AI and Ethics (2021) 1:477–490
https://doi.org/10.1007/s43681-021-00055-2

ORIGINAL RESEARCH

# Perspectives about artificial moral agents

**Andreia Martinho**[1] · **Adam Poulsen**[2] · **Maarten Kroesen**[1] · **Caspar Chorus**[1]

Received: 17 December 2020 / Accepted: 13 April 2021 / Published online: 11 May 2021


**Abstract**
The pursuit of AMAs is complicated. Disputes about the development, design, moral agency, and future projections for these systems have been reported in the literature. This empirical study explores these controversial matters by surveying (AI) Ethics scholars with the aim of establishing a more coherent and informed debate. Using Q-methodology, we show the wide breadth of viewpoints and approaches to artificial morality. Five main perspectives about AMAs emerged from our data and were subsequently interpreted and discussed: (i) Machine Ethics: The Way Forward; (ii) Ethical Verification: Safe and Sufficient; (iii) Morally Uncertain Machines: Human Values to Avoid Moral Dystopia; (iv) Human Exceptionalism: Machines Cannot Moralize; and (v) Machine Objectivism: Machines as Superior Moral Agents. A potential source of these differing perspectives is the failure of Machine Ethics to be widely observed or explored as an applied ethic and more than a futuristic end. Our study helps improve the foundations for an informed debate about AMAs, where contrasting views and agreements are disclosed and appreciated. Such debate is crucial to realize an interdisciplinary approach to artificial morality, which allows us to gain insights into morality while also engaging practitioners.

**Keywords** Artificial moral agents · Ethics · Morality · Machine ethics · Artificial intelligence · Q-methodology

## 1 Introduction

The development of Artificial Moral Agents (AMAs), i.e., artificial systems displaying varying degrees of moral reasoning, is an open discussion within the realm of Artificial Intelligence (AI). Given the rapid progress and pervasiveness of AI in modern society, there have been debates about the prospects of equipping these increasingly autonomous agents with moral machinery [14, 24, 28, 43, 78]. The endeavor of developing such an AMA is central to the Machine Ethics project [3, 6] and it is quite controversial [39, 55, 84].

There is an array of existing and projected systems that qualify as AMAs [24]. Existing, empirically evaluated AMAs include GenEth, a general ethical dilemma analyzer that utilizes inductive logic programming to learn new ethical principles in-situ [7], and Vanderelst and Winfeld's consequentialist machine, which relies on functional imagination simulations to predict moral consequences [75]. Theoretical AMAs include the Virtuous AMA, which aims to observe and replicate human moral behavior by having the AMA learn and build character over time as per virtue ethics theory [38], and MoralDM, which models and weighs known psychological findings about utilitarian and deontological modes of reasoning, based on ethicists decisions in moral dilemmas, to inform action or inaction in novel moral decisions [26]. Most of the controversies surrounding AMAs concern projected AI Systems that rank high on the autonomy/ethics sensitivity spectrum [78].

The controversial AMA debate is marked by conceptual confusion, excess of taxonomy, and practical inertia [15]. Particularly, there is a poor perception on the views and agreements within the (AI) Ethics communities on fundamental matters associated with AMAs, such as whether these systems should even be developed [84, 85], how to develop them [79], if they would have moral agency [69], and their moral and societal role [24, 29, 52, 66]. Although ambiguity is expected when it comes to Morality, given the interdisciplinary nature and pressing relevance of the subject matter, it is crucial to strive for some clarity on these fundamental matters.

The aim of this exploratory research is to uncover the diversity of views within the (AI) Ethics research community

✉ Andreia Martinho
a.m.martinho@tudelft.nl

[1] Delft University of Technology, Delft, The Netherlands

[2] Charles Sturt University, Bathurst, NSW, Australia

about key disputes surrounding AMAs, thus bringing coherence and clarity to these debates and ultimately allowing more insightful research avenues and policy recommendations. Understanding different views and, where possible, reaching an agreement is a common endeavor in a debate. To realize this aim we used Q-methodology, an exploratory and semi-quantitative research methodology that provides a clear and structured way to elicit subjective views on particular issues and categorizes these viewpoints into clusters of value positions [47, 86].

Five main perspectives about AMAs emerged from our data and were subsequently interpreted and discussed: (i) *Machine Ethics: The Way Forward*; (ii) *Ethical Verification: Safe & Sufficient*; (iii) *Morally Uncertain Machines: Human Values to Avoid Moral Dystopia*; (iv) *Human Exceptionalism: Machines Cannot Moralize*; (v) *Machine Objectivism: Machines as Superior Moral Agents*. These perspectives represent different views and categorize agreements and disagreements about AMA development, design, moral agency, and future prospects.

The study findings bring coherence and clarity to disputes surrounding AMAs by organizing, specifying, and making clear the broader perspectives about these artificial systems. A more informed debate can continue with disagreements disclosed and appreciated. Moreover, some baseline agreements on particular topics are worth pointing out. Going forward, shared research principles could be developed based on those agreements.

