The Possibility of Normative Moral Decision Making Artificial Intelligence and its Effects on the Practical Nature of Morality.

A Postphenomenological Analysis

Emmeke Vierhout

University of Wales

Trinity Saint David

Distance Learning Program

Januari 2024

Emmeke Marleen Vierhout (10803420)

January 2024

This dissertation is submitted in partial fulfilment of the requirements for the degree of Masters of Arts in Philosophy at the University of Wales Trinity Saint David, Lampeter Campus

Supervisor: Tristan Nash

Number of words: 14966 (excluding title, current page, abstract, table of contents and bibliography)

Abstract:

Background: Researchers from various background are working on the development of Artificial Intelligence (AI) that is designed to support professionals in moral decision making. These types of AI are to predict our moral beliefs or judgments in specified choice situations. This paper explores the potential of such AI-technologies to be normative for its users and how that may affect the practice of moral decision making in specified choice situations.

Methods: An account of the practical nature of morality is given based on the ideas of Mary Midgley in her work on the nature of morality. Against this background the possibility Moral Decision Making AI (MDMAI) to be normative for its users is assessed for three different ways of designing such MDMAI. As an example of an MDMAI that has the possibility of being normative for its users, the technology of BAIT is used as a starting point in a postphenomenological analysis to assess how such normative MDMAI may transform the practical nature of moral decision making.

Conclusions: BAIT as a normative tool in specified moral choice situations, may transform the moral decision making into becoming conservative in nature. This effect may be enhanced by using BAIT as a tool for justifying our decisions. In relying on BAIT in our moral choices we leave out important though hard to identify motives in our deliberations. This may affect the social relation between the persons involved in the decision. Having more and more accurate technologies like BAIT available in the future in more specified situations of moral decision making, this may affect social relations on a larger scale and also the freedom of professionals in making moral decisions.

Table of Contents

Introduction

Throughout history technology has had an enormous impact on the lives and the development of human beings. We might even argue that we became human beings through technology. AI is just the latest in a history of high-impact technology such as the hand-ax, man-made fire, steam-engines and long-distance communication devices. The impact of technology is not merely practical, the deployment of technology can plausibly undermine or promote specific moral values; technology is value laden.

That people recognize AI as a non-neutral technology is reflected in the explosion of thinking we've seen in articles about what is called 'Ethical AI' or Machine Ethics. Most mentioned in this regard are values or principles like respect for autonomy, non-maleficence, fairness, transparency, explainability and accountability. Task forces and organizations like the EU High Level Expert Group on AI have come up with moral standards, moral frameworks, ethical principles and values along the lines of which AI should operate. This has resulted into the EU AI-Act that 'aims to ensure that fundamental rights, democracy, the rule of law and environmental sustainability are protected from high risk AI' (https://www.europarl.europa.eu/, 2023). These regulations concern mostly AI that have a high potential of being abused for undemocratic purposes, such as facial recognition and ethnic

profiling. However there is another way in which researchers are working on developing 'Ethical AI', which is AI that is not merely being developed regarding our moral values, but AI that actually gives us our moral values; AI that supports us in making moral decisions and of which the outcomes have the potential of being normative for its users. I will refer to such AI as Moral Decision Making AI (MDMAI).

In this dissertation I will investigate three different types of MDMAI and the possibilities of these different kinds to be normative for its users. Furthermore, I will investigate what the impact of such possible normative MDMAI may be on the way we practice morality. I want to investigate these impacts along the lines of the postphenomenological theory of technology as described by Don Ihde (1990, 1995) and further developed by Peter Paul Verbeek (2005, 2016). In chapter three I will briefly introduce the postphenomenological way of approaching technologies and its main concepts. In the last chapter I will provide my postphenomenological analysis of a possible normative MDMAI regarding our practical moral reality.

In order to be able to understand possible effects of any technology on moral reality we need a clear picture of what our practical moral reality entails. So I will begin this dissertation by setting out a theory of a practical nature of morality based on Mary Midgley's work on this

topic. This account will provide the basis upon which I will conduct the postphenomenological analysis of MDMAI in the final chapter.

Chapter 1

The nature of morality

There are two ways in which we can talk about the nature of morality:

1) The first way is with respect to the ontology of morality or what I would like to call the Nature of Morality. This concerns questions about whether moral claims can be truth apt and if they are, in what way can they be true or not? What is the Good and why should we do what is good? What *is* moral obligation?
2) The second way in which we can talk about the nature of morality is in a more worldly practical sense, without the capital N. When is a claim a moral claim, and what makes a claim a moral claim? What is the system behind morality, or better said the practical nature of morality. How does it work?

Although I think that views about the nature of morality have implications for possible answers to the questions concerning the Nature of Morality I don't want to make any claims about the latter. In this dissertation I am looking at the influence of MDMAI on the practical nature of morality, I am looking at the influence of MDMAI on how morality works for us human beings or how may MDMAI affect our moral decisions.

How does morality work?

No matter what your ontological stance is regarding the Nature of morality, even an extreme moral anti-realist can identify a (supposed) moral claim from other types of claims. If you look at the following claims, most human beings know which of the two is a moral one and which one isn't.

1) "Andy! Don't kick that boy, it is not allowed in judo!"
2) "Andy! Don't kick that boy, you are hurting him!"

Comparing both claims may give us an entry into examining the more practical nature of morality, how does a moral claim come about? In the first example, Andy probably just started practicing judo and when he kicks the other boy, he thought that was allowed in judo. The trainer explained to him that it wasn't. Kicking is allowed for example in karate or MMA. What about the second example? The answer to that question is much less simple. Why is hurting someone else an argument against kicking that person? Well the most likely answer someone would give is: "because it is wrong to hurt someone." And with this answer we come into the practical domain of morality, which is notoriously hard to define. But I would say in essence, the domain of morality concerns the questions that pop up when taking "the significant other" into account in our deliberations and acts. Where *significant* refers to the other not as a means but as an end in itself. A significant other could be another person or animal, parts of nature, or something more abstract as society or future generations.

How does such a moral claim about what the right or wrong actions are come about? How does the practical reality of making moral judgments work? To answer this question we might gain some insight from a conceptual model presented in fig 1. by Chorus (2015) which is an attempt to summarize, structure and mutually relate the myriad of findings emerging from a literature review of both empirical studies and studies from the social sciences into moral choice behavior. The model takes the perspective of the moral choice behavior of a considered individual when confronted with a (morally) loaded choice situation and with the expected and actual behaviors of relevant others (Chorus, 2015).

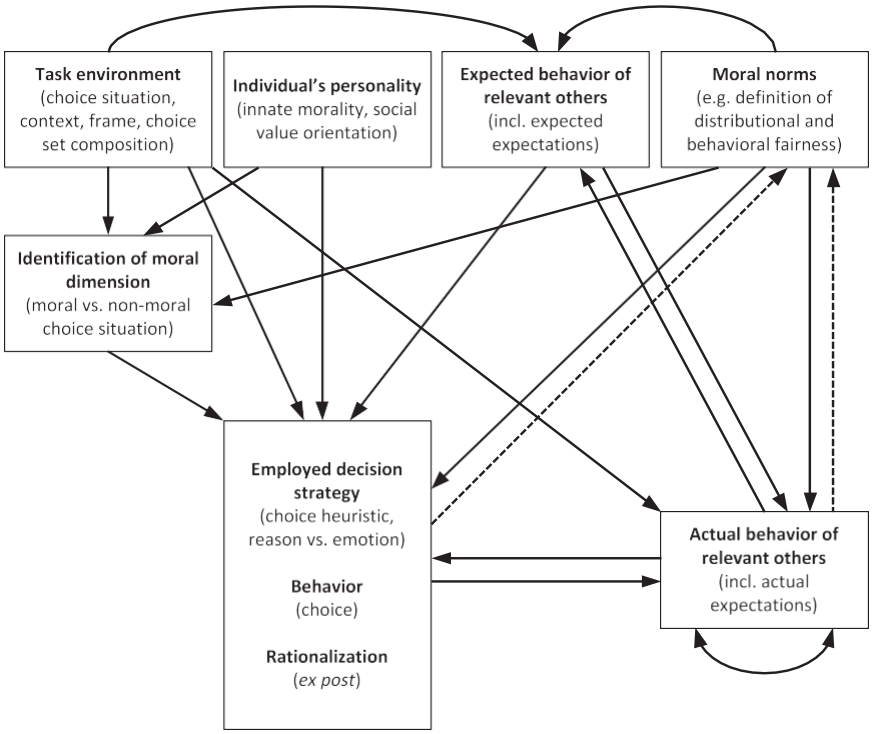

Fig. 1. Conceptual model of an individual's moral choice behaviour (Chorus, 2015)

If Chorus is right, our actual moral judgment comes from a myriad of both conscious and unconscious factors, such as cultural, personal, social and practical factors combined with the information we actually have and our capacity to interpret this information. As we can see in this model, the moral dimension of the decision is not a given. According to Chorus' model, a decision becomes a moral decision because at some point in the process, we identify it as such. I would say that it is the recognition of the "significant other" in the decision as being a stakeholder is what gives a decision this moral dimension.

