Socrates, Aristotle and the Near Future of AI Ethics

Maarten Mintjes

Graduation Project
Media Technology MSc program, Leiden University

Thesis Advisors:
Peter van der Putten and Eduard Fosch-Villaronga

Abstract. After its first stage of establishment, AI ethics has moved into a period of providing structure for the creation of regulation for the development and application of Artificial Intelligence. By now, legal documents are coming into action on a global scale. With this, AI ethics finds itself in a transition towards a stage of answering questions of direction within the constructed boundaries. As AI ethics becomes increasingly normative, we are in need of new constructive frameworks and methods focusing on the shift from restricting poor use of AI towards promoting good and better ways to use AI. In this paper, we locate one of the main obstacles on the path that leads towards the inclusion of all stakeholders in this dialogue. We present a novel conceptual framework and a roadmap that work towards resolving this obstacle, such that the values of the whole AI ecosystem can make their way to the needed ethical debate.

Keywords: Artificial Intelligence · AI Ethics · Virtue Ethics · Socratic Method

1 Introduction

With programmable computers, there came the possibility of creating reasoning from mathematics almost directly. As mathematics seemed near endless in its expressive power, the dream of machine intelligence appeared to have found its means. With this came immediate concerns about intelligence of this kind and about those in power to employ them. Already in 1948, Wiener wrote words that have always remained true in the study and use of Artificial Intelligence: “If we program a machine for winning a war, we must think well what we mean by winning.” [31, p.177]

To provide background for the answers to such questions, the field of AI ethics was established. In this paper, we consider this to be the sub field of machine ethics concerned with the development and application of systems generally considered Artificial Intelligence. Recently, the field has played a large role in laying the ethical foundations for monumental AI regulations on a global scale. Now that these legislative structures are being finalised and implemented, AI ethics will again become more focused on structuring and establishing frameworks that help answer questions of how to act within these structures. With the boundaries set in place and the questions of "what should not be" out of the way, we will move towards direction and questions of "what ought to be within these limits."

We believe that constructive answers to this question will follow from a combination of moral dialogue and ethical frameworks. Furthermore, this dialogue should include all stakeholders to ensure that difficult questions of implementation and design are being handled in a way befitting all members of society. Therefore the research question of this paper is as follows: What are the current obstacles on the path towards a stakeholder-wide moral dialogue surrounding Artificial Intelligence, and how can we begin to resolve these?

Through writing both within and outside the scientific sphere, we have located a gap of shared definitions between all stakeholders as the main obstacle on this path. We present a novel framework and roadmap with the objective of being a structure and trigger for such shared definitions for an inclusive dialogue.

The framework is based on principles as well as virtues. We believe that virtue ethics will play a major role besides the principlism that has been the ethical approach of choice for frameworks to base regulation on in the past years. The nature of virtues is a welcome addition to that of principles, as this gives more ground for resolving existing tensions between the principles. Besides this, we add a layer of topics to allow for a more inclusive nature of the ethical debate. For this, we build on work by Hagendorff [14], earlier work on virtues for a technological society [27], and scoping reviews of principles within the current field of AI ethics [17][12].

The roadmap, specifically the step of coming to shared definitions, is based on the method of Socratic questioning which we see as the best fit for promoting structured and constructive general dialogue. For this, we look back to the earliest applied ethics of Western philosophy [24] and take inspiration from earlier applications of the method to AI ethics [19][28].

In the following section, we discuss the field of AI ethics and its frameworks and ecosystems. We also show our findings regarding the lack of shared definitions that forms the named obstacle. In the third and fourth sections we present and justify our framework and roadmap, respectively. After this, we offer a wider reflection on the approach and results found in this paper before concluding the paper.

2 Moral Philosophy and Ethics for AI

In this paper, we take the pragmatic stance that moral philosophy and ethics are defined differently. We see moral philosophy as the field concerned with the inherent values that we hold, whereas ethics is a formalisation of these values. With this, we mean constructing a frame of discussion for the underlying moral questions. From such a frame we can then, for example, create rules and guidelines, and form the world according to the ethical debate and finally to moral values. This resembles the distinction between the intrinsic and instrumental value of ethical reasoning as suggested by Bietti [6], albeit in a slightly more system-oriented manner.

With this definition in place, she argues for an approach more aimed towards the intrinsic value (investigating moral philosophy) rather than merely focusing on instrumentalising ethical language (solely practicing ethics). This comes in response to criticism of machine ethics as losing its substance. This argument lays the foundation for the stance that we hold for the field of AI ethics: Besides ethical debate, there is a need for moral dialogue, which should include all those involved with AI.

2.1 The Role of Frameworks in AI Ethics

The focus on dialogue has already been present in the first stage of the field of AI ethics. This specifically led to legal regulation, for which we focus on the European Union. In the first stages of exploration by the European Commission, the importance of "dialogue with all relevant stakeholders from industry, research institutes and public authorities" was pointed out [26, p. 1]. In our contribution we find gaps in the structuring of said dialogue, specifically in the definitions of terms used by different stakeholders. This is similar to what has happened in this earlier exploration, although we add a focus on the general public as an active stakeholder in the field.

In combination with other methods, this dialogue, as proposed by the European Commission, led to the first draft of key requirements as proposed by the AI high-level expert group, from which the Ethics Guidelines for Trustworthy AI followed [15]. In turn, the AI Act was drafted which is now in the process of coming into action [11].

This process of ethical implementation following an inclusive moral dialogue exemplifies the different faces of morality and ethics. From moral dialogue follows the construction of ethical frameworks, and from this, legal regulation was created. Therefore this construction can be seen to go hand in hand with the context of the moral questions which the framework is intended to provide structure for. This stance is further worked out in figure 1 later in this paper.

