economical or societal issues of the larger scale.

8.6.1 The concept of Ethical Competence

In health care in particular, ethical competence has been a concept for some time. In other fields, ethical competence has only been raised a little later, and the challenges posed by artificial intelligence, for example, have increasingly brought ethical thinking to the field of technology as well. Ethical competence can be said to consist of six components, ethical sensitivity, awareness, reflection (reflection), decision-making, action, and behavior. Ethical sensitivity is the ability to identify ethical problems. It is a prerequisite for all ethical reflection, as the identification of a problem or conflict is a prerequisite for action and decision-making. Ethical awareness, on the other hand, is an individual’s theoretical and philosophical as well as practical ethical competence, which creates a frame of reference for decision-making and reflection. Reflection, on the other hand, is a holistic reflection on an ethical problem and solution options that takes into account the perspectives, values, and beliefs of those involved. Ethical decision-making is a process that seeks to make a responsible and sensible decision among a number of different options. (Lechasseur, 2016.)

In recent years, indicators have been developed to identify and support ethical competence, which can be divided into two, counseling and guidance-based and reflective methods. These methods have mainly been developed for use by healthcare professionals. Counseling and guidance are mainly the use of external experts to find ethical solutions and support decision-making. Reflective methods, on the other hand, are based on supporting the ethical choices of health care professionals and strengthening ethical competence. Reflective methods have been found in previous studies to increase collaboration and interaction between actors and to increase their ethical sensitivity. These methods include Ethics rounds and Moral Case Deliberation (MCD). The purpose of these methods is to help actors identify and act in various ethically challenging and problematic situations. (Nikunen, 2018.)These methods are not directly methods of ethical effectiveness. However, they are methods aimed at increasing the ethical actions of operators. It can be said that an evaluation of the effectiveness of these methods would be appropriate, but so far I have not found such a study.

Moral Case Deliberation (MCD) is the method used to face ethical dilemmas in healthcare. It is used mostly on cases when the professional ethics and guidelines are not giving a certain anwer to the dilemma at hand. MCD has been developed to support the staff that faces moral dilemmas. This framerork provides support in a structured way in discussion groups led by experienced facilitator. The aim is to have open and diverse discussion on the ethical issue and this way to develop ethical competence of the participants. (Tan, Ter Meulen, Molewijk & Widdershoven, 2018, 181.)

Ethics rounds is a very similar method which also uses discussion on ethical issues that health care professionals face during their work. In their study Marit Silen, Mia Ramklint, Mats G. Hansson and Kristina Haglund found out that the health care professionals were very positive about Ethics Rounds, even though their work stress or work satisfaction remained the same level than before the Ethics Rounds were introduced. However, they experienced that Ethics Rounds helped them to broaden their thinking and to see the situation from the different perspectives. In this study the Ethics Rounds did not improve the ethical climate of the hospital in question. (Silén et al. 2016, 203.)

8.6.2 Ethical competence in SHAPES

In the case of SHAPES project my view is that raising ethical awareness and competence is crucial. We know that the field of AI is rapidly changing and the solutions are more and more complex, we cannot be certain what kind of ethical issues we will be facing. Then the only way to prepare the developers and other professionals is to make them aware that they might be facing moral dilemmas on the development journey. The ethical competence is important from view of the ethical contradiction. The solution might tick some ethical checklist boxes, but still have a contradictions between them. For example, the AI solution might aim for common good, but at the individual level, it might violate privacy or self-determination right. Raising ethical competence of all stakeholders and developers makes ethical matters visible and promotes discussion.

As a methods to raise ethical competence of the developers I would suggest regular ethics trainings. These trainings could be for example web-based trainings that include common ethical awareness sessions and specific trainings of the issues that are relevant in development process of AI systems. These could also include use cases and best practises.

Another way of raising ethical awareness among stakeholders is regular ethical rounds. In these sessions the developers can discuss the real life ethical issues that they are facing in a group situation that is facilitated. This method is especially fruitful in situations when there are contratictions between the ethical guidelines or when a new ethical issue appears. It also can be interesting to engage technical developers into wider ethical discussion since the societal and economical issues can be difficult to produce in the form of simple ethical guidelines. When participating in wider ethical discussion on how the artificial intelligence effects the society and what kind of issues it might have on macro level, might help developers to see new and different solutions to these issues or even focus the development to different areas.

Katie Shilton's study that we discussed earlier brought into light that the developers were aware of the ethical issues that might be related to the system they were developing, but they felt that the handling those ethical issues were not their job. (Shilton, 2012.) In this project it is important that we can assure developers that ethical issues are responsibility of everyone involved and that if the system has an ethical issue, it must be resolved before continueing the development of the system.

