DECIDE-VerA

Clinical DECIsion support system voor het carDiovasculaire risico managEment in de eerstelijnsgezondheidszorg: een Verantwoordelijk- en Aansprakelijkheidsperspectief

Formative ethical analysis

Version: 1.0

Date: 1 oktober 2023

Authors: André Krom[a] & María José Villalobos[b]

[a] Leiden University Medical Center, Department of Medical Ethics & Health Law

[b] National eHealth Living Lab (NeLL)

ZonMW project no. 08540122120004

Table of contents

List of abbreviations 3

Summary 4

1. Introduction 5

1.1 The DECIDE-VerA project 5

1.2 (Expanded) scope of the formative ethical analysis 6

1.3 Structure of this report 6

2. Capability Sensitive Design: key characteristics 6

2.1 A capability theory focused on human capabilities central to human flourishing 7

2.2 Value Sensitive Design 11

3. Capabilities that seem particularly relevant in case of the AI-CDSS DECIDE 15

3.1 Selecting relevant capabilities and functionings 15

3.2 Tacking stock 21

4. Combining Capability Sensitive Design with Guidance Ethics 22

4.1 Introduction 22

4.2 Guidance ethics workshops: main components 22

4.3 Guidance ethics workshops as empirical investigation 23

4.4 Combining guidance ethics workshops with Nussbaum's capability theory 23

5. Next steps 24

5.1 Ethical analysis in Phase II 24

5.2 Towards integrating the results of the ethical, legal and design analyses 24

Literature 26

List of abbreviations

AI-CDSS	AI-based Clinical Decision Support System
CDSS	Clinical Decision Support System
CSD	Capability Sensitive Design
DECIDE	Clinical DECIsion support system for the carDiovascular risk managEment in primary health care
DECIDE-VerA	Clinical DECIsion support system voor het carDiovasculaire risico managEment in de eerstelijnsgezondheidszorg: een Verantwoordelijk- en Aansprakelijkheidsperspectief
EGE	European Group on Ethics in Science and New Technologies to the European Commission
EU	European Union
UNESCO	United Nations Educational, Scientific and Cultural Organization
VSD	Value Sensitive Design
WHO	World Health Organization

Summary

This report presents the results of a formative ethical analysis related to an AI-powered clinical decision support system (AI-CDSS) that is being developed with the goal of better managing the risk of cardiovascular disease in patients under 50 years of age. It is part of Phase I of the ZonMw project "Clinical DECIsion support system voor het carDiovasculaire risico managEment in de eerstelijnsgezondheidszorg: een Verantwoordelijk- en Aansprakelijkheidsperspectief" (DECIDE-VerA). In parallel to this analysis, two other teams in the consortium have carried out analyses of the design and legal aspects relevant for the AI-CDSS DECIDE.[1]

The initial goal of the formative ethical analysis was to analyse the AI-CDSS DECIDE from the normative theoretical perspective of the so-called "capability sensitive design" framework. Specifically identifying which "capabilities" (as defined by Martha Nussbaum, 2000) are relevant for the AI-CDSS. A key characteristic of a capability approach is that it intends to improve people's real freedom to achieve well-being. These results are presented here.

In addition, we extended the formative ethical analysis to also include considerations and recommendations on:

- How to combine the normative theoretical analysis based on Capability Sensitive Design (Phase I), with the *empirical* ethical activities that are part of Phase I (interviews) and Phase II (guidance ethics workshops); and
- How the results of the formative ethical, design and legal analyses (Phase I) can be integrated during Phase II.

All expansions of the scope of the formative ethical analysis aim to contribute to a central aim of the project, namely to develop a truly interdisciplinary approach by *integrating* the results of the analyses of the ethical, design and legal aspects of the AI-CDSS DECIDE.

[1] In Phase I of the project the results of the formative analyses (legal, ethical, design) will be reported on *separately*. The results of all (further) analyses will be integrated in Phase II of the project.

1. Introduction

Clinical Decision Support Systems (CDSS) are tools that can assist healthcare professionals and patients in making better clinical and life-style decisions. However, up to this day, CDSS mostly focus on assisting healthcare professionals, limiting their potential to promote shared decision-making and strengthen patients' autonomy. Although CDSS are not new, digitisation and AI have greatly changed their underlying mechanisms and increased their abilities. We refer here to this type of AI-based systems as "AI-CDSS".

The design, implementation and adoption of AI-CDSS has proven challenging to the point that experts in the field consider that the current level of implementation of AI in healthcare is, in the best case scenario, modest. The nature of the barriers that limit the translation of AI to clinical settings is diverse. Barriers identified include the limited validation and lack of efficient integration into clinical workflows which is related to the insufficient engagement of end users (e.g., healthcare professionals and patients) (Yu et al., 2018; Topol, 2019). Similarly, a long list of ethical challenges are recurrently emphasized (EGE, 2018; WHO 2021; UNESCO 2021). Finally, regulatory gaps and legal insecurity, lack of (global) standards and technical challenges regarding data sharing have been recognised (Jiang et al., 2017). Efforts from the European Union (EU) to responsibly regulate AI applications are quickly taking form. For example the upcoming EU AI Act requires AI- systems to be trustworthy, entailing that they must be technically robust, and must comply with ethical criteria and with relevant laws and regulations (COM, 2021; European Commission, 2019).

1.1 The DECIDE-VerA project

AI in healthcare poses multiple challenges which are complex and interdependent. Often these challenges are addressed independently, failing to recognise the interactions between the ethical, legal and technical aspects. The DECIDE-VerA project applies an interdisciplinary approach to map the ethical and legal challenges of AI-CDSS and aims to formulate practical answers that can be adopted in the development and implementation of CDSS through design strategies.

For this aim we chose an AI-CDSS as a case study: the Clinical DECIsion support system for the carDiovascular risk managEment in primary health care (DECIDE). This CDSS is still in the early stages of development.[2] Specifically, an algorithm is being developed for predicting the cardiovascular risk of women and men under 50 years of age. Its intended foreseeable use is to screen populations in primary care to better and timely address cardiovascular risks.

To project uses a mixed-methods approach and consist of two phases. Phase I consists of a formative analysis of the ethical, legal and design aspects of the AI-CDSS DECIDE. The basis for the analyses is relevant (scientific) literature (for the formative ethical and legal analyses) and semi-structured interviews with key patient representatives and healthcare professionals (for the formative design analysis). The initial analyses will define the second phase (Phase II), where more in-depth activities will take place i.e., group sessions with patients, potential patients (healthy participants), and healthcare professionals. Group sessions include guidance ethics workshops. Based on the results of Phase I and II, we will provide recommendations for improving the design and implementation plan of the AI-CDSS DECIDE. It is expected that at the end of the project we will be able to provide an interdisciplinary and practical approach to explore and address AI systems from ethical, legal and design perspectives.

This report presents the results of the formative ethical analysis related to the AI-CDSS DECIDE.

[2] DECIDE-VerA follows up the research project DECIDE: Clinical DECIsion support system for the carDiovascular risk in primary health care. The AI-CDSS DECIDE is being developed in the DECIDE project, which is an independent initiative.

1.2 (Expanded) scope of the formative ethical analysis

The intended formative ethical analysis proposed to focus on analysing the AI-CDSS DECIDE from the perspective of the so-called "capability sensitive design" framework (see Chapter 2) Specifically we aimed at identifying which "capabilities", from the list of 10 central human capabilities by Martha Nussbaum (2000), seem particularly relevant for the use case of AI-CDSS DECIDE. AI-CDSS DECIDE is developed in the context of management of cardiovascular disease in primary care. A key characteristic of the capability approach of Nussbaum is that it aims at improving people's real freedom to achieve well-being.

During Phase I, the extent of the interdisciplinary collaboration has become more clear and we considered necessary to widen the formative ethical analysis to address the following:

- How to combine the normative theoretical analysis based on Capability Sensitive Design (Phase I), with the *empirical* ethical activities during Phase I (interviews) and Phase II (guidance ethics workshops); and
- How the results of the formative ethical, design and legal analyses (Phase I) can be integrated during Phase II.

The expansions of the formative ethical analysis aim to contribute to a central aim of the project, namely to develop a truly interdisciplinary approach by *integrating* the results of the analyses of the ethical, design and legal aspects of the AI-CDSS DECIDE.