In this process, the connection between moral values and ethical debate can falter. Selbst et al. [25] point out the presence of abstraction errors: acts of considering technologies while not paying enough attention to their social context. Especially excessive formalisation as a part of the nature of abstracting shows this lack of context for our case, as the full meanings of important concepts are ignored. The researchers take as an example the mathematical character that has been ascribed to fairness. However, fairness is inherently social in many uses of the term and there might not be one final definition. Therefore, technical actors should move from seeking a solution to the issue of fairness to "grappling with different frameworks" [25, p.63].

This is a conclusion we take as a starting point, and we aim to further the ability of technical actors to do so. We do this by introducing a framework and a method of grappling with it. Therefore, we also follow the objective of frameworks as put out by the researchers, to “enable process and order, even if they cannot provide definitive solutions” [25, p.63]. From our research, we locate areas in which such process and order are lacking, and we define a framework to enable a transition to a more structured inclusive dialogue following the underlying moral debate.

2.2 The Background of Ethical Frameworks

The framework put out by the High-Level Expert Group (HLEG) on artificial intelligence should be seen in its context, that of the process of creating regulation [15]. This influences the choice of approach in applied ethics. This approach - with the objective of legal regulation in mind - has been one of principlism. As legal regulation has been a large topic in AI ethics over the past years, principlism has played a big role, as can be seen in recent reviews of the field of AI ethics [17][12].

The general use of principles in applied ethics has been noted to be more connected to regulation. This is the case in biomedical ethics, which has a long history of working with a principle-based framework. Beauchamp, one of the main proponents of principlism in bioethics, describes principles as connected to ”regulative guidelines” [4]. In more recent ethical debate specified to Artificial Intelligence, the advantages of principles are also praised for their ability to form a ”useful starting point from which to develop more formal standards and regulation” [30]. Regulation comes back in both these fields as a characteristic of a principled approach.

The connection between principles and regulation can also be noted from a similar focus on restriction that can be found in the field of AI ethics in the past years. In the 84 documents containing principles or guidelines reviewed by Jobin et al. [17], *non-maleficence* finds greater use than *beneficence* as a principle. At the same time, *sustainability*, *dignity* and *solidarity* are underrepresented compared to more restrictive principles. The AI Act relies on a risk pyramid for its regulation, where there are many more rules for high-risk systems than for systems that form little to no risk [11].

Besides the focus on regulation in the past years, there are more moral dilemmas to be answered about Artificial Intelligence and the broader field of machines and technology. Regulation is only one of the steps towards ethical AI, and the ethical debate should lend itself to discussion on the topics of design and the use of AI as well. Next to this, a system of regulation can also not function as a standalone ethical system without the means to put its contents into action.

This wider area of issues leads to the necessity of more approaches than mere principlism for the field of AI ethics. Whereas principles are an important, even necessary, step towards the instrumental value of ethical debate, they do not offer the ability to form a completely sound and complete system on which we build laws, regulations, and further goals. The reality in this is that principles can not always be consistent with each other at a high level of abstraction, and therefore tensions necessarily arise in the context of use cases [30]. These tensions are exactly what should be the focus of the debate, as they are not easily caught in legal documents or clearly agreed upon moral design values.

Without the history or wide database of use cases that AI ethics still lacks, we thus need to rely on different backgrounds to translate principles into practice and arrive at an agreement on how to deal with such tensions [20]. We see a prominent role for virtue ethics in this. This will help us pave the way to a complete debate that includes all stakeholders in a meaningful way, by forming one of the main pillars of our approach in rounding out the ethical background necessary for such debate.

We build on earlier virtue ethics approaches in the field of philosophy of technology. 2016 marked the proposal of Shannon Vallor's widely acknowledged "technomoral virtues," and we will follow her work in placing virtues within a technological society [27]. Rather than relying on the sparse history of ethical debate surrounding modern technology that is pointed out in recent literature [20], Vallor focuses on the global history of morality relating to the actors that are involved with technology. She specifically focuses on Aristotelian, Confucian and Buddhist traditions, and there are more sources to draw from in the debate of what makes up a morally desirable character. This background allows for an approach based on the *how* rather than on the *what* of moral action concerning technology.

Exactly such an approach is what we believe should be the focus of the field of AI ethics. With the acceptance that the complete field of technological morality cannot be caught in one coherent ethical framework, let alone a system of regulation, there comes unavoidable ethical debate. Within the regulation that is agreed upon, there is still space for consideration in how to act concerning both the design and use of AI. This is unavoidably linked to the context in which the action takes place, and therefore the actors are an important constant. Returning to the older field of biomedical ethics, we should not consider virtue ethics and principlism as competitors. Instead, we take the stance of Beauchamp, who notes that "moral philosophy rooted in the virtues complements a framework of principles" [4, p.193].

We are not the first to defend the position of virtue ethics specifically for the ethical debate surrounding Artificial Intelligence. Working on the extensive groundwork laid out by Vallor with the introduction of technomoral virtues, there have been attempts at introducing virtues for technical actors to do with AI. Noteworthy is the framework as proposed by Raquib et al. that is based on Islamic virtues and specified for use in AI ethics [23]. The approach we follow is based on deriving the virtues specifically from the most prominent set of principles, combined with a historical scoping view of the wider field of virtue ethics. This is the approach taken by Hagendorff [14].

2.3 Considering the whole AI ecosystem

However, it should be noted that these virtues are specifically aimed at so-called AI practitioners, which only includes domain experts. This is not the aim of our framework, as we consider virtues as applying to all involved with Artificial

Intelligence. This goes from lawmakers to developers to users: all stakeholders of Artificial Intelligence. Therefore, it is also important to consider the context in which these stakeholders relate to the system with respect to which they act. This forms what we call the AI ecosystem, which can be very expansive or narrow for cases of wide implementation or specific implementation respectively.

For example, a model used by the national tax authorities to detect fraud has an expansive ecosystem as the developers, government, and complete population already form a first circle of stakeholders through their immediate involvement. On the other hand, a chat bot used by a hospital to aid in answering questions will have a more narrow ecosystem as the only possible users will be the group of people that go to this hospital and can communicate via text. In all cases, the values held by different stakeholders in the ecosystem are dependent on the position of these stakeholders in the ecosystem and the specific context in which the technology should be seen.