8.7 Towards ethics for artificial intelligence

Artificial intelligence has strongly emerged into discussions in past ten years. Rapid advancement of AI and especially neural networks has emerged from three grounds. Firstly, the cost of physical hardware and computing memory power has been decreasing while computing capacity has been rapidly increasing. Secondly, the amount of information available has become enormous. Digitalisation and different kinds of IoT solutions produce a massive data archive. Thirdly, AI has matured to the point that it is not only a theoretical subject matter but there are also practical solutions that are using AI. This also means that more educational materials produced and shared. (Kananen & Puolitaival, 2019, 35-36.)

AI effects society broadly and in-depth. All effects of AI are not easily estimated and the benefits and risks might be exaggerated or underestimated depending on who is speaking. Rapid emergency of AI has generated discussion about communication between machines and humans, which generates discussion of what grounds AI operates and what kind of ethical background is applied.

New technologies can however amplify or deepen already existing ethical problems or create new ones. Considering AI, we most definitely must examine the ethics of technology, but since AI is utilised in multiple fields of study, we need the consider studying specific ethical matters of the field as well. For example, when AI is applied in a field of health care, we must consider the special ethical issues involving this field.

In literature there is several approaches to the ethics for AI. The most common one is to examine the ethics for AI via standard ethical theories. This means that the examination of ethics for AI is not required to start from the beginning, but it can be located as part of the theoretical tradition. In this study I will not go deeply into the philosophical traditions of ethics. I will only conclude that ethics of artificial intelligence is bringing the issues of morality, moral code and good and right ways of action into focus. It forces us to consider the boundaries of humanity in a way that we have not contemplated ever before. What is the point when the machine might have a consciousness? Or can machine be held responsible? These are the very questions we need answers for sooner than we might think, but in this study, I will take a bit more hands-on approach since the project needs more pragmatic ethical background. However, all the aspects considered here are leading towards answering the questions about humanity and integrity.

9 Conclusions

The purpose of this study was to review the studies on ethics of the artificial intelligence in comparison to SHAPES project and to specify the special features of the project. Another goal was to design ethical guidelines for the project AI developers. Soon it was discovered that there are theoretical knowledge on artificial intelligence itself and ethics, but there are not too much literature directly on the ethics of the artificial intelligence. This meant that theoretical knowledge was gained by studying common ethical theories, human rights and machine ethics to be able to have theoretical background for the studying and designing ethical guidelines.

The ethical guidelines for the SHAPES project were designed by using Alan Hevner's Design Science Research Approach. Theoretical background and environment were leading the design process in which the ethical guidelines were designed through the iteration process. Environment in this case included studying the specific requirements that the target group, the ageing, and the SHAPES ecosystem being a health care system brought into using artificial intelligence.

It was also useful to study other ethical guidelines for the artificial intelligence and compare the different approaches on the ethical concerns in that way. Also the webinars on the subject of ethics of the AI were giving perspective on the matter to be able to find the most important matters for the guidelines for the SHAPES project.

The guidelines like European Commission's guidelines for trustworthy AI are important since they are good way to create standard prosedures for development and use of artificial intelligence. Sharing the best practices and technical solutions for ethically behaving AI are suited for creating standards and policies that are less rigor than binding legistlation. This enables flexibility for future solutions that we are not yet able to foresee. The guidelines promote self regulation of the field.

It must be said, that legistlation is to the point a good way to regulate the development and use of AI. For example GDPR has given boundaries by obligation to data security and privacy, but it is important to regognize that the legistlation might be also limiting. It is also important to notice that it is crucial to concentrate on solving the ethical issues where there is no legistlation or regulation.

It can be argued whether the ethical guidelines are the best way to produce ethical awareness. It might be better to have a some sort of combination of ethical guidelines, ethical training and promoting ethical competence by ethics rounds or some other methods that use real cases that arise from the work of the developers. Also the technical methods like ethics-by-design or values-by-design are important when considering making artificial intelligence to act responsibly. One subject for the further studies could be to study the effectiveness of the ethical guidelines or to compare different methods of promoting ethical awareness.

This study has not field tested ethical guidelines with the developers of artificial intelligence of the SHAPES project. It would be important to continue this study by field testing and gaining feedback from the developers and let the feedback mold the ethical guidelines that in this study are formed mostly based on theoretical knowledge and other guidelines that already exist. This is also one of the weaknesses of this study. The real life experiences of development of tha artificial intelligence, especially in health care, are not included.

The problem with ethical guidelines the are designed for the developers is the issue of the level of the ethical concerns. Many ethical issues related to artificial intelligence are not on the level of the individuals, but on the level of the society as a whole. Artificial intelligence has possibility to change our societies in ways that we cannot yet even imagine. This means that to solve ethical concerns related to AI should be discussed widely on societal level.

It is important to discuss also the possibility that AI might be able to make more accurate and less biased decisions without human intervention. Do we want AI to make decisions like human being even though we know that humans make bad decisions or do we want to AI make decisions like a machine can even though we do not always understand on what basis the AI has reached the decision? How we know it is a good decision? More discussion is needed when human intervention or guidance is needed and when there is robust enough technical solutions to let AI work independently without human intervention or decision-making.