Mary Midgley and the nature of morality

A philosopher whose philosophical view of practical morality is very compatible with the above figure is Mary Midgley. According to Midgley, if we want to say anything useful about ethics, we need to understand how morality *actually* works for us humans and why. We need to understand the nature of morality in its full blown reality, without trying to reduce it to simple mechanism or theory like utilitarianism, emotivism. Midgley resists philosophical practices that attempt to reduce the complexity of our reality in damaging ways. This resistance against simplification is especially important in her work on morality (Midgley, 1998, 2002, 2003, 2015, 2017) When looking for morality, she begins with looking how humans work, and from there she explores how morality works. Her starting point is the observation of what kind of beings humans are. And for Midgley, human beings are first and foremost animals and more

specifically *social animals*; social animals who are situated in a complex world which is very complicated and muddled. (Midgley,2002)

Human nature, natural needs and learned behavior

In "Beast and Man" (2002) Midgley argues that ethology and biology show us that humans are social animals. In spite of the fact that violence and warfare are daily ingredients in the headlines of our news, it is obvious that we are much better fitted to live socially than to live in anarchy or alone. To know this is true we only need to go back to the Covid pandemic, which has shown us, that all of our most interesting and most valued occupations are social ones. So Midgley is right when she says that 'society is the condition of man's living at all, let alone living naturally. (Midgley 2002,p47) and that 'Rousseau's or Hobbes's state of nature would be fine for intelligent crocodiles, if there were any. For people it is a baseless fantasy.' (Midgley 2002,p47)

To live successfully as *social* animals, per definition, there must be some sort of natural affection and communicativeness; it is in the nature of social animals to cooperate and 'care', even if in a very basic sense. With the natural basic need to take 'the other' into account we can see how the basis for morality lies in our nature as social animals. This claim is supported by extensive observational research of the behavior of other social animals like wolves, dolphins and

elephants, but more relevant for the human case is the research of biologist Frans de Waal who shows in "Primates and Philosophers" (2016), that in our close evolutionary relatives essential building blocks of morality like empathy and a sense of reciprocity are already present.

When it comes to our natural needs, Midgley points out that for humans, as for any kind of animal, a quite limited way of life is suitable, a range that it must quickly find and keep to. And so for each animal their basic needs and with that basic wants are given (Midgley, 2002). And if we want to say something is good or bad for human beings, or for any other species, we must take our or the other species' actual needs and wants as facts, as something given. Since our basic repertoire of needs and wants is given, we could say that in a basic sense our behaviour is determined. As human beings we cannot want or need things that lie outside of our human nature. We can reject needs and wants natural to us (like refusing to eat, or live a solitary life) but we can not acquire a new set at will (Midgley, 2002).

However, this kind of determinism does not imply that our human behavior is also practically determined. Midgley gives the example of bees (Midgley 2002, p50-51), whose behavior is genetically programmed into great detail. Bees will show the same behavior correct in every detail towards other members of their species, like their honey dance, even when they were

reared in isolation from other members of their own species. They are programmed in a very 'closed' way, without much room for improvisation and learned behavior. In contrast, when we look at humans, we don't need to rear children in isolation from other humans to know that those children once matured will not show normal social behavior when brought into contact with other humans. While parts of the human behavior pattern are innately determined, others are left to be filled in by learning from experience. Unlike bees, humans are programmed in a more general way, rather than in full detail. Midgley argues, based on observations in the field , that we can say that the more complex, the more intelligent creatures become, the more those creatures are programmed in this general way (Midgley 2002, p51). And as a species, humans are very complex and intelligent and are born with certain powers and a strong wish to use them. But it will need practice, time and (often) some example before we can develop those powers properly. The bottom line is that compared to other species, for humans there is, within the boundaries of their species' needs, significant room for learned behavior.

Conflict, motives and reason

As said above, human nature comes with natural human *needs*, these natural human needs express themselves in wants and those *wants* that are sufficiently strong that they move us into action or deliberate inaction express themselves into *motives*. And whenever we want to decide what is good or bad for humans, we are confronted with the fact that our needs and wants

conflict and we can have powerful motives for different actions at the same time. And according to Midgley it is *reason* that provides the solution to such conflict (Midgley, 2002).

Midgley emphasizes that wants are not the same as random impulse. Wants are articulated, recognizable aspects of life; they are the deepest structural constituents of our characters. Wants are personal expression of our natural needs. Conflict arises when we need to *decide* on which want to act. What goes on in deciding what to do, is not just any inexpressible clash of feelings. A clash of feelings could be the problem, but does not provide the solution. The solution needs articulate thinking. Unlike jumping one way or the other because you feel like it, deciding involves thought. And so choosing to act one way deliberately, with a full understanding that you could do otherwise, and after explicit reflection on the alternatives, is a very different thing from doing it unthinkingly.' (Midgley 2002, p249) It is the openness in our genetic programming that provides us with this opportunity.

In dealing with a conflict of motives, we have no other option, Midgley argues, than to *reason* from the facts about our wants and needs that those motives originate from (Midgley 2002, p180). She argues vehemently against the popular idea that we cannot reason effectively in moral matters; reason for Midgley is a name for organizing oneself. When there is a conflict of

motives, one desire *must* be restrained to make way for the other. It is the process of choosing which that according to her is rightly called reason. For Midgley, reason and feeling should not be seen apart from each other. Both are aspects of all our motives:

> *Feelings themselves have a form, and one that fits the matter. In fact, of course, it can be our duty to feel in one way rather than another … Practical reasoning would be impossible were not some preferences "more rational" than others. Rationality includes having the right priorities. And deep, lasting preferences linked to character traits are formally a quite different proposition from sharp, isolated impulses.(Midgley 2002, p249)*

Because of our nature as social animals, naturally we will have feelings of empathy and care *for* the other. we will have needs and wants that concern significant others, and those needs and wants can conflict with our other needs and wants. In such a case we would say that the conflict is a moral conflict. But we also have needs to receive care and empathy *from* the other. Selflessness is as natural to us as the wish to be cared for. But also wants that are seemingly less in line with our social nature, like aggression must in a way be part of our nature; as are more basic egocentric wants that are related to needs for immediate survival, like water, food and shelter. And in choosing, we need to deal with all of them.

In order to be able to choose between our conflicting motives, we have to set them in some order of priority, and in doing so we, as Midgley argues, may be said to be valuing (Midgley 2002, p176) Because in prioritizing, what we do is taking sides. We decide for example that

affection matters more here than honor, or honesty than safety. But at the same time it is also forming a factual judgment about what are actually our deepest needs. Such prioritizing of our motives must make sense in the context of our whole lives, as a social animal, living in a society, over time.

Midgley refers to Butler's work when she argues that, 'Conscience is not a colonial governor imposing alien norms; it is our nature itself, becoming aware of its own underlying pattern' (Midgley 2002, p263). Meaning that there needs to be some preexisting balance and structure among our motives in order to be able to get a grip on them. There must be something in the natural needs of humans, that makes that for them for example honesty and justice are important and valuable elements. Reason recognizes this, and does not create these values. (Midgley 2002)

And in choosing what to do when in conflict, we sometimes need to sacrifice one part of our life, and maybe of other people's lives too, to another. And the only ground on which we can do this, Midgley says, is that the one want is more fundamental, more central to us and to humanity, than the other (Midgley 2002, p182). We, in effect, have to decide what person we want ourselves and other persons to be, what our life and society would look like. And this is

basically what it means to act upon your conscience. And in the case of kicking another kid, from a certain age we believe that under normal circumstances, a child should be able to prioritize the desire not to hurt someone, recognizing the other kid as a significant other with its own needs, above the desire to kick someone out of anger or frustration; the child's conscience tells him not to hurt other people. Both desires are equally natural, but differ in priority. If a ten year old shows no care at all for the feelings of other persons we tend to think that the child is not healthy. But most people would equally agree, that if a ten year old is never angry or frustrated we also tend to think that something's wrong. And being able to choose one motive over the other, to intentionally *not* kick the boy for good reasons is what makes humans agents. We are free to prioritize one motive over the other, and as such are responsible for our choices and can be held accountable for them.

Pre-existing systems, cultures

Human needs are multiple and so are our wants, and as said, all those wants are bound to conflict. These conflicts do not only take place in rare extraordinary cases; they are present all day round while we are leading our quiet everyday lives. And because most conflicts, though they can all be seen as an expression of our human needs, due to our complex natures and the complex societies we live in they are not necessarily easy to get a grip on. So in order to lead our everyday lives while trying to deal with these conflicts of motives, in order to be able to *live our*

lives, we need right from the start a system that helps us to make sense of it all somehow. We need some pre-existing scheme of priorities and a *culture*, according to Midgley, is a device for coordinating, fixing, and developing such systems (Midgley 2002). Luckily we can use the whole range of experience our society has stored for us together with the responses of those around us to help us settle what are after all not just private questions, but questions that concern everyone.