This is also where the importance of considering a complete ecosystem comes from. All involved parties hold different values, and these values should be taken into account if we are to hold a constructive ethical debate. This position is for example defended in the applied ethics approach of guidance ethics [28], scientific advisory reports [21][29], and scientific literature [20][30]. These sources form the starting point of the rest of this paper, in which we work towards the inclusive moral dialogue that puts all values held in an ecosystem on the map.

2.4 The Lack of Shared Definitions

The main obstacle on the path towards the inclusive dialogue that we have located in this paper is the lack of shared definitions throughout this ecosystem. The values that different stakeholders hold are not clearly defined enough in terms of the ethical debate that ensues from them. This mostly follows from using terms that are either too specific or too general in their current definition. The former category includes principles as they are being used by domain experts, but are not understood throughout the ecosystem. The latter includes the expression of values in terms of virtues which are not yet discussed enough in the context of Artificial Intelligence.

To locate and further understand the gaps, we took an approach of co-creating both scientific and non-scientific text. For this, we considered viewpoints representing different parts of AI ecosystems as news media, scientific articles and governmental discussion. This consideration was then developed into written texts for further reflection and discussion. The resulting series of essays can be found on `https://www.maartenmintjes.com` [2]. As is discussed in one of these essays, an example of a definition gap can be found in the childcare benefits scandal that took place in the Netherlands.

The childcare benefits scandal included the use of an algorithmic model to classify people according to the perceived risk that they were committing fraud. When it came out that many of those suspected of fraud were not in fact receiving benefits illegally on purpose, the model became heavily criticised and was discontinued. This lends itself very well for a discussion on the tension between accuracy on the one hand and fairness on the other. However, such a discussion requires clarity in the terms that are being discussed. That agreement of definition was not present in this case.

From the discussion within the Dutch governmental sphere, it can be induced that the definitions as used are not quite the same as for the scientific sphere. For example, there has been a long discussion on the prohibition of black box systems, but these are not clearly defined at any point. This can be seen in the report as produced by Amnesty International for example, which has been discussed at length in the States General [3][16]. The report calls for a ban of black box systems in high-risk environments by governments. A black box system is defined as "an algorithmic system where the inputs and outputs can be viewed, but the internal workings are unknown" [3, p. 6].

This contrasts with the way that the discussion has moved on to the different terms of transparency, explainability or predictability as have been present in the domain literature from two years before the report by Amnesty International was published [17]. These terms are a more specific and quantifiable way of speaking of the same features as these that define the black box, but this has not spread through all layers of the ecosystem.

There thus exists a lack of shared definitions throughout the ecosystem. This forms an obstacle on the path towards a moral dialogue between all stakeholders of Artificial Intelligence, which ought to play an important part in ensuring all values are met in the outcome of the ethical debate surrounding these new technologies. Furthermore, this debate should be built on a basis of virtues as well as principles.

Both this inclusive nature and underlying foundation come together in the deliverables of this paper. In the following two sections, we will present both our framework and roadmap that aim at working towards the resolution of the gap of shared definitions. With this, we move closer to an ecosystem-wide moral dialogue on Artificial Intelligence.

3 The Framework for Structuring Inclusive Dialogue

We have laid out our stance on the objective of frameworks in the previous section: to enable structure and process rather than to be a definitive solution to questions of morality. With this comes the importance of seeing frameworks in the context of that part of the ethical debate for which they are designed. This also determines the relevant AI ecosystem, which the framework influences and is influenced by. For example, the framework as published by the HLEG should be seen in the context of leading to legal regulation in the form of the AI Act [15][11] to which its ecosystem corresponds.

This stance on the position that frameworks take in the ethical discussion surrounding AI is visualised in figure 1. The figure also shows our twofold focus of this paper on a framework and a roadmap of further connecting the ethical debate and the concerned ecosystem.

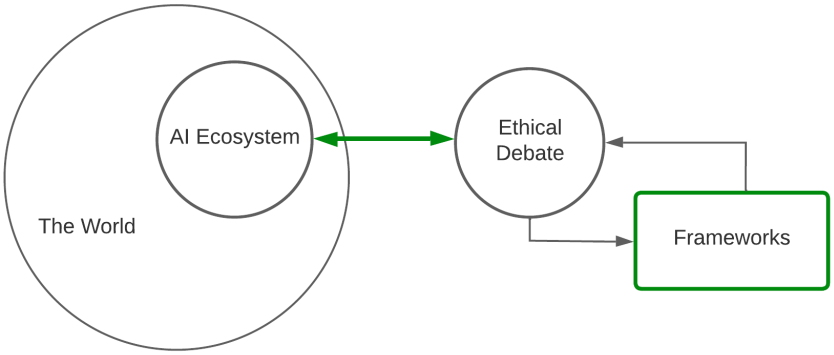

Fig. 1. Ethical discussion structure with our focus in green

3.1 The Context of the Framework

The aim of our framework is to create a structure that can be used to come to shared definitions which allows for dialogue on the moral values that are important for all those that are involved with Artificial Intelligence. This should *not* be seen as an attempt at creating a complete and consistent framework for regulation or guidelines for ethical AI. Rather, this framework is and should be inherently evolving.

Through the connection of principles, as they have been present in the field of AI ethics in the past years, and virtues, as they are historically present, we point at the gaps in understanding and perceived importance of topics in the ethical debate. This ultimately leads to the construction of more shared definitions on which a structured dialogue can be built that includes domain experts as well as the general public. This is a transition for which this framework lends structure.

We address the gaps in shared definitions by adding an extra layer to our framework besides those formed by virtues and principles. This additional layer consists of a collection of 'topics.' These are more approachable than principles and more specific than virtues, with the aim of including stakeholders throughout the AI ecosystem in a structured debate. To our knowledge, this is the first framework aimed at including the whole ecosystem while still including work on principles and virtues from within the scientific field.