10 References

Anderson, Michael and Anderson, Susan Leigh (Eds): Machine Ethics. 2011. Cambridge University Press.

Brey, P. 2012. Anticipatory Ethics for Emerging Technologies. NanoEthics, 6(1), pp. 1-13. doi:10.1007/s11569-012-0141-7.

Cabrera, M., Cabrera, M., Malanowski, N. & Malanowski, N.: Information and Communication Technologies for Active Ageing - Opportunities and Challenges for the European Union 2009.

Dignum, Virginia, Baldoni, Matteo, Baroglio, Cristina, Caon, Maurizio, Chatila; Raja, Dennis, Louise, Génova, Gonzalo, Kliess, Malte, Lopez-Sanchez, Maite, Micalizio, Roberto, Pavón, Juan, Slavkovik, Marija, Smakman, Matthijs, van Steenbergen, Marlies, Tedeschi, Stefano, van der Torre, Leon, Villata, Serena, de Wildt, Tristan, Haim, Galit: Ethics by Design: necessity or curse? (Agre , 1997)https://www.aies-conference.com/2018/contents/papers/main/AIES_2018_paper_68.pdf

European Commission’s High-Level Expert Group on Artificial Intelligence: ETHICS GUIDELINES FOR TRUSTWORTHY AI. 2019.

Gassmann, Oliver and Keupp, Marcus M.: The "Silver Market in Europe": Myth or Reality? Assistive Technology Research Series 23:77-90. 2009. DOI 10.3233/978-1-58603-937-0-77

Gilhooly, Mary L M, Gilhooly, Kenneth J & Jones, Ray B.: Quality of Life: Conceptual Challenges in Exploring the Role of ICT in Active Ageing. Information and Communication Technologies for Active Ageing 49 M. Cabrera and N. Malanowski (Eds.).IOS Press. 2009 DOI 10.3233/978-1-58603-937-0-49.

Gregor, S. & Jones, D: The Anatomy of a Design Theory. Journal of the Association for Information Systems, 8(5), pp. 312-323,325-335. 2007. DOI 10.17705/1jais.00129

Hagendorff, T. The Ethics of AI Ethics: An Evaluation of Guidelines. Minds And Machines, 30(1), pp. 99-120. 2020. DOI 10.1007/s11023-020-09517-8

Hausknecht, Simone, Schell, Robyn, Zhang, Fan and Kaufman, David: Older Adults Digital Gameplay: A Follow-up Study of Social Benefits in in Markus Helfert, Andreas Holzinger, Martina Ziefle, Ana Fred, John o'Donoghue, Carsten Röcker (Eds.): Information and Communication Technologies for Aging Well and e-Health. First International Conference, ICT4AgeingWell 2015 Lisbon, Portugal, May 20-22, 2015 Revised Selected Papers. 2015. Springer International Publishing Switzerland.

Hevner, Alan & Chatterjee, Samir: Design Research in Information Systems Theory and Practice. Springer New York. 2010. DOI 10.1007/978-1-4419-5653-8.

Hevner, Alan, March, Salvatore T., Park, Jinsoo & Ram, Sudha: DESIGN SCIENCE IN INFORMATION SYSTEMS RESEARCH 1. MIS Quarterly, 28(1), pp. 75-105. 2004. DOI 10.2307/25148625.

Howard, Ayanna & Borenstein, Jason: The Ugly Truth About Ourselves and Our Robot Creations: The Problem of Bias and Social Inequity. Sci Eng Ethics. 24:1521–1536 .2018 https://doi.org/10.1007/s11948-017-9975-2

Hursthouse, R. : On Virtue Ethics. Oxford University Press, Incorporated. 2000.

Huyck, Christian, Augusto Juan, Xiaohong, Gao and Botia, Juan A.: Advancing Ambient Assistes Living with Caution in Markus Helfert, Andreas Holzinger, Martina Ziefle, Ana Fred, John o'Donoghue, Carsten Röcker (Eds.): Information and Communication Technologies for Aging Well and e-Health. First International Conference, ICT4AgeingWell 2015 Lisbon, Portugal, May 20-22, 2015 Revised Selected Papers. 2015. Springer International Publishing Switzerland.

Kaakinen, Juha & Törmä, Sinikka: ESISELVITYS GERONTEKNOLOGIASTA Ikääntyvä väestö ja teknologian mahdollisuudet, TULEVAISUUSVALIOKUNNAN TEKNOLOGIAJAOSTOTEKNOLOGIAN ARVIOINTEJA 5, https://www.eduskunta.fi/FI/naineduskuntatoimii/julkaisut/Documents/ekj_2+1999.pdf Referred 7.5.2020

Kaaslainen, Karoliina, Ruohonen, Toni & Neittaanmäki, Pekka:Interventiot ja tekoäly terveydenhuollossa. Loppuraportti vol.3. Jyväskylän yliopisto.Yliopistopaino. Jyväskylä. 2019.