And so humans, unlike bees, need to be reared in contact with other humans to grow up as full blown human beings. They need to *learn* this scheme of priorities in order to be able to live successfully together with other humans. Most of the time, throughout the day we manage to do so pretty easily. Our scheme of priorities guides us through the day without much trouble. However, sometimes, people do disagree about what is good; it is, indeed, the kind of thing that makes them argue. Not just do people disagree with each other, humans also disagree within themselves. And sometimes even after extensive reasoning these conflicts remain insoluble. In such cases we must pick and choose, leaving at least one person or one part of ourselves dissatisfied even within a scheme of priorities, within a culture. There is not always a solution to a conflict of wants, not even within a shared culture. There is an openness in morality *because* of the openness in our genetic programming. Our human natural needs may be shared as human

beings, the openness in our genetic programming makes that our wants, despite that there often is a lot of overlap with other people's wants, are at the same time in fact very personal.

All choosing in Midgley's account, is valuing. But I would say that only those decisions are moral decisions, that concern in some way "the significant other" meaning something outside of ourselves that we consider to be so significant that we attribute value to it independent of ourselves. The significant other has value as an end in itself, not as a means. Most obvious such significant others are other human beings, but in fact it could be anything: the nature around us, animals, but even a holy rock, or something more abstract as our society. A lot of what we add value to outside of our close relatives is dependent on the scheme of priorities that I talked about above, a scheme that we tend to fixate in a culture. What we add value to, and with that how we prioritize our motives can differ enormously between different cultures and different contexts. Despite these differences, if morality is the expression of our natural needs, all priority schemes fixated in cultures and in different contexts can be traced back to our human natural needs. This is why I believe that, cross culturally and in varying contexts, we can still have meaningful discussions about morality even if on the surface our moral priority scheme seems to differ insurmountably.

Objective versus subjective judgments

When it comes to the moral judgments I want to make a distinction which I believe is relevant for this research and which I would like to add to Midgley's account. This is the distinction between *objective* (distanced) and *subjective* (in the moment) judgments. Objective moral judgments are characterized by the fact that the actual context in which the decision needs to be made is absent. Objective moral judgments are hypothetical or retrospective in character, no decisions are made that have immediate impact. Objective moral judgments are used to assess our own moral decision making, are being used in discussions on moral matters and also objective judgments are used in making and assessing laws, protocols and guidelines to be used in specified situations. Objective moral judgments are very much what we would think the world should be like on a general scale. How we should shape society. And I would like to argue that a large part of the motives in objective judgments respond to our natural need for fairness; we use objective judgments in order to prevent undesired personal reasons or lack of knowledge to play a role in the choice made. The call for fair judgment goes back I would say to our need as social animals for a social structure. Humans tend to accept that there is a hierarchy in society and accompanying norms with respect to that hierarchy, but we want to be treated accordingly. Human beings have the need to be treated equally among their equals, which makes perfect sense from a social perspective. Why would you stick to the social rules especially if there is no immediate reward from this behavior, if other individuals from the same social

position do not do their job? Of course the definition of what is fair and what isn't in practice is something that changes along with changes in social structure. For example, in strongly egalitarian countries like the Netherlands and Denmark, not much difference in treatment between persons on basis of the position in society is generally accepted. My use of the word objective need not mislead us into thinking that such a judgment is in any sense more 'objective' as in being more truth apt. As I said, what is considered fair and what isn't is related to what is considered to be the socially accepted structure at the time.

Subjective judgment is not what we would do or should do as seen from a distanced point of view; it is what we think we should and would do in the moment taking everything into account that is going on *in* that moment. Subjective moral judgments are those judgments that lead to immediate action. Even if, in the moment, we follow up on an objective moral judgment described by law or protocol, it is still a subjective judgment doing just that; the objective moral judgment should be seen as input for the subjective moral judgment. A subjective moral decision need not be an isolated process, it can be the result of other objective and subjective moral decisions; the result of a chain of decisions and judgments. In making subjective moral decisions we often take objective moral judgments into account as one of our motives to be weighed and amongst others the need for fairness will thus be accommodated for in our reasoning. When considering taking the job of being the executor of the death penalty,

we will probably take into account our objective general stance on the death penalty. However when actually handling the needle, at that moment there is room for other motives in the reasoning when deciding whether or not to proceed with the injection. For example, feelings of empathy with the prisoner, or a strong repulsive feeling against killing a person, properly understanding the consequences of your decision. Or in other words having your conscience speak. These feelings are context specific, in-the-moment motives for acting, and so need to be weighed in. And many of these motives are highly personal and trace back to the openness in our genetic programming. They can not be known, they may be predicted, but we only know for sure what they are when the actual moment is there. These motives are as much originating from an underlying pattern as motives in objective judgments are.

And it is here that there is room for another expression of our natural needs. The need to be seen and recognized as an individual, as a person with her own specific needs and wants; not only for the person on the receiving end of a moral decision but also for the moral agent herself. Even though human beings are part of a social structure, simply being recognized as part of society is not enough. There is also the need for real connection, a need to be *seen* by the other person as a person and be cared for by the other person as the individuals that we are. And we want to be recognized as moral agents who actually make the decisions; humans in a more or lesser sense want to have responsibilities and we pride ourselves with having done well in

making those decisions. It is not without reason that we speak about mind-numbing work and the effects such jobs have on mental well-being. How strongly this need for personal connection and individuality is expressed in morality differs across cultures.

Taking into account the messy nature of reality, we cannot expect the distinction between objective and subjective judgments to be a hard drawn line of course. But the main point I want to make with regard to the subject of this dissertation, is that in making moral decisions there is always a certain tension between the needs expressed in objective judgments concerning a situation and the needs expressed in subjective judgment. Most people will recognize the situations where we need to sacrifice our more emphatical motivations with respect to a particular person and act for the greater good, but likewise we recognize the opposite. That is reason at work in weighing our motives.

In this chapter I've tried to give an account of what I believe to be the nature of moral decision making. In short, human morality is the expression of our natural human needs; moral judgments arise as the result of conscious reasoning while prioritizing our motives for action that trace back to those natural needs. And moral judgments can roughly be divided into objective and subjective judgments. In this account of human nature and morality the concept of freedom entails the ability to *choose* between conflicting motives by the use of thought and reason. Who we are as human beings defines what our possible motives are, the openness in our

programming makes us free in the sense that we can recognize and choose between our conflicting motives. The openness in our programming and with that our freedom of choosing, provides for different people making different choices in the same situations. And the fact that we can consciously choose what we do in certain situations is gives us agency and what makes us accountable and responsible for our acts.

In the next chapter I will assess what kinds of MDMAI may have the potential to be normative for its users given the account of the nature of morality, after which I will move on in second part of this dissertation to analyse how such normative MDMAI may influence the way we make our moral judgments and decisions and what that entails for freedom, responsibility and accountability.

Chapter 2

Moral Decision Making Artificial Intelligence

For my research I will consider different types of possible Moral Decision Making Artificial Intelligence (MDMAI) and the possibility of such MDMAI to be normative for the agent using it. But before I can do that, it needs to be clear about what I understand under Moral Decision Making Artificial Intelligence. Beginning with what I understand as Artificial Intelligence: Kaplan and Haenlein (2018) define Artificial Intelligence "as a system's ability to correctly interpret external data, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation". Using this definition, when the AI is designed to support us in moral decision making by giving us the right course of action, I would like to call it Moral Decision Making AI. This doesn't mean that the AI is itself the moral agent, because taking into account the nature of moral decision making as described in the previous chapter, for the AI to be considered as such the AI would have to identify a significant other with respect to itself in coming to the right moral decision, and should be considering its own motives in prioritizing. While the probability of this is definitely worth philosophical investigation, for this research MDMAI is considered as a technology *being used* by moral agents as a tool for moral decision making.

For such MDMAI to be normative for its users, the outcomes of the AI need to be recognized as the right course of action in specified moral choice situations. If this is the case, the AI could be said as being actually making the moral decision for us, and thus could truly be called MDMAI. Below I will consider three different types of MDMAI and to what extend the outcomes of the MDMAI could be considered as being normative for its users.

Possible types of MDMAI

I make a distinction between three possible ways to develop MDMAI: *deep-design* and *shallow-design* MDMAI and *hybrid-design* MDMAI. In deep-design MDMAI the procedure leading up to acts or decisions of the MDMAI follows moral reasoning at every step following a top-down approach in which we have explicit implementation of (moral) decision rules into the artificial agents. Such a theory or rule loaden deep-MDMAI would typically be a static AI, a model that is trained offline and used for a while without continuous feedback. Shallow-design MDMAI would be AI in which a bottom-up approach is being followed. It avoids reference to specific moral theories or rules, instead empirical data concerning moral situations are 'fed' to the system as neutral empirical data, and the system predicts what we, in certain situations identified as having a moral dimension, consider to be the right actions. Such *empirical* shallow-MDMAI needs to be dynamic, continuously fed with fresh data in order to adapt to new situations. Hybrid-design MDMAI merges insights from both approaches. Moral decision rules

are explicitly implemented in the system, in combination with empirical data that inform the AI of the 'common sense' moral judgments of the rule based outcomes.

Deep-design MDMAI

The first possible way to develop MDMAI is rule-based. Through implementation of rules the AI can tell us exactly what, according to these rules it is that we should do in a specified situation. The rules in question could be along the lines of a moral theory like utilitarianism or deontological in nature, or some other moral theory or set of rules. The bigger computing power of computers in comparison to humans makes that in following rules the AI is probably more reliable in its outcomes than humans can be in trying to follow those rules. Provided that the rules are unambiguous and not open for interpretation.