1.3 Structure of this report

This report is structured as follows. **Chapter 2** explains the key characteristics of the ethical framework used in DECIDE-VerA, known as Capability Sensitive Design (CSD). CSD combines two well-known methodologies, namely Value Sensitive Design, and Martha Nussbaum's capability theory. The latter focuses on promoting people freedom to achieve wellbeing, and consists of a list of 10 central "human capabilities" deemed essential for human flourishing **Chapter 3**, in turn, identifies which of these capabilities seem particularly relevant for the development and use of an AI-assisted Clinical Decision Support System (AI-CDSS) in shared decision-making between doctors and patients to manage the risks of cardiovascular disease (DECIDE). **Chapter 4** explains how Capability Sensitive Design can be combined with a procedural approach for engaging stakeholders called the Guidance Ethics Approach. **Chapter 5**, finally, briefly outlines crucial next steps, specifically regarding how the results from the legal, ethical and design analyses can be integrated in Phase II of the project.

2. Capability Sensitive Design: key characteristics

Jacobs (2020) has developed a framework for Capability Sensitive Design (CSD), specifically for health and wellbeing technologies. The aim of CSD is to normatively assess technology design in general, and technology design for health and wellbeing in particular. The framework combines a well-known design methodology called Value Sensitive Design (VSD) with Martha Nussbaum's capability theory.

In this chapter, we will briefly present core ingredients of Martha Nussbaum's capability theory and *Value* Sensitive Design. The reason for discussing them in this order, is that some knowledge of Nussbaum's capability theory is required in order to understand how Jacobs (2020) uses elements of *Value* Sensitive Design in her theory of *Capability* Sensitive Design.

2.1 A capability theory focused on human capabilities central to human flourishing

Nussbaum's capability *theory* is one of many possible ways to specify what is more generally known as the capability *approach*. These theories share certain properties (making them part of the capability approach family), while also making choices that set them apart from other interpretations of that approach (turning them into specific capability theories). Hence, there is one capability *approach*, that can be specified in may different ways into specific capability *theories* (Robeyns & Byskov, 2023). After introducing the main concepts and commitments of the capability approach, we will briefly indicate the specific choices that are made in Nussbaum's capability theory.

At its core, the capability *approach* is based on two central normative claims. First, what is of central moral importance is the freedom of people to achieve wellbeing. Second, the wellbeing of people should be understood in terms of their "capabilities" and "functionings" (Robeyns & Byskov, 2023).

2.1.1 Capabilities

Capabilities are things that people can actually "do" or "be" if they so choose, for instance being able to be well-nourished, being able to get married, being able to be educated, being able to travel, et cetera.[3] In other words, capabilities refer to the "real freedoms" that people have to achieve wellbeing (Robeyns & Byskov, 2023).

In principle there are several possible versions of which items are included on the list of capabilities (see e.g., Byskov, 2020; Robeyns & Byskov, 2023). Like many other authors, Jacobs (2020) employs Martha Nussbaum's list of ten capabilities thought to be central for human flourishing (Nussbaum, 2000). Text box 1 provides an overview of that list.

[3] To make a clear distinction between explaining our theoretical and methodological approach, on the one hand, and 'applying' it to the case of AI-CDSS DECIDE for the purposes of the formative ethical analysis (see Chapter 3 ff), on the other hand, the examples given in Chapter 2 are deliberately unrelated to that case.

In order to be able to flourish as a human being, people should have the following capabilities, according to Nussbaum (2000):

1. Life – Able to live to the end of a normal length human life, and to not have one's life reduced to not worth living.

2. Bodily Health – Able to have a good life which includes (but is not limited to) reproductive health, nourishment and shelter.

3. Bodily Integrity – Able to change locations freely, in addition to, having sovereignty over one's body which includes being secure against assault (for example, sexual assault, child sexual abuse, domestic violence and the opportunity for sexual satisfaction).

4. Senses, Imagination and Thought – Able to use one's senses to imagine, think and reason in a 'truly human way'–informed by an adequate education. Furthermore, the ability to produce self-expressive works and engage in religious rituals without fear of political ramifications. The ability to have pleasurable experiences and avoid unnecessary pain. Finally, the ability to seek the meaning of life.

5. Emotions – Able to have attachments to things outside of ourselves; this includes being able to love others, grieve at the loss of loved ones and be angry when it is justified.

6. Practical Reason – Able to form a conception of the good and critically reflect on it.

7. Affiliation

A. Able to live with and show concern for others, empathize with (and show compassion for) others and the capability of justice and friendship. Institutions help develop and protect forms of affiliation.

B. Able to have self-respect and not be humiliated by others, that is, being treated with dignity and equal worth. This entails (at the very least) protections of being discriminated on the basis of race, sex, sexuality, religion, caste, ethnicity and nationality. In work, this means entering relationships of mutual recognition.

8. Other Species – Able to have concern for and live with other animals, plants and the environment at large.

9. Play – Able to laugh, play and enjoy recreational activities.

10. Control over One's Environment

A. *Political* – Able to effectively participate in the political life which includes having the right to free speech and association.

B. *Material* – Able to own property, not just formally, but materially (that is, as a real opportunity). Furthermore, having the ability to seek employment on an equal basis as others, and the freedom from unwarranted search and seizure.

Text box 1: Central human capabilities, according to Nussbaum (2000)

Nussbaum's account entails that this list should be regarded as a *threshold*. According to Nussbaum governments in all nations should guarantee these ten central capabilities (i.e. real freedoms) to their citizens, as a matter of social justice (Robeyns & Byskov, 2023, Section 4; Jacobs, 2020, p. 3368). Nussbaum (2000) justifies this list by appealing to the dignity and equal moral worth of every human being, arguing that each of the ten capabilities is needed to prevent a human life from becoming '…so impoverished that it is not worthy of the dignity of a human being' (p.72). If, for instance a technology design fails to bring a particular stakeholder group to the threshold level of one or more of these capabilities, then, according to Jacobs (2021) 'the technology design is not only inadequate but could

also be morally unjust. In other words: CSD is able to signal whether there is a (structural) injustice at play in a technology design when a particular stakeholder group for whom the technology is (partly) intended is not being brought up to the threshold level of one or more capabilities that have been identified to have moral value in the particular design context.' (p. 3372).

Note that the description of the capabilities list is still quite general and abstract. As indicated by (Robeyns & Byskov, 2023, Section 3.3) this leaves room for translating the central capabilities to implementation and policies at a local level, taking into account local differences, as advocated for by Nussbaum. This need for translation also means that further *specification* is required, in order to understand what Nussbaum's capability theory normatively requires in a specific case.

2.1.2 Functionings

While "capabilities" refer to things that people *can* do or be, if they so choose (people's real freedoms), "functionings" refer to capabilities that have been *achieved* (Robeyns & Byskov, 2023). For example, being able to vote is a capability (Political control over one's environment; item 10A on the list), actually voting is a functioning (Robeyns & Byskov, 2023). Likewise, the ability to travel is a capability (part of Bodily integrity; item 3 on the list), while actually traveling is a functioning (Jacobs, 2020).[4] Finally, the ability to have a good life including, e.g., nourishment is a capability (part of Bodily health; item 2 on the list), actually being nourished is a functioning, as is for instance avoiding escapable morbidity (Robeyns & Byskov, 2023).

2.1.3 Should we focus on opportunities or achievements?

The capability approach can be used to assess how people are doing in terms of wellbeing, by comparing them with others (interpersonal comparisons), or by looking at the wellbeing of an individual either at one point in time or over a period of time. Importantly, in each case the capability approach requires us to take a comprehensive or holistic approach, by looking at which sets or combinations of capabilities are open to a person. It might be the case that a person has two capabilities that are real freedoms when taken separately, but cannot be combined. For instance: '[C]an I simultaneously provide for my family and properly care for and supervise my children? Or am I rather forced to make some hard, perhaps even tragic choices between two functionings which both reflect basic needs and basic moral duties?' (Robeyns & Byskov, 2023, Section 2.6)

Should we focus on capabilities (i.e. real opportunities) or include functionings (i.e. actual achievements) as well when making interpersonal comparisons regarding wellbeing, or when looking at the wellbeing of an individual either at one point in time or over a period of time? Nussbaum (2000) chooses *capabilities* as the appropriate well-being metric, i.e. (real) opportunities, instead of achievements. This is based on liberal political persuasions: 'by focusing on capabilities rather than functionings, we do not privilege a particular account of good lives but instead aim at a range of possible ways of life from which each person can choose.' (Robeyns & Byskov, 2023, Section 3.2)[5]

Whichever choice we make in this regard, capabilities and functionings are not just seen as means to an end, but as ends in themselves when assessing for instance the effects of human behaviour and

[4] Capabilities should not be mistaken for "internal capabilities" of individuals. The difference is relevant because internal capabilities such as intellectual or emotional traits are not necessarily accompanied by the opportunity to exercise it, which is a condition of a capability under this approach.