Kananen, Heidi & Puolitaival, Harri:Tekoäly-Bisneksen uudet työkalut. Alma Talent. Helsinki. 2019.

Kasanen, E., Lukka, K. & Siitonen, A. 1993: The constructive approach in management accounting research. Journal of management accounting research, 5, p. 243.

Knobel, C. & Bowker, G.: Computing Ethics: Values in Design. Association for Computing Machinery. Communications of the ACM, 54(7), p. 26. 2011. DOI 10.1145/1965724.1965735

Lechasseur, K., Caux, C., Dollé, S. & Legault, A. 2018. Ethical competence: An integrative review. Nursing Ethics, 25(6), pp. 694-706. doi:10.1177/0969733016667773.

McNamara, A., Smith, J., Murphy-Hill, E. (2018). Does ACM's code of ethics change ethical decision making in software development?" In G. T. Leavens, A. Garcia, C. S. Păsăreanu (Eds.) Proceedings of the 2018 26th ACM joint meeting on european software engineering conference and symposium on the foundations of software engineering—ESEC/FSE (pp. 1–7). 2018. New York: ACM Press.

Moor, James H.: The Nature, Importance, and Difficulty of Machine Ethics in Anderson, Michael and Anderson, Susan Leigh (Eds): Machine Ethics. 2011. Cambridge University Press.

Neittaanmäki, Pekka, Tuominen, Heli, Äyrämö, Sami, Vähäkainu, Petri & Timo Siukonen (toim.):Tekoäly ja terveydenhuolto Suomessa. Loppuraportti vol 1. Jyväskylän Yliopisto. Yliopistopaino. Jyväskylä. 2019.

Neubert, M. J. & Montañez, G. D. Virtue as a framework for the design and use of artificial intelligence. Business Horizons, 63(2), 2019. DOI 10.1016/j.bushor.2019.11.001

Nevanperä, M., Rajamäki, J. & Helin, J. :Design Science Research and Designing Ethical Guidelines for the SHAPES AI Developers. Procedia computer science, 192, pp. 2330-2339. doi:10.1016/j.procs.2021.08.223. 2021.

Nikunen, Outi: Eettisen keskustelun koulutus terveydenhuollon henkilöstön eettisen osaamisen tukena: Kyselytutkimus osallistujille. Pro gradu -tutkielma. Itä-Suomen yliopisto. 2018.

Ojasalo, Katri, Moilanen, Teemu & Ritalahti, Jarmo: Kehittämistyön menetelmät-Uudenlaista osaamista liiketoimintaan. 3.-4. painos. Sanoma Pro. Helsinki. 2015.

Omari, R. M. & Mohammadian, M. Rule based fuzzy cognitive maps and natural language processing in machine ethics. Journal of Information, Communication and Ethics in Society, 14(3), pp. 231-253. 2016. DOI 10.1108/JICES-10-2015-0034.

Rajamäki, J., Sarlio-Siintola, S, Alapuranen, N. & Nevanperä, M.:Privacy and data protection in Open Source Intelligence and Big Data Analytics: Case ‘MARISA’ in Karoliina Nikula, Sari Sarlio-Siintola & Valdemar Kallunki (eds.) Ethics as a resource Examples of RDI projects and educational development. Laurea Julkaisut 144. p.23-29. 2020.

Shilton, K. Values Levers: Building Ethics into Design. Science, Technology, & Human Values, 38(3), pp. 374-397. 2013. DOI 10.1177/0162243912436985

Silén, M., Ramklint, M., Hansson, M. G. & Haglund, K. 2016. Ethics rounds: An appreciated form of ethics support. Nursing ethics, 23(2), pp. 203-213. DOI 10.1177/0969733014560930.

Swanton, Christine: Virtue Ethics: A Pluralistic View. Oxford University Press, Incorporated. 2003.

Quinn, Michael J.: Ethics for the Information Age. Pearson Education Limited. 2015. Harlow.

H.J.Toh, J. Chia, E. Koh, K. Lam, G.C. Magpantay, C.M. De Leon and J.A. Low: Virtual Geriatric Care: User perception of telegeriatrics in nursing homes of Singapore in Markus Helfert, Andreas Holzinger, Martina Ziefle, Ana Fred, John o'Donoghue, Carsten Röcker (Eds.): Information and Communication Technologies for Aging Well and e-Health. First International Conference, ICT4AgeingWell 2015 Lisbon, Portugal, May 20-22, 2015 Revised Selected Papers. 2015. Springer International Publishing Switzerland.

Tan, D., Ter Meulen, B., Molewijk, A. & Widdershoven, G. 2018. Moral case deliberation. Practical Neurology, 18(3), p. 181. DOI 10.1136/practneurol-2017-001740.

Van Den Hoven, J., Lokhorst, G. & Van De Poel, I. Engineering and the problem of moral overload. Science and engineering ethics, 18(1), p. 143. 2012. DOI 10.1007/s11948-011-9277-z

Viirkorpi, P.: Ikäteknologian hyvät käytännöt. Helsinki: Vanhus- ja lähimmäispalvelun liitto ry. 2015.