How normative could such an AI be? Not very normative, I would like to argue. In her work, Midgley argued vehemently against the simplification of morality through theories. And her arguments against simplification also apply to deep-MDMAI. Like simplified moral theories in general do, MDMAI based on moral theories would be totally out of sync with the reality of moral decision making; it would propose ethical actions that are fundamentally out of touch with our common moral practice. Practices which are counterintuitive for many people,

considered to be *immoral* or in other cases *unfeasible* for human agents. The outcomes of such an AI would only be normative if the input was considered to be normative. Only for those who adhere to the implemented theories would the outcome of the AI be truly normative.

I would argue that even if, as Bogosian (2017) outlines, an AI is developed in which several moral theories would be incorporated in the algorithm, the problem of being out of sync with reality remains. The core idea of such an AI is that there is always a moral theory that comes up with the right action for a specified moral choice situation. Such an AI would for a specified choice situation compute different options along the lines of different moral theories and then each option is scored with a credence-weighted average, thus providing an expected choiceworthiness for each option in a specific situation. The effect of an AI developed along these lines will be that the output is such that it aims to take or promote 'the most broadly desirable action in accordance with various people's values, and will avoid the very counterintuitive actions which are considered to be problems for various moral theories' (Bogosian 2017, p598).

Even though this kind of programming seems more in line with how we actually make moral decisions, this type of MDMAI faces serious problems. First there is the question of how the

credence weighted averages of moral theories are judged, if not against some external standard? What if there is disagreement about those standards? And second, we don't tend to make moral decisions on the basis of moral theories alone. Our natural needs *may* be expressed through the use of moral theory, but are expressed in many other ways as well such as character traits, or pre-existing fixed systems of priorities, such as (sub)cultures or religions which should be given weight in moral decision making as well. This type of AI would not very likely be considered as normative for its users since the output would still have to be weighed against other non-theoretical reasons for acting. This type of deep-design AI would only tell us what we should do according to the implemented rules, whether or not that is normative for us, depends on how we personally value these rules.

Shallow-design MDMAI

Shallow-design MDMAI, unlike deep-design MDMAI is not based on any moral theories, at least not directly. It is rather an empirical attempt to address and embed human moral reasoning in AI Systems. It relies on the idea that actual human morality should be reflected in the AI algorithms. Human morality would therefore first need to be empirically identified and subsequently embedded in the AI system (Awad et al. 2018). So shallow-design MDMAI is grounded in empirical data of how we actually practice morality or on our moral beliefs of how we should act morally. The basis of such AI is empirical data of our actual moral decisions

and/or of our moral believes on which machine learning models are trained to extract implicit relations.

Schramowski et al (2020) proposes for example to develop a Moral Choice Machine that is trained on different temporal news and books corpora from the year 1510-2009 which demonstrates the evolution of moral and ethical choices over different time periods. By training the machine on different cultural sources such as the Bible and the constitution of different countries, the dynamics of moral choices in culture, including technology are revealed. Or the AI might take shape along the lines of the Moral Machine Experiment (Awad et al 2018), an empirical research in which preference data of 1.3 million respondents from over 200 countries and territories of the world was compiled in the context of moral dilemmas faced by autonomous vehicles. Global moral preferences are then summarized and individual variations in preferences, based on respondents' demographics, are documented. As a result they can report cross-cultural ethical variation, and show that these differences correlate with modern institutions and deep cultural traits. Through feedback, the system could also be taught to revise the moral bias picked up from the original data or even manipulate the moral bias itself (Schramowski et al 2020). These types of AI will be based on Machine Learning models and will be able to *predict* what we would do, or believe we should do in specified choice situations.

Could such a prediction of our own beliefs or actions be normative? The first question that comes up is how accurate these predictions are and the second even more important question is how we should value these predictions? In theory such Machine Learning systems can be extremely accurate in predicting, however in order to be that accurate such AI systems as described above require vast amounts of historical data and capturing every preference gives rise to practical runtime problems. Secondly, and more importantly, the models used in the learning process are opaque and so we remain ignorant about how a moral decision came about, thus hampering interpretability and accountability. This is problematic for both objective and subjective moral decision making. As Midgley has shown, moral decision making is not simply acting on a whim, it is the result of conscious reasoning leading to objective judgments or subjective judgments (Midgley 2002). As we have seen part of human morality is not just the making of the right decision, but also reflection on moral decisions and on the reasoning leading up to a moral decision; moral decision making is the prioritizing of motives for action through reason. So in order for AI to be normative, we need to know why it should be normative; we need decisive reasons for the advised acts. Most self-learning systems, though highly accurate in predicting decision making, they are 'black-boxes' when it comes to interpretability (Smeele, 2023). Just an accurate prediction of what we would or believe we should do, is not enough for us to be normative. In order for the outcome of an AI to be normative we need to agree with the reasons behind the outcome. What we need for

normativity is an AI that can tell us our believes *along* with providing the right reasons for holding that belief.

Hybrid-design MDMAI

As both deep and shallow AI have been eliminated as possible candidates for normative MDMAI, and taking the criticisms outlined above to both deep and shallow-MDMAI into account, what we might be looking for is an AI that reflects actual human morality more; AI that is empirically based in that it is grounded on actual human moral decision making, while at the same time being transparent about the rules and considerations that are used in coming to the moral decisions.

One possible way to design such MDMAI is to use discrete choice analysis to codify human moral preferences and decision rules and embed them into a moral decision making AI-system (Martinho et al, 2021). Discrete choice analysis works with mathematical models that formally describe and explain the decision-process of an agent or a group of agents that make a choice between two or more mutually exclusive discrete alternatives from a finite but collectively exhausted choice set. Since many moral behavioral responses are discrete or qualitative in

nature, in those cases discrete choice models are able to analyse and predict the moral decision making behavior of an individual or a group of individuals.

According to Martinho et al (2021), discrete choice analysis does not only provide the possibility to embed human moral preferences and decision rules into an AI-system, it is also able to give an empirical illustration of moral uncertainty, the fact that there is not always general agreement on what the right course of action is; it gives us the weighing of reasons leading to a moral decision. Building on the ideas of Bogosian (2017) as described above, in their model, the choiceworthiness of an action given a certain theory as described by Bogosian, is reconceptualized as the *utility* of an action given a moral theory. The credence of a theory is re-conceptualized into the *share* of a population that adheres to that theory, which corresponds to the probability that a randomly sampled individual from that population adheres to that theory. This can be done of course not just for theories, but for any kind of explicit reasoning and the weighing of motives (Martinho et al, 2021).

Discrete choice models are probabilistic in that they generate choice probabilities for each alternative in a choice set, not because individual behavior is viewed as intrinsically probabilistic, but because moral decisions are almost by definition a result of weighing multiple

attributes (Bartels et al, 2015) and human moral choice behaviour is not a linear process, but 'the result of a subtle interplay between various short- and long-run processes including feed-back loops of behaviors and expectations.' (Chorus, 2015). Amongst other the moral choice behaviour of an individual is influenced by the task environment, the individual's personality, perception of the prevailing norm and expectations concerning the behaviour of relevant others (Chorus, 2015). These empirical observations are as we saw completely in line with Midgley's account of morality. Capturing *all* the information that may be relevant to a single choice for an individual in society in order to have the moral heterogeneity between all individuals reflected in a system is far too complicated for any system to accomplish and such an attempt will almost definitely run into runtime problems.

The approach of discrete choice moral modelling therefore hinges on the assumption that within a population there exist a number of classes featuring different preferences, while assuming relative homogeneity amongst individuals within these classes in terms of behaviors and preferences (Martinho et al, 2021). In this way it is able to reflect the pre-existing systems of prioritizing motives such as (sub)cultures. Moral heterogeneity of society is then reflected in the different classes of the model, but because of the assumed relative homogeneity within classes, the discrete choice model avoids runtime problems. In non-moral settings such as consumer behaviour, discrete choice modelling has already proven to be quite accurate in predicting and

analyzing human choice behaviour, so we have reasons to expect that these discrete choice modelling may be able to do the same for moral choice behaviour.

An AI-system designed on the basis of a utility-based latent class discrete choice model is thus capable of embedding moral theories or moral decision-rules while being transparent about those, and at the same time being able to embed actual moral reasoning based on the weighing of different relevant factors. The different classes in the model define different trade-offs that members of a class are willing to accept, which reflects the prioritizing of motives in Midgley's view of moral decision making. How normative may the outcomes of such an MDMAI be? Assuming that the MDMAI accurately predicts our moral decisions and gives us also our reasons for why we will decide in such and such way, including information about which class I belong to, it may seem that such an AI may give me information about my own motives *and* does the weighing for me. And as such seems to be the most promising candidate in order to be normative for its users. I will go deeper into this question using an existing Hybrid-MDMAI system called BAIT (ten Broeke et al. 2021), which is one example of what the shape and purpose of such a system may entail.