[5] Robeyns & Byskov (2023) discuss a further range of considerations regarding whether the appropriate well-being metric should be capabilities (opportunities) or functionings (achievements). See Section 3.2 of their entry in the *Stanford Encyclopedia of Philosophy*. For the purposes of this formative ethical analysis, it suffices to understand Nussbaum's position on this issue.

policies: 'The capability approach evaluates policies and other changes according to their impact on people's capabilities as well as their actual functionings. It asks whether people are able to be healthy, and whether the means or resources necessary for this capability, such as clean water, adequate sanitation, access to doctors, protection from infections and diseases, and basic knowledge on health issues, are present.' (Robeyns & Byskov, 2023, Section 2.4).

If equality of capability is the focus, like it is in Nussbaum's account, then once the relevant capabilities (real freedoms) are in place, it could be argue that each individual should be held responsible for his or her own choices (Ibid., Section 3.2)

2.1.4 Real freedom requires adequate resources

The examples of *actually* voting, travelling and being nourished from section 2.1.2 make it clear that people need *resources* to actually realize the capabilities that they value. For instance, in the Netherlands voting at the very least requires a (red) pencil, a ballot, a ballot-box, and a valid ID. Typically, people need to travel to a specific location and a designated building in order to vote. Travelling, either for voting or for other reasons, requires resources as well. For instance, we need nourishment to provide us with the energy to be able to travel. Moreover, travelling typically involves wearing at least some clothes, including footwear. Depending on, for instance the travelling distance, our fitness and the time we have to get from A to B, we may also require a means of transportation such as a bike, a bus or a car. Our means of transportation, in turn, may require resources as well to operate, i.e. a type of fuel such as gasoline or electricity. Et cetera.

Taking fuel as a metaphor now, we could say that resources are what fuel functionings (realized capabilities). It is resources, commodities et cetera that are converted into achievements like *actually* voting, travelling, being nourished, et cetera.[6]

2.1.5 Conversion factors & human diversity

Essential as they are, though, resources are not enough. They are necessary, but not sufficient for people to have the real freedom to achieve wellbeing. This has to do with human diversity.

An important way in which the capability approach acknowledges human diversity is that it explicitly recognizes that there are multiple factors that can either hinder or promote the extent to which people are able to actually "convert" specific resources (e.g., money, food, materials, technologies, et cetera) into functionings. Factors that hinder our freedom to achieve wellbeing are called *negative* conversion factors, factors that promote our freedom to achieve wellbeing are called *positive* conversion factors (Robeyns & Byskov, 2023).

Conversion factors are commonly grouped in three categories: personal, social and environmental conversion factors. In the words of Robeyns and Byskov (2023, Section 2.3):

- '*Personal* conversion factors are internal to the person, such as metabolism, physical condition, sex, reading skills, or intelligence. If a person is disabled, is in bad physical condition, or has never learned to cycle, then [a] bike will be of limited help in enabling the functioning of mobility.
- *Social* conversion factors are factors from the society in which one lives, such as public policies, social norms, practices that unfairly discriminate, societal hierarchies, or power relations related to, for example, class, gender, race, or caste.

[6] This means that the freedom to achieve wellbeing, which is central to the capability approach, is at least in part a matter of distributive justice (Robeyns & Byskov, 2023, Par. 4): does everyone have the required means to be converted into functionings that are constitutive of human flourishment, that make up a good life?

- *Environmental* conversion factors emerge from the physical or built environment in which a person lives. Among aspects of one's geographical location are climate, pollution, the proneness to earthquakes, and the presence or absence of seas and oceans. Among aspects of the built environment are the stability of buildings, roads, and bridges, and the means of transportation and communication.'

What this overview shows is that multiple conversion factors can affect our freedom to achieve wellbeing. As a result, '[e]ach individual has a unique profile of conversion factors.' (Robeyns & Byskov, 2023: Section 2.5). To see how this works, take an example from Jacobs (2020) of a woman who buys a wearable fitnesstracker to help increase her capability of bodily health*:* 'This woman could have the personal conversion factor of having a sufficient physical condition to be able to walk and run and in that way use the fitnesstracker. She might also have the right environmental factors needed, such as having broad sidewalks and a park nearby to exercise in. But she might lack the social conversion factor needed if she lives in a neighborhood where it is unsafe for women to go out on their own.' (p. 3337)[7]

The capability approach therefore encourages us to always check whether, in a specific case, there are personal, social and/or environmental conversion factors that hinder or promote people's real freedom to achieve wellbeing.

Focusing on using a capability approach for designing technologies, e.g., Jacobs (2020) concludes that unique features of this framework are that it can account for human diversity (e.g., differences in terms of which capabilities people deem important and how important they are to them given their idea of a good life, but also in terms of e.g., personal, social and environmental conversion factors), and that it can potentially counteract structural injustices that may be (unintentionally) embedded in the process of technology design (an example of a negative socio-political conversion factor).

2.2 Value Sensitive Design

Value Sensitive Design (VSD) has a rich tradition of approximately 30 years. It commonly uses an iterative methodology consisting of three parts that mutually inform and are being informed by the other investigations: a conceptual, an empirical and a technical investigation. In a recent systematic review of almost three decades of VSD, these investigations are described as follows:

- 'The *conceptual* phase is often the preferred starting point for initial value elicitation through philosophical and conceptual investigations and clarification of which direct and indirect stakeholders to involve.
- In the *empirical* investigations, the value perspectives of both direct and indirect stakeholders are included by applying design methods from social science and design studies …, including the use of creative design tools and methods, such as, e.g., the renowned VSD envisioning cards …, and the Value Dams and Flows method to address value tensions in design …
- Informed by insights from the conceptual and empirical investigations, *technical* investigations focus on the technology itself and proactively design for values or reactively redesign existing technologies.' (Gerdes & Frandsen, 2023).[8]

[7] We welcome attempts to make our language more inclusive, for instance with regard to the broadest range of possible gender identifications. We are currently not aware of formulations that would be able to adequately do so, other than by consistently replacing the unnecessarily dichotomous "s/he" with for instance, 'human', 'individual', or 'patient', which would not improve readability. Absent a good alternative, we hope that the way in which we discuss the topic is open enough for everyone to be able read the gender identification(s) thought to be most appropriate for their situation whenever we write "s/he", "him/her" .

[8] Italics added.

Jacobs (2020) proposes that *Capability* Sensitive Design follows the same approach.

2.2.1 *Meeting challenges facing procedural approaches such as Virtue Sensitive Design*

Following others, Jacobs (2020) argues that VSD should be combined with a normative ethical theory. This has to do with several challenges facing procedural approaches that do not make any substantive ethical commitments. Jacobs mentions the following challenges in particular: '(1) obscuring the voice of its practitioners and thereby claiming unfounded moral authority; (2) taking stakeholder values as leading values in the design process without questioning whether what is valued by stakeholders also ought to be valued; and (3) not being able to provide normative justification for making value trade-offs in the design process.' (p. 3368).

In principle Value Sensitive Design can be combined with different normative ethical theories. Part of the rationale for combining VSD with Nussbaum's capability approach is that: '[It] provides such needed substantive normative foundation by defending that all people are morally equal and deserve a life worth living, which entails that every human being should have access to ten central capabilities. By explicitly complementing VSD with this substantive normative foundation of Nussbaum's capability theory, the CSD framework is able to provide sources of justification and argumentation for moral claims and considerations, which are needed to make principled judgments, to attend to a set of bounded and principled values, and to avoid conflating facts with values… Nussbaum's capability theory helps VSD to overcome the challenge of obscuring the voice of its practitioners as well as the naturalistic fallacy, …' (p. 3368)[9]

Importantly, Jacobs (2020) does not claim to provide *overriding* reasons for choosing to combine VSD with Nussbaum's capability approach, instead of with any other normative ethical theory.[10] She does provide reasons to think why CSD – which includes Nussbaum's capability theory – is particularly *well suited* to ethically evaluate technology design for health and wellbeing.