Earlier frameworks do not specifically address this gap. Virtue-based frameworks do not offer a way towards the inclusivity of dialogue [14][23][22]. We build on these frameworks and use the virtues as proposed by Hagendorff as we agree with these and aim towards a more continuous debate. The relevant framework by Buhmann and Frieseler does focus on communication and deliberation and is taken as a source of inspiration for the introduced topics, but builds less on existing work as the main focus is not to connect the ecosystem but rather to introduce a new path towards responsible AI [8].

Furthermore, the used principles are collected from scoping reviews of varying areas of the research as has been performed in AI ethics in the last years [17][12][13]. The groupings of the principles into virtues has been performed through a study of earlier works that have made use of such an approach [15][12][14].

In most cases, we have found that our groupings do closely represent the groupings as are presented in the earlier literature. That being said, differences in the groupings can be found. For example, *access to technology* is not necessarily grouped with *beneficence* as is the case in Fjeld et al.[12] but rather with *inclusiveness in design*. We see this as an interpretative choice, even one that forms part of the gap as we have found in the ecosystem as we are part of this as well.

For the further categorisation of these principles into topics, we started from the set of principles per virtue. These were then split up into subgroups that represented similar moral questions in the public discussion that we have studied and written about for our co-creation approach. For this, the aim was to find a middle ground between the characters of the used virtues and principles. The leading question to answer with the topics, therefore, was “What is the core of principles X, Y and Z that contributes to the specific virtue A?” besides the consideration of whether this would fit within an approachable base of dialogue for the whole ecosystem.

An example to clarify this is the topic of *honest systems*, which covers the core of the principles *transparency*, *predicability* and *explainability*. Furthermore, the design and application of honest systems directly contribute to the virtue of *honesty*. Lastly, this topic fits within a wider dialogue as the topic is already talked about throughout different parts of the ecosystem like the earlier discussed report on the Dutch childcare benefits scandal.

3.2 The Framework and its Further Evolution

The resulting framework is presented in figure 2, with a supporting list of the principles as they are grouped under topic and virtue in table 1. This is a first step towards a more complete model which is designed to be evolving with the current state of the ethical debate. With this also come some points of consideration for the intended further evolution of the framework.

These can mostly be seen to be connected to the newly introduced layer of topics. This is due to the novelty of this layer, in combination with the inherently changing nature of the topics themselves. As our choice of topics is based on the current state of discussion throughout the ecosystem surrounding Artificial Intelligence, the suitable topics for a given time change with the evolution of the dialogue. This is not only an accepted limitation of the introduction of this framework, it is a desired one as the furthering of this wider dialogue is our main objective.

Furthermore, there is room for further exploration within the levels of abstraction that this framework contains. With this, we mean introducing additional layers between the abstract virtues and the more applied principles. Introducing a layer of topics can be seen as a first step towards an ethical structure that allows for consideration of subjects of discussion in various levels of concreteness. For example, the ethical debate might benefit from considering the legality of a specific algorithmic system through questions ranging from abstract to more concrete as the following respective examples:

- "Would a responsible government allow for this?"
- "What legal framework would allow for such a system?"
- "Does this algorithm comply with the formulated principles of safety and reliability?"
- "What specific data protection rules must be in place to allow for this system?"

We strongly recommend looking into these subjects of further consideration. The evolution of the used topics and the possible introduction of more layers of abstraction could form the next step in the ethical debate cycle as presented in figure 1.