BAIT

BAIT as a technology is designed to be used as a medical decision support system; to make medical expertise available to support medical decision making that involve ethical dilemma's and are characterized by a high degree of complexity, uncertainty, and time pressure. BAIT is presented as a support system that captures and codifies medical expertise, which they define as 'knowing what to do in a certain situation, and being able to explain why' (ten Broeke et al. 2021, p614).

'The objective of BAIT is to make accessible to an expert or group of experts the combined expertise of their peers in the context of a particular decision problem. (2021, p164) Bait is an example of hybrid-MDMAI which uses choices from a pool of experts to identify their expertise. According to its developers, the indirect process of using choices instead of asking the experts to explicate their expertise directly 'is aligned with the notion that humans find it very difficult to explain why they made a certain decision, especially when this involves moral judgments.' (ten Broeke et al 2021, p617)

BAIT is developed as follows: First, after the expert decision is specified, factors are identified that presumably play a role in the expert decision. A choice experiment is designed and the

group of experts is invited to make a series of hypothetical choices based on scenarios mimicking the real decision situation. Then the observed choices are used to estimate the importance weights of all factors, after which the factor weights are visualized, showing how each factor contributes to the experts' decisions in the experiment, and these results are presented back to the experts for feedback. The choice model can then be used to assess particular artificial choice situations where the generated assessments take the form a probability statement. Feedback loops are further used to optimize the model (2021).

As an example of a situation in which BAIT can be used (ten Broeke et al, 2021), is the decision to proceed to surgery versus comfort care for a critically ill neonate with a specific syndrome (NEC), where each case presents the treating medical team as well as the parents with the dilemma of whether proceeding to surgery will still be in the child's best interest. I will use this specific example for my analysis of the BAIT technology in the last chapter of this research. But first I will assess how normative the outcomes of BAIT may be for its users.

The question of normativity

As we saw, BAIT is designed as a support tool in moral decision making, not as a tool for actually making the moral decisions. So BAIT is not designed as normative AI, but as a support

tool to be used in co-existence with guidelines and protocols because according to ten Broeke et al (2021, p619) protocols and guidelines are '*more prescriptive (focusing on what the individual expert "should" do)'* whereas BAIT is *'being 'more descriptive (focusing on what the pool of experts "would" do)'.* However, I would assume that in making the choices by the pool of experts, they come to their preferred choice by using guidelines and protocols as well. What is the advise of an expert worth if not arrived at by interpreting all the relevant data, including protocols and guidelines? In order to have any meaning to a choice maker, the prediction should be based upon reasons that are also reasonable to the actual choice maker.

How does this relate to the idea of the developers that a system like BAIT tells you only what you (being an expert yourself) would do, while the protocols and guidelines tell you what you should do? When an expert who needs to make the final moral judgment in a difficult choice situation, is reading from BAIT that say, ninety percent of the expert pool would do A after having familiarized themselves with all the facts concerned and having consulted the guidelines and protocols, could that not also tell the expert what she should do? Couldn't that be read as, being an expert herself, she most likely *would* also do A and conclude from this that she therefore *should* do A? We could off course say that the experts choices are made as if there were no protocols and guidelines, but of what use would that be? What would be their added expertise? I believe that a fundamental part of expertise is to be able to interpret the available

information and knowing how to apply them in following (more or less formalized) rules and guidelines, while at the same time being able to know when to *deviate* from those rules and guidelines. Then, depending on what latent class the expert belongs to, what the expert herself *would* do according to BAIT may then easily been seen as what she *should* do. For these reasons I believe that even though BAIT is designed as a support tool for making moral decisions, depending of the accuracy of the model, the outcomes of a system like BAIT may likely be considered as normative for agents using the technology. In the last chapter of this research I will asses how the technology of BAIT and its potential normative character may affect our moral decision making given a Midgleyan stance on the nature of morality. I will do all this through the lens of Postphenomenology, which I will briefly introduce in the next chapter.

Chapter 3

Postphenomenology

'*Postphenomenology is the practical study of the relations between humans and technologies, from which human subjectivities emerge as well as their meaningful worlds*. (Rosenberger 2017, p12)

In his book "Technology and the life-world" (1993) Don Ihde offers us a new philosophical approach to technology. This postphenomenological approach weds *phenomenology with pragmatism* (Ihde 1993). The phenomenological in postphenomenology is the focus that lies on human-world relations. From phenomenology, it inherited a way of thinking that concerns itself with human experience and therefore like phenomenology, postphenomenological studies approach philosophical questions concerning technologies from the starting point of deep descriptions of human experience. (Rosenberger 2017, p1) What it also takes is the phenomenological insight, from Husserl's phenomenology, that no perspective on technology is ever a neutral starting point. To 'experience' is always to experience *something*, and it is this realization that gives the phenomenological and the postphenomenological philosopher the firm belief that in order to understand properly human-world relations, we need to overcome the object-subject dichotomy which has played, since the enlightment, such an important role in Western Philosophy (Rosenberger 2017).

Where the postphenomenological approach to technology departs from phenomenology is in its refutation of the idea that there is an original world independent of human experience that somehow we *can* have access to (Rosenberger 2017, p11) and it does not adopt what Ihde calls a 'naïve objectivist account' (Ihde, 1993) in which an instrument or technology presents itself as 'other'. From the postphenomenological perspective there is no pre-given subject in a pre-given world of objects in which technology should be placed in the realm of quality-bearing objects as opposed to and perceived by the human subject. Postphenomenology does away with the idea of technology as an entity that stands *between* the human subject and the world as some alienating force that keeps us from experiencing the world as it is.

Rather, technology should be seen as 'the source of the specific shape that human subjectivity and the objectivity of the world can take' in a specific situation. (Rosenberger 2017, p12) Technology in its role as technology should not be placed in the realm of objects, but as such should be considered in terms of mediation. (Rosenberger 2017, p10) Technology *is* mediation, and subject and object are constituted in their mediated relation (Verbeek, 2005). It is this focus on mediation and mutual constitution that sharply demarcates the postphenomenological approach to technology from classical phenomenology. (Rosenberger 2017, p12) Against the phenomenological idea that technology *alienates* human beings from themselves, postphenomenology places the idea of technologies *shaping* human subjectivity and objectivity

of the world. The idea of 'a privileged access to the things themselves becomes an impossibility within the postphenomenological approach' (Rosenberger 2017, p12).

To analyze how this mediation takes place, postphenomenological studies turn towards "the thing themselves" (Ihde, 1993). This is where we come across the influence of the American pragmatist tradition of philosophy in postphenomenology. Postphenomenological studies find their starting point in empirical analysis of actual technologies. (Rosenberger 2017, p9) As in pragmatism "postphenomenological claims are never about the absolute foundations of reality or knowledge, and never about the 'essence' of an object of study. (Rosenberger 2017, p1)

They look at actual technologies and technological developments and study the relations between humans and technologies. This is where the constitution of subject and object happens, humans and technologies are the result of the interaction between them, they mutually shape each other in the relations that come about between them (Verbeek, 2015). And in many cases, the human-technology relation is part of a larger relation, a relation between human beings and their world. In these cases there is no direct relation between humans and the world, only an indirect relation in which technologies play a mediating role. (Verbeek,

2015). The human – world relation in many situations, is typically a human-technology-world relation (Ihde 1993).

So postphenomenological studies typically 'approach technologies as mediators of human experiences and practices' and investigate 'technologies in terms of *the relations between human beings and technological artifacts*, focusing on the various ways in which technologies help to shape relations between human beings and the world.' (Rosenberger 2017, p9)

Don Ihde (1993) offers us a language to analyze the structure of the various ways in which technologies can organize human-world relations:

1) Embodiment relations: (human-technology) – world

   Technological artifacts are taken into the experiencing, broadening the area of sensitivity to the world; An example of an embodiment relations are *glasses*

2) Hermeneutic relations: human – (technology-world)

   You read the world through technology; An example of a hermeneutic relation is a thermometer

3) Alterity relations: human – technology (world)

The interaction is *with* the technology and not with the world; An example of an alterity relation is a humanoid robot

4) Background relations: human – (technology/world)

The technology forms the context for your experience; An example of a background relation is the sound of a fridge

Multi-stability is another central concept developed by Ihde in post-phenomenology (Ihde, 1993). As explained above, in postphenomenology technology is not thought of as having an essence; *what* technology is, is the outcome of the relations that humans have with it. Technology becomes meaningful in the context of the relations we have with the technology, in this relation both human and technology mutually shape each other. That is why a technology has no essence, a technology is multi-stable in its meaning. Mediations are not *in* the technologies, mediations are the outcomes of the relations that we have with the technologies (Verbeek, 2016). In the relation human-technology-world, all three elements are mediated through the relationships between them.

A theory of mediation

The above explained concepts give us a suitable framework to analyse and understand how it is that technological artifacts mediate and constitute humans and their world. How do specific artifacts mediate human experience and practices? How does the mutual constitution of humans and their world takes places through specific technological artifacts? How do specific technological artifacts shape knowledge or morality? These are all questions concerning *technological mediation* which are of central importance in the work of Peter Paul Verbeek (2005, 2016). In "What things do" (2005) Verbeek shows how technological mediation can be analyzed in two separate dimensions: a hermeneutic dimension and an existential dimension.