2.2.2 Well-suited to evaluate technology design for health and wellbeing

Jacobs (2020) provides a number of reasons why CSD is particularly well suited to ethically evaluate technology design for health and wellbeing. Firstly, CSD focuses primarily on people's capabilities and this fits quite well with the common aim of technology design to enhance and expand what people are able to be and do.[11] Secondly, CSD pays explicit attention to conversion factors, i.e. people's abilities to *converse* resources into capabilities, and by doing so CSD is able to account for human diversity.[12] Thirdly, and lastly, CSD is particularly well suited to the domain of technology design for *health and*

[9] Committing a "naturalistic fallacy" entails deriving what we should do (an 'ought') from what is empirically the case (an 'is').

[10] Faced with the challenges as described in the main text – which we acknowledge – we do not think that there is anything specific to Nussbaum's capability theory on the basis of which we could make a decisive choice between combining VSD either with Nussbaum's capability theory or with any other normative ethical theory, as far as meeting the challenges facing procedural approaches such as VSD goes. This has to do with the *generality* of both the challenges and the rationale given for why Nussbaum's capability approach can help meet these challenges. Concerning the latter: being 'able to provide sources of justification and argumentation for moral claims and considerations, which are needed to make principled judgments, to attend to a set of bounded and principled values, and to avoid conflating facts with values…' et cetera are general requirements of *any* normative ethical theory.

[11] See sections 2.1.1 and 2.1.2 of this report on capabilities and functionings, respectively.

[12] See section 2.1.5 of this report on conversion factors and human diversity.

wellbeing, according to Jacobs, because it aims to normatively assess technology design based on whether the design expands human capabilities that are identified as valuable (p.3365, 3373).

Jacobs goes on to discuss adhering to a specific concept of health, namely health as a person's ability to realize one's vital goals and to achieve or exercise a cluster of basic human activities (a definition previously proposed by other authors). Subsequently, she argues that: 'Given that we adhere to this conception of health, then CSD seems to be particularly suited to normatively assess technology designs for health and wellbeing.' (Ibid., p. 3365)

We agree with Jacobs that CSD is well suited to the domain of technology design for *health and wellbeing*. We also agree that it can be of *practical* use if there is a match between the concept of health and wellbeing that is used, on the one hand, and the normative commitments inherent in the method that is used to normatively assess e.g., technologies designed to promote health and wellbeing (here: CSD), on the other hand. That said, we see no reason to think that CSD would be less well suited to assess technologies that are designed with a different conception of health and wellbeing in mind. Nor to think that CSD would be less useful in situations where people have different conceptions of health or wellbeing. If we accept *value pluralism*, as Jacobs (2020) does (p. 3367), there seems to be room for, amongst other things, different interpretations of what we value, and for different interpretations of concepts such as health and wellbeing. Put positively, we think that, at least in theory, CSD could be well suited to assess even a broader range of cases than argued for by Jacobs (2020). Specifically, also to assess technology design in cases where the focus is on health and wellbeing technologies but where stakeholders have different notions of health and wellbeing in mind.

2.2.3 Additional challenges relevant for VSD

A recent systematic review of almost three decades of Virtue Sensitive Design provides useful information about additional challenges for VSD as a procedural approach involving stakeholders, that have not been covered explicitly so far. In their review Gerdes & Frandsen (2023) reiterate a number of challenges for VSD as formulated by Friedman *et al.* (2021). The challenges concern how to (1) account for power, (2) evaluate VSD, (3) frame and prioritize values, (4) enhance professional and industry appropriation, (5) influence and inform tech policy, (6) account for values and human emotions, (7) account for challenges related to AI, and finally, (8) how to settle value tensions. We will now highlight some of the challenges we think are particularly relevant for the purposes of the formative ethical analysis.[13]

Regarding the challenge how to account for power (challenge 1) Friedman *et al.* (2021) report that '[v]alue sensitive design theory and projects to date have not directly addressed the issue of power.' Fundamentally, the concern is that 'failing to address power relationships during the design research process risks limiting the ability of people, in all of their diversity, to thrive.' (p. 7). Key questions raised in relation to this challenge are:

- Which theories of power are well aligned for adoption or adaptation by value sensitive design?
- What are appropriate methods (either existing, or to be developed) for identifying and potentially addressing power relationships influenced by the outcomes of design?
- What are appropriate methods for identifying and potentially addressing power relationships throughout the process of design and design research?

[13] It is outside the scope of this formative ethical analysis to provide an exhaustive overview and an in-depth discussion of all challenges mentioned here. An important next step for Phase II of DECIDE-VerA is to further explore whether the way in which we proceed accounts as well as possible within the limits of the project to account for relevant methodological and practical challenges.

- What types of training and support will better position design researchers to identify, articulate and take responsibility for power dynamics in their work and the outcomes of their work?
- How can value sensitive design be applied as a method for evaluation, review and exposure of power and technological relationships? (Ibid., p. 8)

Challenges (3), (6), and (8) have already been touched upon briefly, at least in part, in the previous sections. Regarding how to frame and prioritize values (challenge 3) it was indicated, for instance:

- That the capability approach is based on two central normative claims. First, what is of central moral importance is the freedom of people to achieve wellbeing. Second, the wellbeing of people should be understood in terms of their "capabilities" and "functionings" (section 2.1);
- That within the capability approach capabilities and functionings are considered to be ends in themselves, not just means to some other end (section 2.1.3); and
- That the capability approach leaves room for value pluralism, including room for valuing things other than capabilities and functionings (section 2.2.2).

The combination of these points gives some guidance as to prioritizing values, in that Nussbaum's capability theory at least seems to provide normative reasons for prioritizing capabilities over values other than capabilities and functionings. Recall, that the theory takes *capabilities* (real freedoms) to be the adequate wellbeing metric, not functionings (realized capabilities) when making interpersonal comparisons or when assessing how an individual is doing in terms of wellbeing at a certain point in time or over a certain period of time (section 2.1.3). This, in turn, seems to leave room for individuals themselves to prioritize certain functionings (realized capabilities) over others, and to prioritize other values over certain functionings. From a liberal point of view, it is up to individuals how to decide whether or not they would like to realize their real freedoms.

Regarding how to account for values and human emotions (challenge 6) it was indicated, for instance:

- That "Emotions – being able to have attachments to things outside of ourselves" is a relevant capability (it is item 5 on Nussbaum's list of central human capabilities). This includes being able to love others, grieve at the loss of loved ones and be angry when it is justified (see Text box 1 in Section 2.1.1).[14]

While this description does seems to place normative boundaries on at least *some* emotions ('be angry when it is *justified*'), no specific emotions seem to be excluded off-hand. Moreover, many other central human capabilities seem inherently tied to emotions, such as sexual satisfaction (part of Bodily Integrity, item 3 on Nussbaum's list); fear, pleasure and pain (part of Senses, Imagination and Thought, item 4 on Nussbaum's list), empathy, compassion, humiliation and feeling recognized (part of Affiliation, item 7 on Nussbaum's list), and joy (part of Play, item 9 on Nussbaum's list).

Regarding how to settle value tensions (challenge 8), it was indicated, for instance:

- That '[b]y explicitly complementing [Value Sensitive Design] with… Nussbaum's capability theory, the [Capability Sensitive Design] framework is able to provide sources of justification

[14] The points relating to values that were put forward in response to challenge (3), are also relevant in response to challenge (6).

and argumentation for moral claims and considerations, which are needed to make principled judgments, to attend to a set of bounded and principled values…' (section 2.2.1)

Note that while challenges (3) and (8) partly overlap with the challenge mentioned by Jacobs of VSD not being able to provide normative justification for making value trade-offs in the design process, part of challenge (6), specifically how to account for human emotions, is new. The same goes for the other challenges identified by Friedman *et al.*, namely: how to (1) account for power, (2) evaluate VSD, (4) enhance professional and industry appropriation, (5) influence and inform tech policy, and (7) account for challenges related to AI.