Walker, Alan: The Future of Ageing in Europe-Making an Asset of Longevity. 2019. https://doi.org/10.1007/978-981-13-1417-9.

Electronic

Artificial Intelligence at Google: Our Principles. https://ai.google/principles/ . Referred 28.4.2020.

Deloitte Belgium. The Socio-Economic Impact of AI in Healthcare. 2020.

Everyday Ethics for Artificial Intelligence, IBM design program office. 2019. https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf Referred 27.4.2020

Google: *Artificial Intelligence at Google: Our Principles.* https://ai.google/principles/ . Referred 28.4.2020.

https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1%3aprimaryr6

https://ec.europa.eu/health/md_sector/overview_en referred 31.1.2021

https://ec.europa.eu/health/policies/overview_en referred 6.2.2021

https://ec.europa.eu/health/md_sector/overview_en referred 6.2.2021

GDPR.EU https://gdpr.eu/eu-gdpr-personal-data/?cn-reloaded=1. Referred 21.12.2020

https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en. Referred 19.1.2021

IEEE: ETHICALLY ALIGNED DESIGN A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems. Referred 27.4.2020

https://shapes2020.eu/ Referred 14.3.2021

THL:https://thl.fi/fi/web/hyvinvointi-ja-terveyserot/tavoitteet/lait-ja-ohjelmat#terveydenhuoltolaki referred 6.2.2021

United Nations: https://www.un.org/en/universal-declaration-human-rights/

University of Helsinki: https://ethics-of-ai.mooc.fi/. Referred 20.12.2020

Unpublished:

Nevanperä, M., Helin J. & Rajamäki, J: Comparison of European Commission’s Ethical Guidelines for AI to Other Organizational Ethical Guidelines. 2021b. Conference paper for the 3rd European Conference on the Impact of Artificial Intelligence and Robotics.

Webinars:

Lehtimäki, Pasi & Seppänen, Anna: Tekoälyn etiikka: Analyyttisiä välineitä eettisiin haasteisiin. YKA-webinar. 29.10.2020

Nieminen, Mika: Kolme pointtia vastuullisesta teknologiasta. VTT. 20.11.2020

Birkstedt, Teemu & Mäntymäki, Matti: Responsible AI- An Emerging Business. University of Turku. TSE Alumni Webinar. 28.10.2020

Figures

Appendix 1: The European Commission's Ethical Guidelines, the Checklist

TRUSTWORTHY AI ASSESSMENT LIST (PILOT VERSION)

1. Human agency and oversight

Fundamental rights:

□ Did you carry out a fundamental rights impact assessment where there could be a negative impact on fundamental rights? Did you identify and document potential trade-offs made between the different principles and rights?

□ Does the AI system interact with decisions by human (end) users (e.g. recommended actions or decisions to take, presenting of options)?

□ Could the AI system affect human autonomy by interfering with the (end) user's decision-making

process in an unintended way?

□ Did you consider whether the AI system should communicate to (end) users that a decision,

content, advice or outcome is the result of an algorithmic decision?

□ In case of a chat bot or other conversational system, are the human end users made aware that

they are interacting with a non-human agent?

Human agency:

□ Is the AI system implemented in work and labour process? If so, did you consider the task allocation between the AI system and humans for meaningful interactions and appropriate human oversight and control?

□ Does the AI system enhance or augment human capabilities?

□ Did you take safeguards to prevent overconfidence in or overreliance on the AI system for work processes?

Human oversight:

□ Did you consider the appropriate level of human control for the particular AI system and use case?

□ Can you describe the level of human control or involvement?

□ Who is the “human in control” and what are the moments or tools for human intervention?

□ Did you put in place mechanisms and measures to ensure human control or oversight?

□ Did you take any measures to enable audit and to remedy issues related to governing AI autonomy?

□ Is there is a self-learning or autonomous AI system or use case? If so, did you put in place more specific mechanisms of control and oversight?

□ Which detection and response mechanisms did you establish to assess whether something could go wrong?

□ Did you ensure a stop button or procedure to safely abort an operation where needed? Does this procedure abort the process entirely, in part, or delegate control to a human?

2. Technical robustness and safety

Resilience to attack and security:

□ Did you assess potential forms of attacks to which the AI system could be vulnerable?

□ Did you consider different types and natures of vulnerabilities, such as data pollution, physical infrastructure, cyber-attacks?

□ Did you put measures or systems in place to ensure the integrity and resilience of the AI system against potential attacks?

□ Did you verify how your system behaves in unexpected situations and environments?

□ Did you consider to what degree your system could be dual-use? If so, did you take suitable preventative measures against this case (including for instance not publishing the research or deploying the system)?

Fallback plan and general safety:

□ Did you ensure that your system has a sufficient fallback plan if it encounters adversarial attacks or other unexpected situations (for example technical switching procedures or asking for a human operator before proceeding)?