From a hermeneutical perspective, the central question in an analysis of technological mediation is 'what role does the artifact play in the manner in which humans interpret reality?' (Verbeek 2005) Mediation is indissolubly linked with a *transformation* of perception, which has the definite structure of amplification and reduction. Technological mediation always strengthens specific aspects of reality and weakens others. Naked perception and perception via artifacts are never fully identical (Verbeek 2005, p131). Such mediations through amplification or reduction takes place in a *microperceptual* dimension of experience which affect experience in a *macroperceptual* dimension (Ihde 1993). It affects how we make sense of the world and vice versa, our macropeceptual experiences in turn affect how we interpret microperceptual

perceptions. A good example of such hermeneutical mediation is the Copernican shift of perception where perception on a microperceptual level induced a shift in macroperceptual experience and from that moment on the renewed macroperceptual framework we found ourselves in has affected how we interpret our microperceptual experiences (Verbeek 2005).

From an existential perspective, what we are looking for is in what way the material environment shapes the way in which humans realize their existence. How do technological artifacts affect the way we spend our day or how we organize our social relations? How do technological artifacts affect the way we see ourselves as human beings? In order to find answers we first need to elucidate the technical mediations of the actions that give shape to human existence, before we can understand how our actions shape the ways in which we realize our existence. However like in the hermeneutical dimension, the mediation is not a one way direction because the form of our existence, in turn, shapes our human actions. Using the examples of speedbumps and bulky key rings on hotel keys, Verbeek (2005) shows how artifacts, through invitation and inhibition shape human actions.

Technological artifacts then, when functioning, appear to be present to human beings in a specific way. Through amplification and reduction, invitation and inhibition they actively shape

relations between humans and the world by transforming both experience and actions. A paradigmatic example in Verbeek's work (2011) is how the introduction of obstetric ultrasound mediated our perceptions on a microperceptual level, constituting the reality of a fetus in a radically new way. With that it brought new ethical questions and responsibilities, reorganizing ethical practices and decisions. Obstetric ultrasound mediated the character of morality with respect to the fetus and the expecting parents.

In the next and final chapter I will analyse how the MDMAI technology of BAIT may shape the and mediate our moral decisions in specified situations and our social relations, applying the postphenomenological concepts described above. And in line with the postphenomenological approach I will take the BAIT-technology and the ways it can be used as a starting point for my philosophical analysis rather than 'applying' theories to BAIT.

Chapter 4

Postphenomenological analysis

In this chapter I will assess the hybrid MDMAI technology of BAIT and its impact on the nature of moral reality as seen from a Midgleyan stance through a postphenomenological lens. In the first section I will establish what place BAIT occupies in the human – technology - world relation and how this relation manifests itself to the world. After which I will move on to asses the technology of BAIT by engaging the central postphenomenological concepts of amplification and reduction and multi-stabilities. I will analyze how these amplifications and reductions and different multi-stabilities may mediate or transform the way users of BAIT may make moral decisions and the way social relations are organized in those specified choice situations.

Human-technology-world relation

'The objective of BAIT is to make accessible to an expert or group of experts the combined expertise of their peers in the context of a particular decision problem.' (Ten Broeke et al 2021, p614)

If we look at the four ways Don Ihde (1993) specified in which technologies can organize human-world relations, BAIT falls I would argue under hermeneutic relations, since we read the world through this technology.

And it does so in two ways:

1 We read through the technology a prediction of the judgment of a pool of experts and which part of this pool of experts would do what under given circumstances.

2 We get an insight of the supposed factors that play a role in these predicted decisions and the weight each factor has in arriving at this decision.

Having established what kind of technology-human-world relation we are dealing with, we can now assess how these two ways in which we read the world through BAIT amplify and reduce different aspects of the human nature of morality.

Amplification of objective judgments

What do these two ways in which we read the world through BAIT amplify or reduce? And what different multi-stabilities may BAIT take? Considering the distinction I made in chapter 1 between distanced objective moral judgment and in-the-moment subjective moral judgment, I would say that the predicted judgment we read from BAIT would fall under the distanced

objective judgment. The input for the model is given by the pool of experts about *hypothetical* choice situations, which makes the judgment necessarily distanced instead of *in* the moment. Taking into account that the feedback loops in finetuning the model would presumably come as input from the *evaluation* in hindsight of the predicted judgments in specific choice situations and from *evaluation* of subjective in-the-moment judgments that were made, all this makes for the predicted expert judgment offered by BAIT to be the experts' objective judgment.

While BAIT as we saw is supposed to be 'more descriptive (focusing on what the pool of experts "would" do)' (Ten Broeke et al, 2021), I would say that taking the above into account, BAIT is being descriptive in telling us in what the pool of experts '*believe they should do in a specified choice situation*'. What the pool of experts would actually do in the specified choice situation, that subjective decision is made in the moment and cannot be captured by BAIT. In comparison with Machine Learning models which are being fed with data about *actual* choices made by the expert *in* the specific choice situations these models would *predict* what the expert would *actually* do; what their real time subjective judgments would be. One may argue that feedback from real-time decisions made by experts, may update the model and make it more predictive of what the experts would actually do, however as I will argue below, the fact that we gain insight in the supposed factors that play a role in our decisions make that feedback on those decisions may nullify this effect.

Given what is said above, BAIT amplifies the objective judgment concerning specified choice situations, while it reduces the subjective judgment in the specified choice situation. And I believe that even though the designers don't mention this explicitly this is also their aim with the technology, to provide the agent with the objective moral judgment of a group of experts concerning a specified choice situation. And in a sense, since the supposed user of the technology is an expert herself, she can read from the technology what her own objective judgment concerning a situation would entail. Below I will assess how this amplification of the objective judgment may mediate the process and outcome of moral decision making.

Amplification of reasons given

BAIT does not just amplify the objective judgment itself concerning a particular moral decision, it also amplifies the reasons the experts give for the objective judgment and the weights they give to these reasons. How does this work? Weren't we supposed to read from BAIT the supposed factors underlying a decision, not the reasons experts would give for their judgments? Yes we were, but I will explain how we may get from supposed factors to reasons given through using the technology.

In the initial design what is implemented in the model are factors that supposedly play a role in the judgment, these factors are specified by the designers and their weights in de outcome are extrapolated from the choices made by the experts. So the factors in the design phase may not resemble the reasons that the experts may give for their choices. Using choices to capture the factors and their weights instead of given reasons is aligned with the notion that humans find it very difficult to explain why they make certain moral choices (ten Broeke et al 2021, p617). And so the factors and weights give us insights in what the reasons behind the choices very likely are. These discrete choice models can only incorporate a limited amount of factors and have difficulty in catching more subtle factors that are hard to identify and capture in hard data, such as motivation or cultural norms. So it is important to understand that though the identification of the factors and their respective weights may give highly accurate predictions for choice behaviour the supposed factors need not be the actual factors and even if they are, there are other factors involved that are not identified by the model. But how do these identified factors become 'reasons given' by the experts as I argued above? Well that is because of the feedback loops that are used to refine the model and through the use of the model as a support system.

While, according to Midgley, moral decision making is the result of conscious reasoning, we can see from the model in fig 1. (Corus 2015) in the first chapter, that this conscious reasoning can be seen as the tip of the iceberg in the entire process of moral decision making. The basis

upon which we build our conscious reasoning is a process that takes place subconsciously which consists of factors that relate to our culture, experience, character, feelings, and so on. Lots of our moral knowledge is 'known' in the same way as we know how to walk or talk. Knowing how these factors influence our conscious decisions is extremely hard to have good insight to, like trying to explain how you walk or talk. This is part of the reason why it is so hard to know why we make certain choices.

Despite this difficulty, humans still tend to want to justify their moral choices by giving reasons for how we got to them. This is important for the forming of our moral judgments; in order to be able to discuss and improve our moral decisions we need to be able to discuss the reasons for our judgments. The call for providing reasons for our moral decisions does not only come from the need to improve our moral behavior, but also because we see the other as agents who are responsible and accountable for our intentional acts and reasons are called for in justification of our acts. However, given the subconscious character of a large part of the reasoning process, the reasons given are not likely to present the whole picture of what lies behind our choices and judgments. This generally does not cause problems in discussions, because many of these subconscious reasons are shared by others in the discussion. But what may be troublesome in making objective moral judgments is that we will not be very likely to give reasons that are not considered to be moral, like racial prejudices. We will give socially accepted reasons for our

decisions and probably sincerely believe that these are our true reasons. To see what I mean consider the idea that for example skin-color is identified as a significant factor in the choices made by the pool of experts. In the feedback loop this would most likely be seen as undesirable and immoral by the pool of experts. Experts in their feedback may therefore likely adjust their judgments in such a way that skin-color does not play a role anymore in their better objective judgment. Undesirable prejudices then will be filtered out in the feedback loops, and the factors then given in the outcome of BAIT will more and more represent the preferred reasons that experts want to give for their judgment. So the supposed factors will at some point in the development stage of the model be transformed into the reasons given for a judgment. And so BAIT will amplify the desired reasons for our choices, the reasons that we are likely to give for our choices. And as we already know, these reasons need not provide the real complete picture since reasons that are hard to identify like those concerning character, internalized knowledge and cultural norms are hard to specify.