Note too that all examples of ways to counter challenges (3), (6), and (8) for VSD as mentioned by Friedman *et al.* (2021) refer to sources *external* to Virtue Sensitive Design, namely to the capability approach or Nussbaum's capability theory.[15]

3. Capabilities that seem particularly relevant in case of the AI-CDSS DECIDE

Previous sections focused on some of the normative commitments and theoretical and methodological issues involved in combining Virtue Sensitive Design (VSD) and Nussbaum's capability theory into a framework of Capability Sensitive Design (CSD). This chapter presents the results of another key ingredient of the formative ethical analysis, as it was originally planned, namely which of the central human capabilities from Nussbaum's capability seem particularly relevant for assessing and designing the AI-mediated Clinical Decision Support System (AI-CDSS) for managing the risks of cardiovascular disease (DECIDE). More specifically, which capabilities seem particularly relevant for supporting shared decision-making between doctors and patients that is mediated by AI-CDSS DECIDE.

Methodologically and practically, this amounts to using Nussbaum's capability theory to perform parts of the first phase of Value Sensitive Design: conceptual investigation. Jacobs (2020) explains that the aim of the conceptual analysis is to: (1) select the capabilities and corresponding functionings that are relevant in the particular design context, (2) get clear who the stakeholders are that are affected by the technology design, and subsequently (3) identify what the relevant conversion factors at play are for these stakeholders (p. 3375). This chapter will cover parts of aims (1) and (3). Aim (2) is part of phase II of the DECIDE-VerA project, in which multiple stakeholder processes will be organized.

3.1 Selecting relevant capabilities and functionings

Recall that – as part of a separate initiative (the DECIDE-project) – an algorithm is being developed for predicting the cardiovascular risk of women and men under 50 years of age. Its intended foreseeable use is to screen populations in primary care to better and timely address cardiovascular risks.

In determining which capabilities and corresponding functionings seem particularly relevant for supporting shared decision-making between doctors and patients that is mediated by AI-CDSS DECIDE, 'relevant' is interpreted as covering two possible situations. 'Relevant' can mean that specific capabilities and/or functionings are *needed* in the case at hand, i.e. for shared decision-making on managing risks of cardiovascular disease mediated by AI-CDSS. It can also mean that specific capabilities and/or functionings can be *affected* using the AI-CDSS to support shared-decision-making

15

on managing the risks of cardiovascular disease. We will now briefly discuss each of the central human capabilities in turn, and indicate whether we think that they are relevant in these senses.

3.1.1 Life

Having the capability Life entails being able to live to the end of a normal length human life, and not to have one's life reduced to not worth living.

It is self-evident that one needs to be alive in order to participate in shared-decision-making on any given topic. In this sense both the capability Life and the functioning of actually being alive are necessary conditions for AI-mediated shared decision-making on how to manage the risks of cardiovascular disease.

Assuming that the aim of the AI-CDSS DECIDE is being able to *better* manage the risks of cardiovascular disease, we take it that the capability Life is also relevant in the sense of the capability potentially being (positively) affected by the AI-CDSS. Whether it actually does positively affect the capability Life, is an empirical matter, depending on for instance the specific workings of the AI-CDSS, including its efficacy.

3.1.2 Bodily Health

Having the capability Bodily Health entails being able to have a good life which includes (but is not limited to) reproductive health, nourishment and shelter.

We can imagine that a certain threshold needs to met in terms both of the capability of Bodily Health and in terms of actually having a good (enough) life including having good (enough) health, both physically and mentally,[16] for someone to be able to participate in AI-mediated shared decision-making on how to manage risks of cardiovascular disease.

Whether the AI-CDSS actually does positively affect the capability Bodily Health, again is an empirical matter, depending on for instance the specific workings of the AI-CDSS, including its efficacy. Consequently, so is the question whether the AI-CDSS positively affects actually having good health, both physically and mentally, if a person agrees with the technology being used. This will depend on the balance between potential positive effects of using the technology and any risks and burdens that may be involved, related to for instance the risk of false positive and false negative results flowing from the algorithm, the specific psychological and emotional response to the results flowing from the algorithm, et cetera.

3.1.3 Bodily Integrity

Having the capability Bodily Integrity entails being able to change locations freely, in addition to, having sovereignty over one's body which includes being secure against assault (for example, sexual assault, child sexual abuse, domestic violence and the opportunity for sexual satisfaction).

As with Bodily Health, we can imagine that a certain threshold level needs to be met both in terms of the capability Bodily Integrity and in terms of actually having sovereignty over one's body including actually being secure against assault, for someone to be able to participate in AI-mediated shared decision-making on how to manage risks of cardiovascular disease. Whether or not the functioning of actually traveling is relevant in this sense, depends on further empirical facts concerning the technology and the way that it can and will be used, including whether or not it is possible to use the technology remotely.

[16] On the notion of thresholds, see section 2.1.1 of this report.

At least at first sight, we see limited ways in which AI-mediated shared-decision-making in this context could affect the capability of Bodily Integrity. This seems to be limited to the possibility of it promoting our real freedom to have sovereignty over our body, for instance by increasing the opportunity for sexual satisfaction (related to our risk of cardiovascular disease, and how we weigh those risks). It seems unlikely, again at least at first sight, that and how the technology could affect our real freedom to change locations freely, or that it could increase our security against assault.

3.1.4 Senses, Imagination and Thought

Having the capability Senses, Imagination and Thought entails being able to use one's senses to imagine, think and reason in a 'truly human way' – informed by an adequate education. It also includes the ability to produce self-expressive works and engage in religious rituals without fear of political ramifications, the ability to have pleasurable experiences and avoid unnecessary pain, and the ability to seek the meaning of life.

Focusing on how shared decision-making using the AI-CDSS DECIDE might affect this capability or related functionings first, we see no possibilities. While we can imagine that using this technology in this context has the potential to (positively) influence whether or not we actually experience pleasurable experiences and not experience unnecessary pain, this pertains to functionings perhaps more akin to the capability Bodily Health, than to Senses, Imagination and Thought.

In another respect, though, this capability seems highly relevant, especially in the context of early engagement with stakeholders in developing technologies. At least when our criteria for highlighting a capability as 'relevant' cover both the potential effects that a technology may have on our capabilities (real freedoms), and specific capabilities themselves being necessary in the context of developing and using that technology.[17] On the latter interpretation: thinking of constructive ways to include values in the design of a technology, surely requires Senses, Imagination and Thought.

We submit that this interpretation of relevance can also be important to consider from the perspective of attempts to promote people's real freedom to participate in design processes of technologies aimed at promoting their real freedom to achieve wellbeing.

3.1.5 Emotions

Having the capability Emotions entails the ability to have attachments to things outside of ourselves. This includes being able to love others, grieve at the loss of loved ones and to be angry when it is justified. This capability seems relevant in the context of the design of the AI-CDSS DECIDE quite straightforwardly, and perhaps in the context of the design of health and wellbeing technologies in general. To start with, assuming that emotions have a cognitive dimension (e.g., Roeser 2006), the ability to have attachments to things outside of ourselves et cetera is an important ability for judging the moral acceptability of technological risks. An example of a relevant functioning would be actually being angry when this is justified, for instance when the AI-CDSS would be clearly used in a way that is contrary to one's needs.

[17] Notably, Jacobs (2020) does not highlight this capability as relevant in the context of the design of an AI-based therapy chatbot to help improve people's mental health. The reason for not including the capability Senses, Imagination and Thought as relevant in this case, might be that Jacobs primarily focuses on the possible effects that a technology may have on a capability in her interpretation of what constitutes a capability being relevant in the context of the design of that technology.

3.1.6 Practical Reason

Having the capability Practical reason entails the ability to form a conception of the good and critically reflect on it. Here too, the capability seems relevant in the context of the design of the AI-CDSS DECIDE quite straightforwardly, and perhaps in the context of the design of health and wellbeing technologies in general. For instance, it is an important precondition for reflecting on whether and to what extent the way in which the AI-CDSS (or any other health and wellbeing technology) is being developed and used aligns with our conception of the good. Results from the algorithm, and the way that the CDSS is used as part of shared decision-making in the context of managing cardiovascular risks might also provide relevant input for our real freedom to form a conception of the good and to critically reflect on it. A relevant functioning would be actually reflecting critically the place of a specific cardiovascular risk in our conception of the good.