□ Did you consider the level of risk raised by the AI system in this specific use case?

□ Did you put any process in place to measure and assess risks and safety?

□ Did you provide the necessary information in case of a risk for human physical integrity?

□ Did you consider an insurance policy to deal with potential damage from the AI system?

□ Did you identify potential safety risks of (other) foreseeable uses of the technology, including accidental or malicious misuse? Is there a plan to mitigate or manage these risks?

□ Did you assess whether there is a probable chance that the AI system may cause damage or harm to users or third parties? Did you assess the likelihood, potential damage, impacted audience and severity?

□ Did you consider the liability and consumer protection rules, and take them into account?

□ Did you consider the potential impact or safety risk to the environment or to animals?

□ Did your risk analysis include whether security or network problems such as cybersecurity

hazards could pose safety risks or damage due to unintentional behaviour of the AI system?

□ Did you estimate the likely impact of a failure of your AI system when it provides wrong results, becomes unavailable, or provides societally unacceptable results (for example discrimination)?

□ Did you define thresholds and did you put governance procedures in place to trigger

alternative/fallback plans?

□ Did you define and test fallback plans?

Accuracy

□ Did you assess what level and definition of accuracy would be required in the context of the AI system and use case?

□ Did you assess how accuracy is measured and assured?

□ Did you put in place measures to ensure that the data used is comprehensive and up to date?

□ Did you put in place measures in place to assess whether there is a need for additional data, for example to improve accuracy or to eliminate bias?

□ Did you verify what harm would be caused if the AI system makes inaccurate predictions?

□ Did you put in place ways to measure whether your system is making an unacceptable amount of inaccurate predictions?

□ Did you put in place a series of steps to increase the system's accuracy? Reliability and reproducibility:

□ Did you put in place a strategy to monitor and test if the AI system is meeting the goals, purposes and intended applications?

□ Did you test whether specific contexts or particular conditions need to be taken into account to ensure reproducibility?

□ Did you put in place verification methods to measure and ensure different aspects of the

system's reliability and reproducibility?

□ Did you put in place processes to describe when an AI system fails in certain types of settings?

□ Did you clearly document and operationalise these processes for the testing and verification of the reliability of AI systems?

□ Did you establish mechanisms of communication to assure (end-)users of the system’s reliability?

3. Privacy and data governance

Respect for privacy and data Protection:

□ Depending on the use case, did you establish a mechanism allowing others to flag issues related to privacy or data protection in the AI system’s processes of data collection (for training and operation) and data processing?

□ Did you assess the type and scope of data in your data sets (for example whether they contain personal data)?

□ Did you consider ways to develop the AI system or train the model without or with minimal use of potentially sensitive or personal data?

□ Did you build in mechanisms for notice and control over personal data depending on the use case (such as valid consent and possibility to revoke, when applicable)?

□ Did you take measures to enhance privacy, such as via encryption, anonymisation and aggregation?

□ Where a Data Privacy Officer (DPO) exists, did you involve this person at an early stage in the process?

Quality and integrity of data:

□ Did you align your system with relevant standards (for example ISO, IEEE) or widely adopted protocols for daily data management and governance?

□ Did you establish oversight mechanisms for data collection, storage, processing and use?

□ Did you assess the extent to which you are in control of the quality of the external data sources used?

□ Did you put in place processes to ensure the quality and integrity of your data? Did you consider other processes? How are you verifying that your data sets have not been compromised or hacked?

Access to data:

□ What protocols, processes and procedures did you follow to manage and ensure proper data governance?

□ Did you assess who can access users' data, and under what circumstances?

□ Did you ensure that these persons are qualified and required to access the data, and that they

have the necessary competences to understand the details of data protection policy?

□ Did you ensure an oversight mechanism to log when, where, how, by whom and for what

purpose data was accessed?

Traceability:

4. Transparency

□ Did you establish measures that can ensure traceability? This could entail documenting the following methods:

□ Methods used for designing and developing the algorithmic system:

o Rule-based AI systems: the method of programming or how the model was built;

o Learning-based AI systems; the method of training the algorithm, including which input data was gathered and selected, and how this occurred.

□ Methods used to test and validate the algorithmic system:

o Rule-based AI systems; the scenarios or cases used in order to test and validate;

o Learning-based model: information about the data used to test and validate. □ Outcomes of the algorithmic system:

o The outcomes of or decisions taken by the algorithm, as well as potential other decisions that would result from different cases (for example, for other subgroups of users).

Explainability:

□ Did you assess:

□ to what extent the decisions and hence the outcome made by the AI system can be understood?

□ to what degree the system’s decision influences the organisation’s decision-making processes?

□ why this particular system was deployed in this specific area?

□ what the system’s business model is (for example, how does it create value for the organisation)?

□ Did outcome that all users can understand?

you ensure an explanation as to why the system took a certain choice resulting in a certain

□ Did you design the AI system with interpretability in mind from the start?