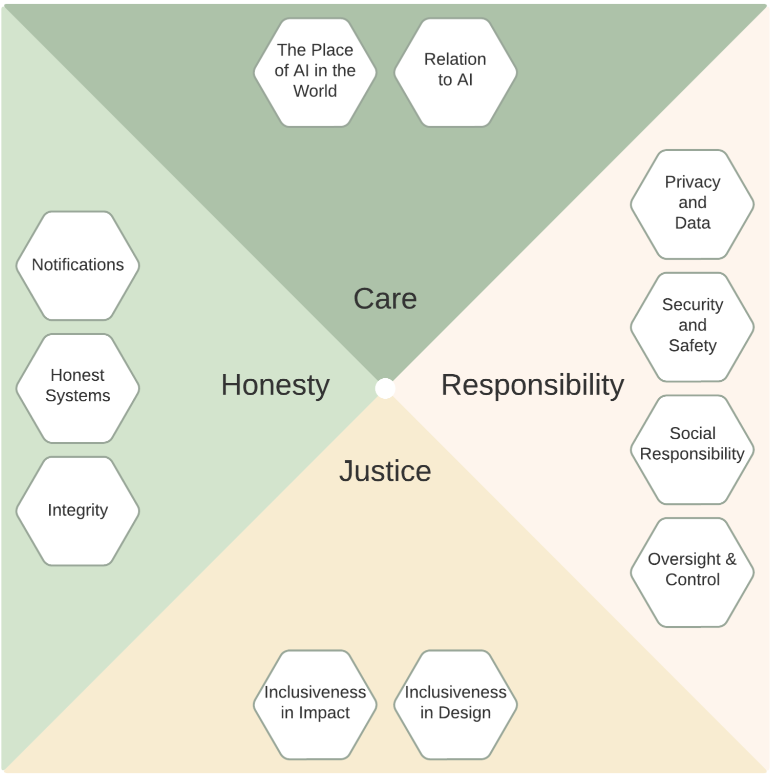

Fig. 2. The Framework for Structuring Inclusive Dialogue

Virtues	Topics	Principles
Care	The Place of AI in the World	Hidden Costs; Non-Maleficence; Sustainability; Leveraged to Benefit Society; Dual-use Problem, Military, AI Arms Race; Environmental Responsibility; Beneficence; Human Values and Human Flourishing; Well-being; Consideration of Long Term Effects
	Relation to AI	Ability to Restrict Processing; Ability to Appeal; Ability to Opt Out of Automated Decision; Consent; Freedom; Human Autonomy; Dignity
Justice	Inclusiveness in Impact	Fairness; Solidarity, Inclusion, Social Cohesion; Justice; Inclusiveness in Impact; Access to Technology
	Inclusiveness in Design	Equality; Non-discrimination; Cultural Differences in the Ethical Design of AI Systems; Inclusiveness in Design; Prevention of Bias; Representative and High Quality Data; Diversity in the Field of AI
Honesty	Notifications	Notification when AI Makes a Decision about an Individual; Notification when Interacting with an AI
	Integrity	Open Procurement (for Government); Scientific Integrity; Open Source Data and Algorithms; Verifiability and Replicability
	Honest Systems	Transparency; Predictability; Explainability
Responsibility	Security and Safety	Security; Security by Design; Remedy for Automated Decision; Certifications for AI Products; Cybersecurity; Safety and Reliability; Human Control of Technology
	Privacy and Data	Privacy by Design; Recommendation for Data Protection Laws; Control over Use of Data; Privacy Protection; Privacy
	Oversight & Control	Recommendation for New Regulations; Evaluation and Auditing Requirement; Regular Reporting Requirement; Human Oversight, Control, Auditing; Human Review of Automated Decision; Accountability; Liability and Legal Responsibility; Legislative Framework, Legal Status of AI Systems; Creation of a Monitoring Body; Science-policy Link
	Societal Responsibility	Right to Information; Responsible Design; Impact Assessment; Responsible / Intensified Research Funding; Future of Employment / Worker Rights; Public Awareness, Education about AI and Its Risks; Right to Rectification; Right to Erasure; Multistakeholder Collaboration

Table 1. List of used Virtues, Topics and Principles

4 The Roadmap Towards Inclusive Dialogue

We have now introduced a framework to structure the ethical debate in the transition of AI ethics. It is our view that this must go hand in hand with an inclusive moral dialogue. For this, we must lay a common ground. Besides relying on the framework for the discussion topics, we also offer a way to put the construction of shared definitions into practice. The combination of these elements is visualised at the end of this section in figure 3 after the contents are discussed.

4.1 Stages of the Roadmap

Establishing a common ground for dialogue throughout the AI ecosystem can be separated into two parts. The first part consists of ensuring all parties have access to a sufficient amount of background knowledge. This includes a broad idea of the technical workings of Artificial Intelligence systems, but also of legal regulations that are in place for the development of such systems. An example of this can be found in the Dutch national AI Course, which is publicly available and aims to provide a complete background picture for the field of Artificial Intelligence [1].

The main objective of the second part is the construction of a base of shared definitions throughout the ecosystem. This is the main focus of this paper, as it is where we have located the main obstacle on the path to an ecosystem-wide moral dialogue. We also see this as the focus of the expert dialogue as the HLEG initiated it in coming to the guidelines that led to the AI Act [15]. Furthermore, an example of this can be found in the approach of guidance ethics [28]. This approach centralises discussion, for which a dialogue between stakeholders is first held on the definitions used. Applications of this have mostly been in the healthcare sector but vary over different projects. Therefore, the concerned ecosystem is inclusive, but on a small scale.

The last stage of the roadmap is the first step of inclusive dialogue. Discussing the tensions between virtues is imperative for coming to an agreed-upon direction of AI ethics for different use cases. In the remainder of this section, we propose the method of Socratic questioning as the best fit for coming to shared definitions in building a common ground for constructive debate.

4.2 The Socratic Method

To come to shared definitions between all stakeholders of systems of Artificial Intelligence, we see great benefit in the use of the method of Socratic questioning. We will first give a brief overview of the method. After this, we will defend the use of this method concerning the characteristics of a suitable method.

The Socratic method is named after Socrates and his dialogues[24]. In this method of questioning, five stages can be discerned[7]:

1. Posing a **question**, generally one with no clear answer;

2. Forming a first **hypothesis** as an answer to the question;
3. Questioning and attempting to refute the hypothesis;
4. Accepting or rejecting the hypothesis after the previous step;
5. Acting on the findings of the questioning.

The method has historically been applied to metaphysical or ethical questions such as "What is courage?" or "Is justice better than injustice?"[24]. Socrates himself held that the answers to such questions must be:

- explanatory of the subject, i.e. what it is that makes a courageous person courageous;
- exclusive to the subject, i.e. what it is that belongs only to the courageous;
- belonging to the whole subject, i.e. that which is the case for all courageous people[5].

A defining feature of the Socratic method is that there is no preconception of an existing final answer, there is no truth that cannot be denied [9]. Whereas there might be answers that meet the three requirements as stated, these may and should always still be questioned. This is beneficial for our objective as the moral dialogue on AI not only contains questions that are inherently unanswerable, values held throughout the ecosystem also differ and it can be helpful to consider this without the idea of a definite solution in mind.

The Socratic method breaks down concepts and introduces them anew, allowing the present members of the dialogue to come to shared definitions and structures in their knowledge organization[18]. This breaking down is especially essential for the topics at hand because there are many differences in existing knowledge organisation and experience with the subject matter for moral questions concerning Artificial Intelligence.

Socratic questioning also helps in the promotion of critical thinking. A case study of this can be found in including a Socratic seminar in an ethical education program developed for middle school children as an initiative of MIT[19]. As part of a broader educational method, the Socratic seminar followed the method of Socratic questioning to come to a discussion in the classroom. On a wider review of this implementation, this was seen to have a positive effect on the students' ability to think critically on the studied case of the ethical design of YouTube[10]. It should be noted here that only the benefits for the students were taken into account, not for the teachers in the situation.

4.3 A Proof of Concept for the Roadmap

In the approach that we have taken for this paper, we co-created both this scientific paper and a series of less scientifically minded essays. Besides locating the obstacle of the lack of shared definitions, this has also resulted in texts that touch on many of the discussed topics. Therefore, the series of essays can be seen to partly fit the structure as we have introduced in our roadmap. This gives a concrete proof of concept for the more conceptual roadmap. The essays are listed below and can be found at https://www.maartenmintjes.com [2].