3.1.7 Affiliation

Having the capability Affiliation entails being able to live with and show concern for others, empathize with (and show compassion for) others and the capability of justice and friendship. Institutions help develop and protect forms of affiliation. The capability also includes being able to have self-respect and not be humiliated by others, that is, being treated with dignity and equal worth. This entails (at the very least) protections of being discriminated on the basis of race, sex, sexuality, religion, caste, ethnicity and nationality. In work, this means entering relationships of mutual recognition.

Given the specific aim of the AI-CDSS DECIDE to support shared decision-making between patients and doctors regarding how to manage the risk of cardiovascular disease, this capability is highly relevant in the context of designing the CDSS. How the technology will be designed and will be used in the context of shared decision-making can have an important impact on many if not all of the components of our real freedom of affiliation. For instance, whether we are treated with dignity and equal worth, whether we are protected against discrimination (e.g., in relation to potential bias in the data used for the algorithm, or in the workings of the algorithm itself; see e.g., European Union Agency for Fundamental Rights, 2022), and is there mutual recognition?

Trust and trustworthiness in the doctor-patient relationship

This allows us to put the general emphasis on trustworthy AI into perspective. We submit that while it is crucial that the AI-CDSS will be worthy of trust, its importance is strongly connected to the importance of trust and trustworthiness in the doctor-patient relationship more generally.

The upcoming AI Act stresses the importance of AI-systems being trustworthy (COM, 2021; European Commission, 2019), i.e. worthy of our trust. One of the reasons why it is important that AI-systems are worthy of our trust can be linked to the importance of trust and trustworthiness more generally in contexts in which the use of a specific AI-system is being considered. For instance, for a doctor-patient relationship to be beneficial for patients maintaining trust and being trustworthy are key, irrespective of the technologies that are used or considered as a result of their shared decision-making. Their importance can be easily understood from the perspective of four generally recognized medical-ethical principles: respect for autonomy, beneficence, non-maleficence, and justice (Beauchamp & Childress, 2019).[18] First, maintaining trust and being trustworthy are crucial in health care because they are necessary conditions for equal access to health care. If a patient does not trust a doctor, s/he might not be willing to visit a doctor, even if s/he might urgently need medical assistance. That would be

[18] We assume that the importance of maintaining trust and trustworthiness holds both for doctors and patients a-like. For the purposes of this formative analysis the examples given are not meant to provide an exhaustive overview of the roles to be played by doctor and patient regarding trust and trustworthiness.

problematic, e.g., on grounds of *justice*. Second, trust and trustworthiness are also necessary conditions for being able to provide good care. If a patient does visit a doctor but does not trust him/her, the patient might not be willing to share the information that the doctor needs in order to be able to provide adequate care (i.e., care that would do more good than harm to the patient). That would be problematic from the perspective of the principles of *beneficence* and *non-maleficence*. Third, this means that doctors maintaining trust and being trustworthy, are important conditions for patients to exercise their autonomy (and for the doctor to be able to show *respect for* the *autonomy* of the patient). Trust and trustworthiness are thus inherently tied to the ability of doctors to fulfil their duty of care. The worst case scenario is that a patient who does not trust a doctor, will not seek medical assistance. It is not at all self-evident that that would qualify as a prime example of an autonomous choice, as a patient exercising his/her self-determination. It also makes it difficult (if not impossible) for the doctor to respect the autonomy of the patient. Not because there is a deliberate lack of respect on the part of the doctor, but because it is unclear whether there is a truly autonomous choice of the patient or what exercising his/her right to self-determination would amount to.

Maintaining trust and trustworthiness in AI-mediated shared decision-making

This brings us straight to the heart of the DECIDE-VerA project, given that it focuses on the ethical, legal and design aspects of the (potential) introduction of a specific technology – AI-CDSS – for the management of cardiovascular disease risk. Given what was said earlier, an important moral (and legal) requirement, then, is that the AI-CDSS should support *shared* decision-making between individuals and doctors concerning how to manage patients' cardiovascular disease risk, while maintaining trust and trustworthiness more generally.

Implications for the process of obtaining informed consent

The capability approach generally provides a rich conceptual and normative framework for assessing, for instance the design and potential use of a health and wellbeing technology such as the AI-CDSS DECIDE. For instance, it allows us to reflect on potential implications for how to put specific and generally recognized norms into practice. What has been said so far, indicates that employing a Capability Sensitive Design framework for assessing health and wellbeing technologies can have implications for, e.g., the process of obtaining informed consent, if we let the distinction between real freedoms (capabilities) and actual functioning (realized capabilities),[19] the importance of having adequate resources,[20] and the potential influence positive or negative conversions factors[21] sink in.

These considerations indicate that if the aim is to promote people's (real) freedom to achieve wellbeing, it may not be enough to provide patients with a technology (the AI-CDSS), with information about how it works, the risks, burdens and benefits involved, and have them choose. This *might* be necessary and sufficient for obtaining informed consent and to show respect for the autonomy of patients. It is an open empirical question, however, whether it is sufficient and if so under what conditions to promote the real freedom of individuals to achieve well-being, even if they voluntarily consent to the AI-CDSS being used in the process of shared decision-making with a doctor. That would require paying attention, for instance to potential personal, social and environmental conversion factors that could promote or hinder the ability of a specific individual, in his/her specific circumstances to realize capabilities s/he considers to be crucial for having a good life. Indeed, it requires attention

[19] See sections 2.1.1 and 2.1.2, respectively.

[20] See section 2.1.4.

[21] See section 2.1.5.

to increase the chance of the AI-CDSS being a positive conversion factor, and to reduce the risk of it being a negative conversion factor.

Implications for what respecting autonomy requires?

A more fundamental question is whether employing a capability theory has implications for what we think is required for doctors to adequately respect the autonomy of individuals. On Nussbaum's account, which focuses on responsibilities of governments to make sure that people have at least *thresholds* of all ten capabilities (see section 2.1.1), the implications for what is required for doctors to respect the autonomy of patients, might be limited. This depends, though, on further discussion and analysis of who should do what, if anything, either when the required thresholds are met, or when they are not met (assuming that we know when this is the case).

3.1.8 Other Species

Having the capability Other Species entails the ability to have concern for and live with other animals, plants and the environment at large. At first sight, this capability, or specific functionings related to it, do not seem particularly relevant in the context of the design of the AI-CDSS DECIDE. There is perhaps at least one notable exception, having to do with the environmental footprint of AI-applications, which has come under scrutiny, both in general and regarding AI-applications in health care. Richie (2022), for instance warns against the potential negative environmental impact and also highlights the need and potential for developing environmentally sustainable AI in health care.

As we have seen in section 2.1.5 our environment is a conversion factor, which – depending on the circumstances – can have major implications for our ability to convert relevant resources into functionings that we deem important. In theory, it can affect all of our capabilities and functionings.

3.1.9 Play

Having the capability Play entails being able to laugh, play and enjoy recreational activities. Depending on the specifics of the case, for instance the severity of cardiovascular risks, and the implications of these risks for our daily lives, this capability could be identified as relevant for the design of AI-CDSS DECIDE.

3.1.10 Control over One's Environment

Having the capability Control over One's Environment, finally, has a political and a material component. The *political* component entails the ability to effectively participate in the political life which includes having the right to free speech and association. The *material* component entails the ability to own property, not just formally, but materially (that is, as real opportunity). Furthermore, having the ability to seek employment on an equal basis as others, and the freedom from unwarranted search and seizure.

In one sense, we estimate that the political component of the capability Control over One's Environment is not relevant for the design of the AI-CDSS DECIDE. We do not expect this specific technology to have an influence on our real freedom to effectively participate in the political life and our having the right to free speech and association.[22] On a less formal understanding of 'political', though, this capability seems highly relevant, especially in the context of early engagement with

[22] Regarding the right to association, also see our considerations in section 3.1.3 (Bodily Integrity).

stakeholders in developing technologies. Specifically, involving stakeholders in the process of designing the AI-CDSS DECIDE itself can support this capability.

Regarding the material component of the capability of Control over One's Environment, this could be relevant in the context of design of the AI-CDSS DECIDE in the sense that e.g., specific agreements about how information about risks of cardiovascular disease is used and disseminated, could potentially impact our ability to seek employment on an equal basis as others, if such risk information would be used in employment practices.