□ Did you research and try to use the simplest and most interpretable model possible for the

application in question?

□ Did you assess whether you can analyse your training and testing data? Can you change and

update this over time?

□ Did you assess whether you can examine interpretability after the model’s training and

development, or whether you have access to the internal workflow of the model?

Communication:

□ Did you communicate to (end-)users – through a disclaimer or any other means – that they are interacting with an AI system and not with another human? Did you label your AI system as such?

□ Did you establish mechanisms to inform (end-)users on the reasons and criteria behind the AI system's outcomes?

□ Did you communicate this clearly and intelligibly to the intended audience?

□ Did you establish processes that consider users' feedback and use this to adapt the system?

□ Did you communicate around potential or perceived risks, such as bias?

□ Depending on the use case, did you consider communication and transparency towards other

audiences, third parties or the general public?

□ Did you clarify the purpose of the AI system and who or what may benefit from the product/service? □ Did you specify usage scenarios for the product and clearly communicate these to ensure that it is understandable and appropriate for the intended audience?

□ Depending on the use case, did you think about human psychology and potential limitations,

such as risk of confusion, confirmation bias or cognitive fatigue?

□ Did you clearly communicate characteristics, limitations and potential shortcomings of the AI system? □ In case of the system's development: to whoever is deploying it into a product or service?

□ In case of the system's deployment: to the (end-)user or consumer?

5. Diversity, non-discrimination and fairness

Unfair bias avoidance:

□ Did you establish a strategy or a set of procedures to avoid creating or reinforcing unfair bias in the AI system, both regarding the use of input data as well as for the algorithm design?

□ Did you assess and acknowledge the possible limitations stemming from the composition of the used data sets?

□ Did you consider diversity and representativeness of users in the data? Did you test for specific populations or problematic use cases?

□ Did you research and use available technical tools to improve your understanding of the data, model and performance?

□ Did you put in place processes to test and monitor for potential biases during the development, deployment and use phase of the system?

□ Depending on the use case, did you ensure a mechanism that allows others to flag issues related to bias, discrimination or poor performance of the AI system?

□ Did you establish clear steps and ways of communicating on how and to whom such issues can be raised?

□ Did you consider others, potentially indirectly affected by the AI system, in addition to the (end)-users?

□ Did you assess whether there is any possible decision variability that can occur under the same conditions?

□ If so, did you consider what the possible causes of this could be?

□ In case of variability, did you establish a measurement or assessment mechanism of the potential impact of such variability on fundamental rights?

□ Did you ensure an adequate working definition of “fairness” that you apply in designing AI systems?

□ Is your definition commonly used? Did you consider other definitions before choosing this one?

□ Did you ensure a quantitative analysis or metrics to measure and test the applied definition of fairness?

□ Did you establish mechanisms to ensure fairness in your AI systems? Did you consider other

potential mechanisms?

Accessibility and universal design:

□ Did you ensure that the AI system accommodates a wide range of individual preferences and abilities?

□ Did you assess whether the AI system usable by those with special needs or disabilities or those at risk of exclusion? How was this designed into the system and how is it verified?

□ Did you ensure that information about the AI system is accessible also to users of assistive technologies?

□ Did you involve or consult this community during the development phase of the AI system?

□ Did you take the impact of your AI system on the potential user audience into account?

□ Did you assess whether the team involved in building the AI system is representative of your target user audience? Is it representative of the wider population, considering also of other

groups who might tangentially be impacted?

□ Did you assess whether there could be persons or groups who might be disproportionately

affected by negative implications?

□ Did you get feedback from other teams or groups that represent different backgrounds and

experiences?

Stakeholder participation:

□ Did you consider a mechanism to include the participation of different stakeholders in the AI system's development and use?

□ Did you pave the way for the introduction of the AI system in your organisation by informing and involving impacted workers and their representatives in advance?

6. Societal and environmental well-being

Sustainable and environmentally friendly AI:

□ Did you establish mechanisms to measure the environmental impact of the AI system's development, deployment and use (for example the type of energy used by the data centres)?

□ Did you ensure measures to reduce the environmental impact of your AI system's life cycle?
Social impact:

□ In case the AI system interacts directly with humans:

□ Did you assess whether the AI system encourages humans to develop attachment and empathy towards the system?

□ Did you ensure that the AI system clearly signals that its social interaction is simulated and that it has no capacities of “understanding” and “feeling”?

□ Did you ensure that the social impacts of the AI system are well understood? For example, did you assess whether there is a risk of job loss or de-skilling of the workforce? What steps have been taken to counteract such risks?

Society and democracy:

□ Did you assess the broader societal impact of the AI system’s use beyond the individual (end-)user, such as potentially indirectly affected stakeholders?

7. Accountability

Auditability:

□ Did you establish mechanisms that facilitate the system’s auditability, such as ensuring traceability and logging of the AI system’s processes and outcomes?