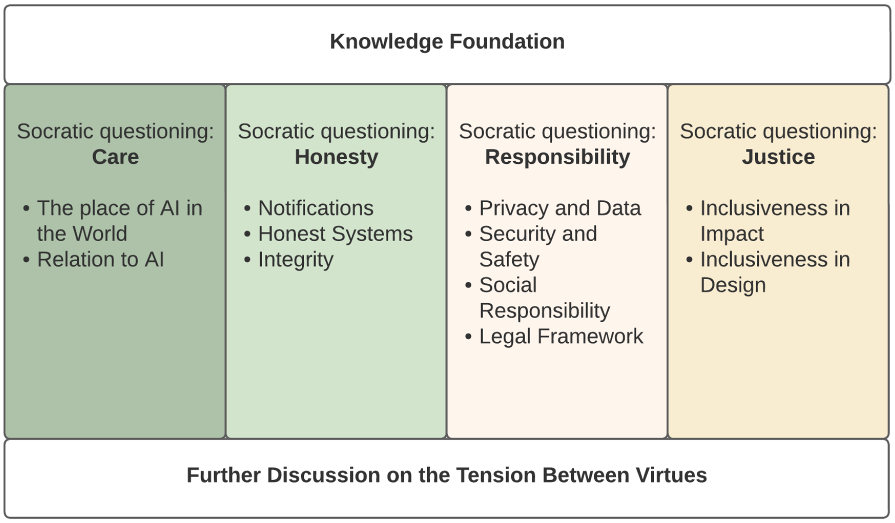
Knowledge Foundation
Socratic questioning: Care
• The place of AI in the World
• Relation to AI
Socratic questioning: Honesty
• Notifications
• Honest Systems
• Integrity
Socratic questioning: Responsibility
• Privacy and Data
• Security and Safety
• Social Responsibility
• Legal Framework
Socratic questioning: Justice
• Inclusiveness in Impact
• Inclusiveness in Design
Further Discussion on the Tension Between Virtues

Fig. 3. The Roadmap Towards Inclusive Dialogue

1. ”On Machine Learning” gives an introduction to the field of Machine Learning and Artificial Intelligence.
2. ”But why, Mr. Robot?” gives an introduction to eXplainable AI and interpretable machine learning.
3. ”Putting robots in boxes” considers the benefits and drawbacks of black box systems and transparency and explainability.
4. ”Die Frage” is the first part of a consideration of the relation that we have to algorithmic systems and technology in general.
5. ”Let’s talk about God” is the second part of this consideration, in which a more philosophical background is given for our relation to technology.
6. ”Concrete Case Study: The Toeslagenaffaire” is a case study on the use of self-learning systems in our society of today, which concerns accuracy, bias, and understanding of what we will call AI.
7. ”To bias or not to bias” considers the benefits and drawbacks of adding more information into a self-learning system.

The essays then follow the roadmap structur as presented in figure 4.

Four of the seven texts are written in a Socratic style: essays 3, 4, 5 and 7 in the given enumeration. The Socratic method has historically been used in dialogues, rather than purely written form [24]. It is preferable to enter discussion directly, something that would be done on a smaller scale. However, these essays should be seen mostly as a general introduction to further Socratic dialogue which is suitable for a large audience at once. They introduce topics, and question them in a Socratic way by asking questions styled in the way of ”What is X?”

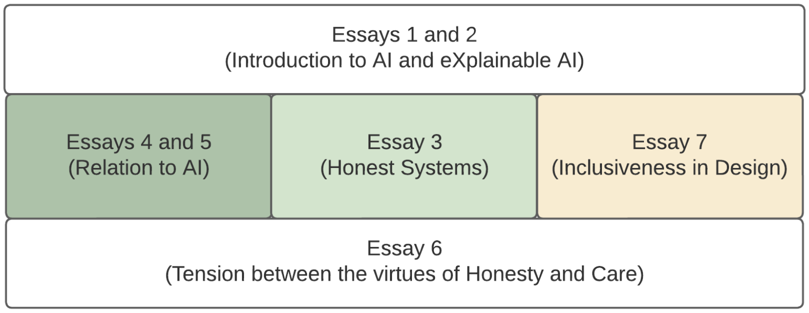
Essays 1 and 2
(Introduction to AI and eXplainable AI)
Essays 4 and 5
(Relation to AI)
Essay 3
(Honest Systems)
Essay 7
(Inclusiveness in Design)
Essay 6
(Tension between the virtues of Honesty and Care)

Fig. 4. Proof of Concept for the Roadmap

or "What ought Y to be?" and attempting to answer and refute these answers in an iterative manner.

An example of this can be found in the essay on bias in systems. First, the idea of bias and unbiased systems is introduced together with the hypothesis that we ought to aim for completely unbiased systems. This is then refuted by stating that an unbiased system would be a system which is of no use as bias is also what makes a system able to distinguish between different instances. This relates to the first steps of the Socratic method of questioning, hypothesising and attempting to refute. Accepting the refutation makes for a void in the useful definition of bias and specifically unbiased systems. Now there is a place for a new, shared definition of bias for which the possible first steps are taken in the essay as well. This completes the last steps of the Socratic method.

This first proof of concept is not yet a complete application of the theoretical structure of initiating dialogue. For this, the introductory and final part should be more extensive. Besides this, there is a clear gap in the consideration of the virtue of responsibility compared to the theoretical structure and conceptual framework. More weight has gone to the virtue of care, even though this encompasses less of the principles as are important in the field of AI ethics as it is today. These gaps offer interesting insights of their own in the challenges that come with structuring the dialogue that AI ethics will require. We will discuss these in the next section, and contextualise our results in a wider scope.