3.2 Tacking stock

This concludes our formative ethical analysis of which capabilities from Nussbaum's capability theory seem particularly relevant for the design of the AI-CDSS DECIDE. It amounts to performing parts of the first phase of Value Sensitive Design: conceptual investigation. Specifically, the results contribute to parts of the first and third aim of what are commonly regarded as the aims of the conceptual analysis, namely to: (1) select the capabilities and corresponding functionings that are relevant in the particular design context, (2) get clear who the stakeholders are that are affected by the technology design, and subsequently (3) identify what the relevant conversion factors at play are for these stakeholders. We have mostly identified relevant capabilities, while giving some examples of relevant corresponding functionings, and some examples of potential conversion factors.

Within this limited scope, the analysis yields several important and useful results. Specifically:

- A contribution to what is required for, for instance a capability to be 'relevant' in the context of designing a specific health and wellbeing technology. We have argued for a broad interpretation, covering both the question whether a capability/functioning is needed, and if so to what extent to be able to use the technology in a desired way, and the question whether capabilities/functionings can be affected by using that technology (section 3.1).

Taking this broad approach, we have also been able to:

- Provide relevant examples of the inherent link between capabilities and conversion factors, that help highlight some of the limitations of the common approach to respecting autonomy of patients, some of the potential implications for organizing processes of shared-decision-making, and that help put the focus on the importance of trustworthy AI in perspective. Specifically, the broader perspective of the importance of maintaining trust and trustworthiness for the doctor-patient relationship more generally (sections 3.1.2, 3.1.3 and 3.1.7); and
- Link two central human capabilities directly to reasoning about the importance of involving stakeholders in processes of designing health and wellbeing technologies (sections 3.1.4 and 3.1.10).

In line with the planning of DECIDE-VerA these and other results will inform the next part of the project, which includes, amongst other activities, in-depth and extensive stakeholder engagement, and the *integration* of the results of the ethical, legal and design analyses. This will no doubt further enrich the analysis of relevant capabilities, functionings and conversion factors.

In the remaining chapters, we will first briefly discuss how one specific way in which stakeholders will be involved in the next part of the project, through Guidance ethics workshops, relates to Capability Sensitive Design, specifically, how they can be combined (Chapter 4). Finally, we will indicate what we foresee as important next steps given the central aim of the project, namely to develop a truly interdisciplinary approach by *integrating* the results of the analyses of the ethical, design and legal aspects of the AI-CDSS DECIDE (Chapter 5).

4. Combining Capability Sensitive Design with Guidance Ethics

4.1 Introduction

DECIDE-VerA combines Capability Sensitive Design (CSD) with Guidance ethics in order to flesh out relevant ethical aspects of the AI-CDSS DECIDE, and to map – in close interaction with stakeholders – ways to responsibly further develop and (potentially) implement the AI-CDSS in clinical care regarding the management of cardiovascular risk. Key characteristics of Capability Sensitive Design were discussed in Chapter 2. CSD, it was explained, combines Virtue Sensitive Design with Nussbaum's capability theory. VSD, in turn, typically consists of three strands of investigation: conceptual investigation, empirical investigation, and technical investigation. As indicated Chapter 3 covers part of the main aims of the conceptual investigations. Technical investigations are reported on separately, in the formative *design* analyses. That leaves empirical investigation. This is what we will be using guidance ethics workshops for.

Outline

We will now first introduce the main components of guidance ethics workshops (section 4.2). Then, we indicate how we will use these workshops as part of our empirical investigation (section 4.3), and how the guidance ethics workshops will be combined with the other component of Capability Sensitive Design, namely Nussbaum's capability theory (section 4.4).

4.2 Guidance ethics workshops: main components

The Guidance ethics approach was developed by Daniël Tijink (ECP | Platform voor de informatiesamenleving) in collaboration with philosopher of technology Peter Paul Verbeek (Verbeek & Tijink, 2020). The approach usually takes the form of a structured dialogue with stakeholders to systematically discuss the possible use of a concrete technology (for instance the AI-CDSS) in a concrete context (for instance shared decision-making between doctors and patients about how to manage risks of cardiovascular disease). The aim is to represent the perspectives of all stakeholders who might be influenced or impacted by the development of a technology and/or its use. The duration of the workshop ranges from 3,5 – 4 hours.

A workshop typically consists of three stages:

1. Description of the technology in context
2. Participants identify: (a) (additional) actors who might have a stake in the development and/or use of the technology and who should ideally be invited to the conversation as well (at a later stage); (b) potential positive and negative effects of the use of the technology, and key values involved in those effects.
3. Participants discuss strategies ("options for action") to promote the positive effects and limit the negative effects of the technology. A distinction is made between three types of options for action , that can complement each other: a) changes in the technology (*ethics by design*); b) changes in the social, physical and institutional environment in which the technology is used (*ethics in context;* this includes protocols, policy, laws and regulations); and c) options for action to promote users' responsible use of the technology (*ethics by user*).

Guidance ethics workshops will be organized in Phase II of DECIDE-VerA. As is standard in using this approach, we will invite participants in the following categories: (medical) professionals; citizens/clients/patients; managers and policymakers, and technology developers.

4.3 Guidance ethics workshops as empirical investigation

Guidance ethics workshops will be used as empirical investigation that make up the second phase of Virtue Sensitive Design. Borrowing from Jacobs (2020, p. 381) the elements of the empirical investigation can be summarized as follows.[23]

Generally, the empirical investigation explores if the findings of the conceptual phase correspond with the experiences and values of the direct and indirect stakeholders.

With the help of various empirical methods such as interviews, surveys or focus groups VSD practitioners develop an understanding of how stakeholders are experiencing current provisions and services of the care in question, what stakeholders currently value and what they are missing, and what their initial impressions are of the technology in question.

Furthermore, in conversation with the stakeholders, VSD practitioners further specify the relevant ethical considerations, adding context-specific content to what may still be abstract considerations from the conceptual investigation.

Subsequently, based on these findings VSD practitioners can make prototypes of the envisioned technology that incorporates the results of the conceptual and empirical investigation. Then, after the making of the first prototype(s) there is, ideally, a second empirical investigation conducted in which the prototype is presented to the stakeholders and their assessment of the prototype is explored. Based on these findings, the prototype is adjusted up to the point that it finds its 'ideal' form, to the extent that is feasible.

4.4 Combining guidance ethics workshops with Nussbaum's capability theory

Consistent with the framework of Capability Sensitive Design, we will, in the next phase of DECIDE-VerA, combine guidance ethics workshops with Nussbaum's capability theory. The central question will be how, to what extent, and under what conditions use of the AI-CDSS in shared decision-making between doctors and patients on how to manage the risks of cardiovascular disease can promote people's freedom to achieve well-being. Following the regular procedure of guidance ethics workshops, we will identify:

- Possible positive and negative effects of the AI-CDSS on the capabilities of stakeholders that influence the doctor-patient relationship and the quality of shared decision-making when dealing with cardiovascular diseases risk management. When making an inventory of positive and negative effects, particular attention is paid to so-called "conversion factors" i.e., factors that could either prevent (*negative* conversion factors) or facilitate (*positive* conversion factors) people from achieving wellbeing, if they so choose.
- Which values are at play in those effects. Here, we will explicitly leave room for value pluralism, entailing both that people might value several other things besides capabilities and functionings, that they might have different interpretations of specific values, and that they might assign different importance to specific values, even if they interpret those values the same way as others do; and
- Potential options for action that might help to ensure that those involved can realise their capabilities, if they so choose. Explicit attention will be paid to possibilities for strengthening or inserting positive conversion factors, as well as for removing or limiting the effect of negative conversion factors. Specifically, we will look at possible options for action that can contribute to this, again distinguishing three types: a) changes in technology (*ethics by design*);

[23] For reasons of methodological clarity, we have omitted references to capabilities in the description of the phase of empirical investigation as part of *Value* Sensitive Design.

b) changes in the social, physical and institutional environment in which the technology is used (*ethics in context*); and c) user requirements for responsible use of the technology (*ethics by user*).

In line with the recommendations for the empirical investigation (section 4.3) several guidance ethics workshops will be organized in the next phase of DECIDE-VerA. Results from the formative analyses will provide input for the first guidance ethics workshop. In turn, results from the workshop will provide input for subsequent activities and publications.