But even if the feedback loops make that supposed factors transform into reasons given for moral judgments, this does not entail that what BAIT offers us is an illusion. In adjusting their moral judgments in order for the factors to resemble the desired reasons, the experts may actually improve their objective moral thinking. And the predicted objective moral judgments by BAIT may indeed give us the objective judgments we want to have, BAIT may improve our

moral thinking in this way. Through the use of BAIT experts gain insight into their own behavior and can adjust their own way of moral thinking to more desirable outcomes. It is not a bad thing I would say to filter out racial prejudices from our moral thinking with the help of BAIT. This is one of the multi-stabilities of the BAIT technology; a technology that gives us insights into our own moral behavior and adjusting that behavior to our better conscious moral standards.

This multi-stability of BAIT as a kind of learning tool is one way how it may mediate the moral decision making of clinicians in the specified situations. I can see more ways in which BAIT could be used as a learning tool, for example for students in order to gain insights into the knowledge of experts concerning certain choice situations. But also for clinicians in general to gain insights in other 'bubbles' and the factors and reasons that underlie the decision making in these bubbles. It could give clinicians insight into how different backgrounds generate different weights to the same factors in the decision making. We could imagine that for example the outcomes of BAIT would be different in the Netherlands compared to India concerning the same specified situation. BAIT offers us great insight into our behavior concerning ethical choices we make, and through studying this we may refine our own moral decision making and explore new directions. However there is also a multi-stability I see for BAIT which may mediate our moral decisions into the opposite direction, by mediating our moral decision making

towards conservatism. Which is when using the multi-stability of BAIT as a support tool instead of as a learning tool.

Conservatism

As I explained in chapter 2 in using BAIT as a technology to support her in making moral decisions, the clinician may very likely consider the outcomes of BAIT as being normative for her. Considering the outcomes of BAIT as normative, this will mediate the moral decision making of its user toward conservatism. Mediation towards conservatism is a feature of all AI of which the outcomes are considered to be normative for its users. AI is by its nature conservative, since in its predictions it can only build upon data of the past and cannot take into account brand new data that have no resemblance to data of the past, let alone the absolute uniqueness of each new situation. And AI cannot on its own deviate from the results of the past. As such, relying on the outcomes of AI will freeze the process of development in decision making. The of this conservatism on our moral decision making, totally depends on its purpose and the specified work area. Consider for example a specified situation with high occurrence and relatively low moral impact, like opting for conservative treatment versus surgery in the case of a ruptured meniscus. Even though a misjudgment by AI could lead to physical discomfort, the moral impact is relatively low and the common prevalence of this injury is so high that a

clinician would very likely not come to different outcomes in her subjective judgment than the AI in the objective (conservative) judgment.

BAIT however was developed specifically for a specified choice situation with low prevalence and high moral impact. And this causes problems for accuracy not only for hybrid MDMAI but also for models based on machine learning specifically designed for such choice situations with relatively low prevalence. Because of this low prevalence of the specified choice situation, data considering such situations is therefor also relatively rare. Not only the available data for the technology, but also the data available for the pool of experts; their expertise is in a sense less expert than the expertise of the orthopedist faced with a ruptured meniscus. The accuracy of the prediction of any behavior including our own is much more reliable in very common situations than it is in rare cases. The question then is, given that morality is not some static process and that judgments change overtime together with changes in society or when new technology and information becomes available, how accurate any model can predict behavior in such cases with low prevalence. Related to this problem of predicting our own behavior, is the observation concerning Discrete Choice Models that 'given the hypothetical nature of the choice experiments, the choice tasks may not have enough consequentiality for decision-makers whose decisions do not reflect their preferences as they would in real-world decisions' (Smeele et al 2023, p6). This, I would say, is especially pressing in rare cases with high moral impact and

where we don't through experience know what the possible consequences may entail, and how we react to that.

Development in moral decision making, the gaining of new insights, acquiring moral knowledge, these things do not come about only through discussions of the reasons we provide for our moral decisions. Human behavior as machine learning AI and discrete choice modeling has shown, is highly predictable, however due to the openness in our genetic programming, human behavior is also unpredictable. And even though this unpredictability affects maybe 0,01% of all our decisions, the impact of these decisions cannot be underestimated I would argue. Many of our new moral insights come from the objective judgments being overruled in our subjective judgments and the unexpected results that come from this. Many of our insights arise from the unpredicted weight of certain context specific motives in our subjective moral decisions, or in other words following your conscience, and the outcomes of these unpredicted moral decisions. We learn from each unpredicted behavior and in these rare specified cases each unpredicted decision and with that each new insight, whether positive or negative, has relatively more weight in our moral development concerning these specified cases in comparison to specified situations with a much higher prevalence. And so the multi-stability of BAIT as a technology of which the outcomes are normative for its users in these specified cases will freeze

the acquiring of new knowledge concerning moral decisions in these cases to a higher extend than in cases with higher prevalence.

Accountability

We may think that this knowledge about the accuracy and the conservative nature of the outcomes of BAIT, would make that the expert is more likely to follow her conscience in specified situations. However, even though that may be the case in obvious mismatches between the outcome of BAIT and the conscience of the clinician, in general I don't think this will the case. This is not only due to the idea that the expert may consider the objective judgment coming from BAIT as her own, but also with the possibility of the multi-stability of BAIT as a tool for justifying our decision. Since moral decisions concern significant others and we consider the one making the decision as an agent who intentionally made her decision, justification of moral decisions is an important part of our social relations. We want to know why someone did something that has affected us, especially when the outcome of that action did not have the desired result, like the suffering or death of our premature child. We hold people that we consider agents accountable for their decisions, and in asking someone for reasons, we like to hear facts, we want to hear reasons that can be checked.

But as is argued above, it is hard for humans to know what their motivations are in coming to a moral decision and this is largely due I would say, not just to the fact that a large part of the 'why' comes from the subconscious framework they operate in, which is hard to come by, but it is also because not all conscious motivations can so easily be quantified in data or easily put into words. We have knowledge and motivations that are for example based on experience (internalized knowledge) or on empathy or on the ability to read other people's capabilities and motivations. And even though this is real knowledge we all know that some things are hard to nail down into numbers or words; it is hard to substantiate this knowledge with hard facts. And when something is hard to nail down in words or quantified data, it is also hard to explain our reasoning when reflecting on our use of this knowledge. It is hard to give reasons for how I walk or talk, but that doesn't mean that I don't know what I am doing. Just like it maybe hard for an experienced fire-fighter to explain why he left the building other then saying that he knew it was going to collapse; experience is hard to quantify. Or ask how we know whether someone is in pain; empathy, people's skills, good judge of character, all extremely important knowledge, but hard to quantify.

And when asked for justification of a decision that turned out to be a bad decision, saying that you knew what you were doing may not convince people. And when we know that we need to reflect upon and substantiate our choices later and will be held accountable for these choices,

we may be reluctant to act upon our more elusive knowledge and instead rely on knowledge that we can quantify and use in justifying our choices. This effect may be amplified when lawsuits lurk in the background and verifiable facts substantiating your judgments are being asked for. Try to explain why you went against the predicted advise of a pool of experts? After all this information, though conservative in nature, is still the objective moral judgment coming from a pool of experts. And as a clinician who has to make the final decision, you need to be very confident of your own expertise and knowledge in order to go against the advise of BAIT, especially when in doubt and without having the opportunity to discuss this particular case with the pool of experts consulted in BAIT in order to fully understand their reasons. Of course, the clinician will have a medical team that also advises her and with whom she can discuss the matter. But the members of the team also need to take the advise from BAIT into account and the stakes are high, letting a child die is a morally pressing question and everyone needs to come up with convincing arguments to go against the advise of BAIT. Thus it seems not farfetched to assume that in hard cases, not being sure and under a lot of pressure the clinician and her team will be pulled towards the outcomes of BAIT not only because of the outcome of BAIT being normative for the expert, but also because of the multi-stability of BAIT as a provider of clear cut information for justification of decisions made in specified choice situations. This may not only freeze the development of our moral decision making because of the conservative tendencies of AI, but also impoverish our moral decision making by leaving relevant though

more elusive information concerning the specified choice situation out of the equation in hard cases.

Social relations

As we have seen above, technologies like BAIT which are being based on discrete choice models can only incorporate a limited amount of factors and have difficulty in catching more subtle factors that are hard to identify and capture in hard data. BAIT therefore in giving the experts' objective judgments, disregards motivations that relate to unquantifiable elusive information and has room only for motivations that are substantiated by quantifiable information and even then only those motivations that carry the most weight in the prioritizing. And on top of that, when we are relying on technologies like BAIT in making difficult moral decisions for the reasons described above, in our subjective moral judgments we will also disregard or put less weight on these elusive motives and the knowledge related to them.