5 Discussion

The main thesis of this paper is that a lack of shared definitions throughout the AI ecosystem forms the main obstacle on the path towards an inclusive moral dialogue on Artificial Intelligence, which is a desirable part of the debate in the near future of AI ethics. To resolve this obstacle, we present a novel framework and a roadmap towards this inclusive dialogue. In this section, we contextualise our findings and discuss the implications of our work. For this, we begin by reflecting on the approach of co-creation that has resulted in the identification of this obstacle and the production of a proof of concept for the roadmap. Afterward, we give a more high-level view of the limitations and possibilities for future work regarding our findings.

5.1 Reflecting on the Approach of Co-Creation

The written essays both laid the groundwork for the approach of this paper and formed a more concrete view of the more abstract result of the produced roadmap. This came to be in an organic manner, as has been described in section 2.4. Therefore this should also be seen as a further reasoning behind the choice for the specific method of Socratic questioning and a topic-based framework, next to being a proof of concept for the usage of such a style.

The differences between the composed framework and method and the organically produced proof of concept should also be noted. The Socratic style of writing did lend itself better for work on the construction of shared definitions rather than on introducing factual knowledge or discussing tensions between different virtues or principles. This has resulted in the point of view that the focus with the Socratic method should be on that stage of initiating dialogue.

Furthermore, there was a gap in the representation of virtues in the proof of concept. This is especially noticeable with regard to the virtue of responsibility. This virtue makes up the group with the most principles, but was not represented in our writing. While this should of course by no means be seen as proof or evidence of a correlation, it is worth considering whether the principles as they now do indeed point to the most pressing concerns to be discussed in the near future of AI ethics. Virtues not only offer a way of coming to more resolution in discussion as was pointed out before, but new frameworks may also be able to offer guidance in the direction of work of AI ethics by pointing out important areas of research and discussion that are yet underrepresented.

Writing about current topics in the general discourse, as well as diving into the domain literature at the same time, allowed us to consider a perspective representative of a larger part of the ecosystem. This is a point of view that, although seemingly straightforward, is not much represented in either of the respective spheres due to too little overlap between discussion groups. Exactly this reflection is important for the meta-discussion at hand, as we find ourselves in a time when we still have the luxury of asking questions about the dialogue to be had rather than being stuck to a certain path forward.

Findings in the writing of the essays and the reactions of peers and wider reader groups have contributed to this feeling and stance within this paper. Many topics were not questioned in any way in general discourse or were not given any attention in the scientific debate. At the same time, reactions from non-expert readers were positive and sparked further discussion containing viewpoints that were not represented yet by experts in the field. This pointed out a base of knowledge and opinion that was currently used very little, as well as pointing out the lack of two-way communication flow between the respective spheres of experts and the public. With this paper, we have placed these viewpoints within the scientific debate, which provided insights for a structure of method to rely upon in the following work.

We hope to have shown by using our research method that the interplay between scientific and non-scientific spheres does not need to be limited to results. As technology and society become more intertwined, so must science and application.

A dialogue of research can support the dialogue in AI ethics.

5.2 In the Context of AI ethics

With the framework as presented in this paper we aim to further along the transition of AI ethics towards providing structure for the answer to the question of where we ought to go with Artificial Intelligence within the agreed upon restrictions. With this, we are part of a larger base of research focused on the exploration and implementation of novel frameworks to guide the near future of AI ethics. It is therefore also an accepted and even desired consequence that we do not present a final complete framework for the field to agree upon. Rather, the framework is designed to help in working towards a dialogue that will result in clearer moral and ethical standards for Artificial Intelligence rather than proposing them directly.

Furthermore, it is an essential characteristic of ethical debate that there is not one agreed upon method of agreeing upon a standard or direction. However, we see dialogue as an inherent part of an accepted and trusted direction for Artificial Intelligence in the world. With this in mind, we aim to defend the Socratic method as we believe that this adds to the shared definitions that we see as required for a constructive and structured dialogue. Even so, this approach only forms part of the common ground that needs to be formed as a first step of initiating dialogue between different stakeholders. There will undoubtedly be varying approaches to this dialogue. With this approach we most importantly aim to make a strong case for the inclusion of the general public as well as for the open nature of such a general debate, which are characteristics which come forward with Socratic questioning.

5.3 Looking Forward

It is recognised in a much broader context that there is a lot of work needed to prepare society for the upcoming time. The Netherlands forms a good example for this. This year, the Dutch Minister of Education, Culture and Science has made available 10 million euros for a new national center of science communication. Over ten times that money is assigned extra this year to new research studying the relation between technology and society in the Netherlands. More funds and efforts are needed, and made available for the transition to an even more technological society.

Specifically, in AI ethics, this means already preparing for what is to come. The field has been in a more reactive state and, even though there are proponents of such a strategy, not all ethical questions can be answered that way.

Frameworks like the one we have presented need to be challenged and improved upon. Especially considering the applied nature into which we are seeing the field of AI ethics transition, it will be of great importance to try out approaches and converge towards frameworks and methods that work in an iterative version. However, there is no time to waste with this. Expanding upon the given proof of concept that is given as part of our applied method, and even expanding on the theoretical framework, gives good handles to work with to make concrete the efforts put into the field. Artificial Intelligence is developing and so are the moral questions that come with it. We must be ready to provide structure for the debate that is beginning to ensue.

6 Conclusion

Moral philosophy is concerned with the study of what is right and what is wrong. These seemingly inseparable topics nevertheless form the basis for the two subjects that make up applied ethics: agreeing on right and wrong and promoting and restricting these respectively. The sub field of applied ethics that is concerned with Artificial Intelligence is now in a transition from a stage of creating boundaries towards a stage of providing direction within these.

We believe this transition should rely on moral dialogue including all stakeholders of Artificial Intelligence. In this paper, we have identified a lack of shared definitions throughout the ecosystem as the main obstacle towards this dialogue. With a novel approachable framework and roadmap, we work towards the resolution of this obstacle and the construction of a structured dialogue to support the AI ethics debate.