5. Next steps

The main aim of DECIDE-VerA is to develop a truly interdisciplinary approach by integrating the results of the analyses of the ethical, design and legal aspects of the AI-CDSS DECIDE. In this final chapter, we will briefly and tentatively indicate what we foresee as important next steps to that aim.

5.1 Ethical analysis in Phase II

In addition to organizing multiple guidance ethics workshops in Phase II of the project (section 4.4), public reports will be made of each guidance ethics workshop, including an overview of the possible positive and negative effects of the AI-CDSS discussed, key values involved in those effects, and options for action. The reports will explicitly discuss the results from the perspective of Capability Sensitive Design. Also, a scientific ethical paper will be written integrating all relevant main results of the stakeholder processes (guidance ethics workshops, interviews[24]), the normative analyses (ethical and legal) and the design analyses contained in DECIDE-VerA.

5.2 Towards integrating the results of the ethical, legal and design analyses

Integrating the results of all project activities in a truly interdisciplinary approach at the very least requires developing specific ideas and approaches as to how the different activities relate to one another, the extent to which they can build on each other, and how.

Without striving to be exhaustive, we will now briefly reflect on how the different elements of the project relate to another, and some of the steps that could be taken to actually integrate their results.

5.2.1 Building on the formative analysis of the Assessment List for Trustworthy Artificial Intelligence (ALTAI)

One of the main outcomes of the analysis of ALTAI is that the applicability of the assessment list for technologies in early stages of development, such as the AI-CDSS, is limited. This is of particular interest for the following further activities: the scoping review of additional normative frameworks, the guidance ethics activities, the legal analysis, and further design activities.

A crucial next step is to monitor possibilities for accommodating some of the limitations of ALTAI, while still contributing to the AI-CDSS DECIDE being trustworthy, in the sense of relevant technical, legal, and ethical requirements. In general, results pertaining to laws, regulations and protocols, can be easily integrated in our guidance ethics activities, given that laws, regulations and protocols are examples of a specific type of options for action, namely *ethics in context* (section 4.4).

[24] Interviews have been conducted as part of a separate formative analysis in Phase I of DECIDE-VerA.

5.2.2 *Building on the (formative) legal analysis*

The formative legal analysis discusses regulations with regard to liability that are relevant for the AI-CDSS DECIDE, and discusses some of the legal implications for using the AI-CDSS in practice. This is of particular interest for the guidance ethics activities and further design activities.

Crucial next steps include connect the legal and the further ethical analysis. Part of that connection is to explicitly reflect on how liability issues can be incorporated in identifying or developing responsible options for action. Generally, considerations having to do with liability are the legal counterpart of moral responsibility. We will explicitly pay attention, for instance in the guidance ethics workshops to ways in which liability issues might impact the doctor-patient relationship in such a way that it may help promote or hinder people's freedom to achieve wellbeing (liability regulations as a social conversion factor).

Another crucial step includes connecting the legal analysis to the further design activities. Specifically, we will explore options to counter potential liability issues by way of (capability sensitive) design.

5.2.3 *Building on the design analysis*

The formative design analysis reflects on ways in which design can incorporate relevant values. This is of particular interest for the legal analysis and the further ethical analysis.

Crucial next steps include integrating the results from the design analysis with the results from the formative ethical analysis (which also contain considerations related to design). This should be done before and as part of preparing the first guidance ethics workshop (February 2024). It is also crucial to reflect on whether the formative design analysis can provide further insights into which potential liability issues having to do with using the AI-CDSS DECIDE are irreducible (i.e. cannot be avoided by means of smarter design). That, in turn, is highly relevant input for the type of ethical and policy decisions that will have to be made when considering whether or not, and if so how, to implement the AI-CDSS DECIDE, once it is ready.

5.2.4 *Building on the formative ethical analysis*

This report presents the results of the formative ethical analysis. The results are of particular interest for the further legal analyses and further design activities.

A first crucial next step coincides with one formulated in the previous section, namely to reflect on whether the formative design analysis can provide further insights into which potential liability issues having to do with using the AI-CDSS DECIDE are irreducible (i.e. cannot be avoided by means of smarter design). On this point, the design team and the ethical team can work together. In the same vein, these teams can compare the results of their analysis to see how their finding and reflections can contribute to a comprehensive view of and grip on relevant ways in which design can (or cannot) help promote people's freedom to achieve wellbeing in the context of using the AI-CDSS to support shared decision-making between doctors and patients to manage the risks of cardiovascular disease. Finally, in preparing the guidance ethics workshops it is important to reflect on ways in which we might be able to accommodate some of the challenges facing Capability Sensitive Design that were discussed in chapter 2, such as the challenge to accommodate the potential of power relations during to limit the ability of people, in all of their diversity, to thrive (which would amount to negative social conversion factor) (section 2.2.3).

Literature

Beauchamp, T.L., J.F. Childress (2019). Principles of Biomedical Ethics. Eight edition. New York: Oxford University Press.

Byskov, M.F. (2020). The Capability Approach in Practice. A New Ethics for Setting Development Agendas. London/New York: Routledge

COM (2021) 206 final 2021/0106 (COD) Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS, https://artificialintelligenceact.eu/the-act/

EGE (2018). Statement on artificial intelligence, robotics and 'autonomous' systems. Directorate-General for Research and Innovation, Brussels. https://op.europa.eu/en/publication-detail/-/publication/dfebe62e-4ce9-11e8-be1d-01aa75ed71a1/language-en/format-PDF/source-78120382

European Commission. (2019). Ethical guidelines for trustworthy AI. High Level Expert Group on Artificial Intelligence. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

European Union Agency for Fundamental Rights, *Bias in algorithms – Artificial intelligence and discrimination*, Publications Office of the European Union, 2022, https://data.europa.eu/doi/10.2811/25847

Friedman, B., Harbers, M., Hendry, D.G. *et al.* (2021). Eight grand challenges for value sensitive design from the 2016 Lorentz workshop. *Ethics Inf Technol* 23, 5–16. https://doi.org/10.1007/s10676-021-09586-y

Gerdes, A., Frandsen, T.F. (2023). A systematic review of almost three decades of value sensitive design (VSD): what happened to the technical investigations?. *Ethics Inf Technol* 25, 26. https://doi.org/10.1007/s10676-023-09700-2

Jacobs, N. (2020). Capability Sensitive Design for Health and Wellbeing Technologies. *Science and Engineering Ethics*, *26*(6), 3363–3391. https://doi.org/10.1007/s11948-020-00275-5

Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., & Wang, Y. (2017). Artificial intelligence in healthcare: Past, present and future. Stroke and Vascular Neurology, 2(4), 230–243. https://doi.org/10.1136/svn-2017-000101

Nussbaum, M. (2000) *Women and Human Development: The Capabilities Approach*, Cambridge University Press: Cambridge, MA.

Richie C. (2022). Environmentally sustainable development and use of artificial intelligence in health care. Bioethics. Jun;36(5):547-555. doi: 10.1111/bioe.13018. Epub 2022 Mar 15. PMID: 35290675; PMCID: PMC9311654.

Robeyns, I. and M.F. Byskov (2023), "The Capability Approach", *The Stanford Encyclopedia of Philosophy* (Summer Edition), Edward N. Zalta & Uri Nodelman (eds.), https://plato.stanford.edu/archives/sum2023/entries/capability-approach/

Roeser, S. (2006). The role of emotions in judging the moral acceptability of risks. *Safety Science 44(8): 689-700.* https://doi.org/10.1016/j.ssci.2006.02.001

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

UNESCO (2021). Recommendation on the Ethics of Artificial Intelligence. France.

Verbeek, P.-P. & D. Tijink (2020). *Guidance ethics approach – An ethical dialogue about technology with perspectives on actions*. The Hague: Platform voor de InformatieSamenleving (ECP), https://ecp.nl/publicatie/guidance-ethics-approach/

WHO (2021). Ethics and Governance of Artificial Intelligence for Health: WHO Guidance. Geneva. Licence: CC BY-NC-SA 3.0 IGO. Geneva.

Yu, K.-H., Beam, A. L., & Kohane, I. S. (2018). Artificial intelligence in healthcare. Nature Biomedical Engineering, 2(10), 719–731. https://doi.org/10.1038/s41551-018-0305-z