This disregard of motives may transform the social relations in these specified situations, because many of these elusive motivations for moral actions relate to our social nature and our natural need to be seen as a person and to be cared for. These motives are especially present in our subjective moral judgments because they come to the fore in our personal connections and

are substantiated by what we may call people's skills, empathy, sensitivity, imagination. And disregarding these motives, or putting less weight on these motives in the prioritizing of motives when making decisions with high moral impact will very likely transform the social relations between those involved. This mediation of the social relation between clinicians and patients may be expressed in a mediation of how the clinician sees the patient and how the patient experiences this mediation. Relying in a difficult moral decision of life and death on the outcome of technology like BAIT, may in the eyes of the clinician transform the patient from an analogue person into a bundle of quantified data as input for the model.

As we saw in chapter 1, humans are social animals and part of the definition of being a social animal is that there is genuine care for other members of society. Humans are not simply beings who endure the presence of the other because it may serve them well at some point, like in a Hobbesian society; human beings want to live with each other, we need social relations as we need air for breathing. And caring for one another is the expression of this natural need. How this care expresses itself can be very different, but the basis lies in recognizing the other as significant and its wellbeing as intrinsically important to us. We are able of caring for the other, because we as humans have the capacity to understand how other people feel. We can see whether someone is in pain, or tired, or sad or annoyed, angry and so on and take this knowledge into account into our deliberations on how to act.

In not making use of our people's skills and seeing our patients as a bundle of data input for our model, may make for fairer judgments overall, but may also make that the patient feels dissatisfied because important natural needs are not met with in the moral judgment of the doctor. Not feeling seen or heard by a professional while being at the receiving end of life changing decisions, is a common complaint in many areas like for example in courtrooms or in dealing with social services, but also in hospitals. When a clinician decides that it would be the best choice not to operate on a child, and the child will die because of that, most people don't want to be informed about this decision by a computer but by the clinician herself. And they would want her to listen to them, they would want her to look them in the eye, showing that she cares. Showing that she cares about their pain and that even though she recognizes this pain, she will still advise not to operate out of care for the child. And in courtrooms it is proven that the defendant is more willing to accept punishment when he has the feeling that he was *seen* by the judge. And in the same way I could imagine that patients are more willing to accept an unwelcome decision from a clinician when they have the feeling that the clinician sees their grief, and cares for them. And there is a reasonable chance, that the more technologies are employed in medical decisions across different disciplines and with more or less moral impact, the more the personal relation between doctor and patient will be mediated. Even though the

use of these technologies may lead to the best possible treatments for the patients, patients may feel more dissatisfied with their treatments.

Freedom

One last thing I would briefly like to mention is how the abovementioned multi-stabilities and the way they may mediate in a microperceptual dimension the moral decision making in the specified choice situations, may in turn mediate the freedom of the experts overall in their moral decision making. This mediation will become more relevant if in the future there will be more an more accurate hybrid-MDMAI technologies available in different complex moral choice situations. This may lead to a generation of doctors who have not developed specific professional people's skills at all having always had BAIT like technology at hand in complex moral choice situations. Why learn how to navigate on the stars if you have a compass, or why learn reading a map when you have Google Maps? People's skills cannot be learned by AI; AI cannot put itself in the other person's shoes and understand that it needs to go against the rules and take a risk based on what it learned from imagination. People's skills, empathy, sensitivity, imagination, alle these capacities stem from our social nature, these are skills that make us human as opposed to a probabilistic model like AI. Developing these people's skills is not a given but something that will need practice and some people will be more talented at this then others. When clinicians rely more and more on predicted judgments like those from BAIT, the

less need there is to learn to *see* the person involved at the other end. The patient will become more relevant as data for the model than as a person we need to get to know and understand as a person in order to know what to do.

Knowledge makes free, a clinician who has expert knowledge of a situation is more free in her choice whether or not to proceed to surgery in comparison to a layman like me. Freedom in moral decision making entails the ability to *choose* between conflicting motives; the question is how loosing knowledge from people's skills in medical decisions may mediate our freedom as clinicians in making medical decisions, and as such may deprive us of coming to new unexpected insights. These are notoriously difficult questions to answer, because diminishing personal freedom in one area, may enlarge freedom in other areas and also because the development of hybrid-MDMAI technologies is still in its infancy. But I would say that gaining insight in these mediations on a macrolevel is worth pursuing in order to know in time what mediations of our freedom and our social relations we are willing to accept considering our nature as social animals.

Summary

In my analysis I have singled out three multi-stabilities for BAIT: as a learning tool, as normative MDMAI and as a reason-providing technology for justification. In all three of these multi-stabilities the amplification of objective judgments and the reasons given for those judgments will mediate moral decision making towards conservatism. Relying on the objective judgments from BAIT will also mediate moral decision making in the sense that more elusive motives are not or on a lesser scale being taken into account into prioritizing motives in the subjective moral judgments. This happens because BAIT cannot accommodate these hard to identify factors, but also because the overvaluing of the outcomes of BAIT for the reasons given above, will undervalue the more elusive motives that tend to come up in our subjective judgments. And since, as I argued, many of these elusive motives come forth of our need for social connection and the need to be seen and heard as individuals, undervaluing these motives or ignoring them, will mediate the social relations between those involved in the specified situations. And finally, in the future, having more and more accurate hybrid-MDMAI technologies available, the loss of knowledge concerning these elusive motives may on a larger scale mediate our personal freedom in making moral judgments overall and may result in the freezing of the process of moral development and with that hampering moral progress.

Bibliography

Artificial Intelligence Act: Deal on comprehensive rules for trustworthy AI: News: European parliament (2023) *Artificial Intelligence Act: deal on comprehensive rules for trustworthy AI | News | European Parliament*. Available at: https://www.europarl.europa.eu/news/en/press-room/20231206IPR15699/artificial-intelligence-act-deal-on-comprehensive-rules-for-trustworthy-ai.

Awad, E. *et al.* (2018) 'The moral machine experiment', *Nature*, 563(7729), pp. 59–64. doi:10.1038/s41586-018-0637-6.

Bartels, D.M. et al. (2015) 'Moral judgment and decision making', *The Wiley Blackwell Handbook of Judgment and Decision Making*, pp. 478–515. doi:10.1002/9781118468333.ch17.

Bogosian, K. (2017) 'Implementation of moral uncertainty in Intelligent Machines', *Minds and Machines*, 27(4), pp. 591–608. doi:10.1007/s11023-017-9448-z.

Chorus, C.G. (2015) 'Models of moral decision making: Literature review and research agenda for discrete choice analysis', *Journal of Choice Modelling*, 16, pp. 69–85. doi:10.1016/j.jocm.2015.08.001.

De Waal, F. (2016) *Primates and philosophers: How morality evolved*. Princeton: Princeton University press.

Ihde, D. (1993) *Technology and the lifeworld: From garden to Earth*. Bloomington, IN: Indiana University Press.

Ihde, D. (1995) *Postphenomenology: Essays in the postmodern context*. Evanston, Ill: Northwestern University Press.

Kaplan, A. and Haenlein, M. (2018) *Siri, Siri, in my hand: Who's the fairest in the land? on the interpretations, illustrations, and implications of Artificial Intelligence*, *Business Horizons*. Available at: https://www.sciencedirect.com/science/article/pii/S0007681318301393.

Martinho, A., Kroesen, M. and Chorus, C. (2021) 'Computer says I don't know: An empirical approach to capture moral uncertainty in artificial intelligence', *Minds and Machines*, 31(2), pp. 215–237. doi:10.1007/s11023-021-09556-9.

Midgley, M. (1998) *The ethical primate: Humans, freedom, and morality*. London: Routledge.

Midgley, M. (2002) *Beast and man: The roots of human nature*. London: Routledge.

Midgley, M. (2003) *Heart and mind: The varieties of moral experience*. London: Routledge.

Midgley, M. (2015) *Wickedness: A philosophical essay*. London: Routledge.

Midgley, M. (2017) *Can't we make moral judgements?* London: Bloomsbury Academic, an imprint of Bloomsbury Publishing Plc.

Rosenberger, R. and Verbeek, P.-P. (2017) *Postphenomenological investigations*. London: Lexington Books.

Schramowski, P. *et al.* (2020) 'The moral choice machine', *Frontiers in Artificial Intelligence*, 3. doi:10.3389/frai.2020.00036.

Smeele, N.V.R. *et al.* (2023) 'Towards machine learning for moral choice analysis in health economics: A literature review and research agenda', *Social Science & Medicine*, 326, p. 115910. doi:10.1016/j.socscimed.2023.115910.

ten Broeke, A. et al. (2021) 'Bait: A new medical decision support technology based on discrete choice theory', *Medical Decision Making*, 41(5), pp. 614–619. doi:10.1177/0272989x211001320.

Verbeek, P.-P. (2005) *What things do.* University Park, PA: Pennsylvania State Univ. Press.

Verbeek, P.-P. (2012) *Moralizing technology: Understanding and designing the morality of things*. Chicago: University of Chicago Press.

Verbeek, P. P. C. C. (2015). Cover story: Beyond Interaction: a short introduction to mediation theory. *Interactions* (ACM), 22(3), 26-31. https://doi.org/10.1145/2751314

Verbeek, P.P. (2016). 'Toward a Theory of Technological Mediation: A Program for Postphenomenological Research'. In: J.K. Berg O. Friis and Robert C. Crease, *Technoscience and Postphenomenology: The Manhattan Papers.* London: Lexington Books. ISBN 978-0-7391-8961-0, pp. 189-204