Currently, principles play a key role in the work that is done in AI ethics. They have been the source of structure and consensus in creating leading guidelines and regulations. However, an approach based purely on principles also has its limitations. We believe virtue ethics will become an important addition to the ethical theory that makes up the foundation of AI ethics. The framework we propose combines virtues and principles and adds a layer of topics to help bridge the gap between different parts of the stakeholder ecosystem. This way, it builds on earlier research and works towards a more inclusive debate on the virtues of all those involved with Artificial Intelligence.

Furthermore, we propose to continue work in the direction of the Socratic method as a way of coming to needed shared definitions. The approach that we have taken to locate and resolve the obstacles of coming to an inclusive dialogue forms a proof of concept for placing the Socratic method on the path towards this ecosystem-wide moral dialogue.

Both our research and experience working on the topic point to the need for an approach to questioning in AI ethics. Now is the time to establish an agreed-upon direction for new developments and implementations of Artificial Intelligence. This involves a critical look at the standing values in the ethical debate. As society moves towards a state in which intelligent systems are ubiquitous, dialogue between stakeholders must also be. The question to ask ourselves should no longer be how we can move forward, but what direction we agree that forward really is.

References

1. De Nationale AI Cursus. https://www.ai-cursus.nl/, accessed: 23/01/2023
2. The project. https://www.maartenmintjes.com/, accessed: 23/01/2023
3. Amnesty International: Xenophobic machines - discrimination through unregulated use of algorithms in the Dutch childcare benefits scandal (October 2021)
4. Beauchamp, T.L.: Principlism and its alleged competitors. Kennedy Institute of Ethics Journal **5**(3), 181–198 (1995)
5. Benson, H.H.: Socratic wisdom: the model of knowledge in Plato's early dialogues. Oxford University Press on Demand (2000)
6. Bietti, E.: From ethics washing to ethics bashing: A view on tech ethics from within moral philosophy. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. p. 210–219. FAT* '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3351095.3372860
7. Boghossian, P.: Socratic pedagogy: perplexity, humiliation, shame and a broken egg. Educational Philosophy and Theory **44**(7), 710–720 (2012)
8. Buhmann, A., Fieseler, C.: Towards a deliberative framework for responsible innovation in artificial intelligence. Technology in Society **64**, 101475 (2021). https://doi.org/10.1016/j.techsoc.2020.101475
9. Delić, H., Bećirović, S.: Socratic method as an approach to teaching. European Researcher. Series A (10), 511–517 (2016)
10. DiPaola, D., Payne, B.H., Breazeal, C.: Decoding design agendas: an ethical design activity for middle school students. In: Proceedings of the interaction design and children conference. pp. 1–10 (2020)
11. European Commission: Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain Union legislative acts (2021)
12. Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., Srikumar, M.: Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. SSRN Electronic Journal (2020). https://doi.org/10.2139/ssrn.3518482
13. Hagendorff, T.: The ethics of AI ethics: An evaluation of guidelines. Minds and Machines **30**(1), 99–120 (2020)
14. Hagendorff, T.: A virtue-based framework to support putting AI ethics into practice. Philosophy & Technology **35**(3), 1–24 (2022)
15. High-Level Expert Group on Artificial Intelligence: Ethics guidelines for trustworthy AI. Tech. rep. (April 2019)
16. van Huffelen, A.C.: Kamerbrief met reactie op Amnesty rapport Xenofobe Machines. Letter of Government (Jul 2022), https://www.tweedekamer.nl/kamerstukken/brieven_regering/detail?id=2013Z19029&did=2013D39431
17. Jobin, A., Ienca, M., Vayena, E.: The global landscape of AI ethics guidelines. Nature Machine Intelligence **1**(9), 389–399 (2019)
18. Lam, F.: The socratic method as an approach to learning and its benefits (2011)
19. Massachusetts Institute of Technology: RAISE. https://raise.mit.edu/, accessed: 23/03/2023
20. Mittelstadt, B.: Principles alone cannot guarantee ethical AI. Nature Machine Intelligence **1**(11), 501–507 (2019)

21. Nederlandse AI Coalitie: Mensgerichte artificiële intelligentie: Een oproep voor zinvolle en verantwoorde toepassingen (2020)
22. Neubert, M.J., Montañez, G.D.: Virtue as a framework for the design and use of artificial intelligence. Business Horizons **63**(2), 195–204 (2020). `https://doi.org/10.1016/j.bushor.2019.11.001`
23. Raquib, A., Channa, B., Zubair, T., Qadir, J.: Islamic virtue-based ethics for artificial intelligence. Discover Artificial Intelligence **2**(1), 11 (2022)
24. Robinson, R.: Plato's earlier dialectic. Cornell University Press (1941)
25. Selbst, A.D., Boyd, D., Friedler, S.A., Venkatasubramanian, S., Vertesi, J.: Fairness and abstraction in sociotechnical systems. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. p. 59–68. FAT* '19, Association for Computing Machinery, New York, NY, USA (2019). `https://doi.org/10.1145/3287560.3287598`
26. The European Commission: Building trust in human-centric artificial intelligence. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions (April 2019)
27. Vallor, S.: Technology and the virtues: A philosophical guide to a future worth wanting. Oxford University Press (2016)
28. Verbeek, P.P., Tijink, D.: Guidance ethics approach: An ethical dialogue about technology with perspective on actions. ECP (2020)
29. Wetenschappelijke Raad voor het Regeringsbeleid: Opgave AI. de nieuwe systeemtechnologie. Tech. Rep. 105, WRR, Den Haag (2021)
30. Whittlestone, J., Nyrup, R., Alexandrova, A., Cave, S.: The role and limits of principles in AI ethics: towards a focus on tensions. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. pp. 195–200 (2019)
31. Wiener, N.: Cybernetics: or Control and Communication in the Animal and the Machine. MIT Press (1948)