□ Did you ensure, in applications affecting fundamental rights (including safety-critical applications) that the AI system can be audited independently?

Minimising and reporting negative Impact:

□ Did you carry out a risk or impact assessment of the AI system, which takes into account different stakeholders that are (in)directly affected?

□ Did you provide training and education to help developing accountability practices?

□ Which workers or branches of the team are involved? Does it go beyond the development phase? □ Do these trainings also teach the potential legal framework applicable to the AI system?

□ Did you consider establishing an ‘ethical AI review board’ or a similar mechanism to discuss

overall accountability and ethics practices, including potentially unclear grey areas?

□ Did you foresee any kind of external guidance or put in place auditing processes to oversee ethics and accountability, in addition to internal initiatives?

□ Did you establish processes for third parties (e.g. suppliers, consumers, distributors/vendors) or workers to report potential vulnerabilities, risks or biases in the AI system?

Documenting trade-offs:

□ Did you establish a mechanism to identify relevant interests and values implicated by the AI system and potential trade-offs between them?

□ How do you decide on such trade-offs? Did you ensure that the trade-off decision was documented? Ability to redress:

□ Did you establish an adequate set of mechanisms that allows for redress in case of the occurrence of any harm or adverse impact?

□ Did you put mechanisms in place both to provide information to (end-)users/third parties about opportunities for redress?

Appendix 2.Abstract Nevanperä, M., Helin J. & Rajamäki, J: Design Science Research and Designing Ethical Guide-lines for the SHAPES AI Developers. 2021.

Abstract

This article targets the design process of ethical guidelines for the SHAPES project (Smart and Healthy Aging through People Engaging in Supportive Systems) which is a H2020 Innovation Action project. The aim of the project is to build solutions that can make it easier for older individuals to live at home, such as, robots, wearables and sensor technologies that apply artificial intelligence (AI). The guiding method of the design process of ethical guidelines is Alan Hevner's Design Science Research. Theoretical background consists of a form of literature overview, which contains the most relevant ethical theories and research on AI ethics, machine ethics and human rights. This article introduces the process of building the ethical guidelines for the SHAPES project and further discussion if providing guidelines is the sufficient tool to developers to take ethical action in development of the AI systems. The SHAPES guidelines include the following themes; accountability, transparency and explainability, diversity, inclusion and fairness, safety and security and societal wellbeing and humanity.

Appendix 3.

Ethical Guidelines for the SHAPES Project
Name responsible person for all stages of AI development and use. Make it easy to contact when AI decisions are unexpectable.

AI system should be developed with ethics by design-approach to provide ethical boundaries that within the AI system can make decisions.
Decide when the human-in-the-loop and human-on-the-loop-approaches are necessary. Keep in mind that information and decision-making in SHAPES includes health data.
Make it easy to overtake AI when it seems to generate unfamiliar decisions.
Design the AI processes reviewable and give the user the information that the decision is based upon.
Enable the user to access an explanation of why the AI system behaved as it did.
Design models for users and use visualization when communicating the algorithmic decision-making to stakeholders. Have the ageing people in mind.
Inform openly ways of using AI and be open for discussion of the working methods and development of AI.
Keep records of the development process and the decision-making.
Offer possibilities to users and other stockholders to give feedback.
Minimize the risks that might appear from transparency for example security risks.
AI’s decision-making process should be explainable. The technology behind AI must be understandable for some, but how it comes to conclusions must be understandable to the most.
The user must have a choice whether he/she acts with a machine or a human being.
Make it easy to decide and interact.
Guide through the process.
The user must be always aware when and how he/she is interacting with AI.
Keep in mind the norms, values, experiences and gains of the user group and value the positive in them.
Be aware of the cultural differences, gender and age and be sensitive not to bias upon them. Use “design for all” if possible.
Review carefully the data used and use ongoing research and diverse data collection to minimize algorithmic bias.
If the bias is detected, investigate carefully to understand from where the bias is originated and how it can be removed.

Collect feedback from the users.
Create and use check lists which promote diversity.
Review and create possibilities to empower the groups that are more likely to be excluded from the resources.
Safety and security must be taken into account in all phases of development and tested thoroughly before release. Provide documentation.
User group, the ageing, must be taken into account.
Promote easy access and conventions familiar to the ageing population.
Promote one user interface whenever possible.
Protect individuals from social, physical and emotional harm.
Create security measures also for the situations that unconventional action or malfunction happens.
Discuss openly the limits of the robustness, security and safety.
Set boundaries on which the AI system can work independently.
Discuss on what grounds AI can be used in health care. For example, must AI make more reliable decisions than human being or should we allow some false decisions?
Design and develop the AI systems in that way that human contact for the elderly will remain or increase.
Aim for diminishing loneliness. Promote social contacts between generations and peer groups.
Promote prevention in all forms.
Promote solutions that maintain functional capacity or increase activity
Development should aim to promote or maintain the relationships with nature and environment. Solutions should promote outdoor activites.
Development should aim for common good and benefit humanity