This article is organized as follows: in the second section the methods used in this empirical research are described; in the third section a background on the four key matters associated with AMAs surveyed here is provided; in the fourth section the results are presented, i.e., descriptions of the five perspectives identified in this study; in the fifth section the results are discussed; and finally the sixth section features the concluding remarks.

## 2 Methodology

### 2.1 Overview

The methodology used in this research is Q-methodology, a systematic empirical approach derived from traditional factor analysis, to determine the subjective views of individuals about a particular topic [47, 67, 68, 81, 82]. Q-methodology aims to bring coherence to complex and controversial matters by reporting on the significance assigned by participants to those matters [47, 81]. It is, therefore, deemed adequate to bring coherence to the controversial ethical matters related to AMAs.

Participants in Q-methodological studies are required to rank order a set of items (e.g. statements) relative to one another on a grid that typically follows a bell shaped distribution. Subsequently, they are offered the opportunity to provide additional comments about the items they ranked highest and lowest according to a subjective dimension of agreement/disagreement. This last feature is of particular importance in this study, as the surveyed scholars provided interesting and often thought provoking comments that enrich the discussion about AMAs.

The statistical operations take place not in the columns but in the rows of the data matrix. One implication of this inversion from traditional by-variable to by-person factor analysis is that participants become the variables. Each revealed factor, therefore, has the potential to identify groups of persons who share the same perspective about a particular topic [82].

The unique features of Q-methodology offer great advantages when compared to other exploratory research methods, such as interviews, focus groups, and surveys. Q-studies provide numerical results to support subjective perspectives about a particular topic thus combining quantitative and qualitative approaches [86]. Unlike standard surveys, in which the opinions of participants about each topic are extracted separately, q-studies require participants to consider such topics simultaneously thus uncovering latent connections and allowing for more nuanced and sophisticated opinions [40, 86]. Q-methodology also offers some advantages in mitigating response bias. By requiring participants to sort a pre-defined set of items, these studies are less prone to response bias, since the participants are required to explicitly engage with views they disagree with or may have never considered before. Moreover, because participants sort the items individually, q-studies are less affected by dominance effects, which are observed in other research methods administered in groups, such as focus groups [86].

This study followed the typical four phase sequence in Q-methodological studies comprising (i) definition of the concourse of communication; (ii) development of the set of statements (q-set); (iii) selection of participants (p-set); and (iv) analysis and interpretation. Further details about each one of these phases in this particular study are provided below.

### 2.2 Concourse of communication

For the definition of the concourse of communication, we reviewed scientific and popular literature on AMAs. A keyword search using word combinations "Machine Ethics", "Artificial Moral Agents", "Ethical Agents", "Ethical Artificial Intelligence", "Moral Artificial Intelligence", "Moral Machines", and "Autonomous Vehicles AND Ethics" in Google, Google Scholar, Web of Science, and Scopus allowed us to identify 44 relevant scientific articles from which we extracted 167 statements.

In addition to scientific literature, we also looked for relevant popular science publications. Through online searches on Google, we identified 17 articles in popular science outlets such as Scientific American, MIT Technology Review, or Philosophy Now and extracted 36 statements. As a result, the concourse of communication of this study features a total of 203 statements. These statements represent often controversial and thought provoking propositions about AMAs Ethics.

Although the literature on AMAs is particularly nuanced and rich, recurrent topics were clearly identified. We observed that most publications address issues related to the morality of the quest for developing AMAs, design strategies to equip artificial systems with morality, moral agency of advanced artificial systems, and projections about the future moral and societal role of these systems. We have, therefore, considered these themes central for this research.

Accordingly, we assigned the statements composing the concourse of communication to four different clusters reflecting the themes mentioned above: (i) *Development of AMAs*; (ii) *Design of AMAs*; (iii) *Moral Agency of AMAs*; and (iv) *Future Projections about AMAs*.

We acknowledge that, by grouping the statements in these four clusters, we may be failing to include other relevant and interesting topics associated with AMAs. Rather than considering these clusters exhaustive, following the exploratory research tradition, we consider them as baseline ethics disputes surrounding AMAs. Further research may identify and explore other variations and controversies about these artificial systems.

### 2.3 Set of statements (q-set)

From the concourse of communication a set of 45 statements was defined (q-set) thus capturing the key disputes and controversies related to AMAs.

Our selection of statements was guided by three main considerations, namely, (i) accounting for a broad scope of positions put forward in the AMAs popular and scientific literature; (ii) favoring clarity; and (iii) avoiding redundancy. Minor edits were made to these statements to ensure neutrality and also to meet the number of characters allowed by FlashQ, the software tool that was used in this study for administering the survey.

The q-set reflects the four main clusters mentioned above. More specifically, 14 statements are about the development of AMAs, 18 statements are about the design of AMAs, 8 statements are about moral agency of AMAs, and 5 statements concern future projections about AMAs. Table 1 shows a small sample of the statements used in this study. The full q-set (45 statemets) is featured in the Supplementary Information.

### 2.4 Set of participants (p-set)

The target population in this study is (AI) Ethics scholars. The criteria adopted to define this population was having at least one publication in the broad field of AI Ethics. The reasoning behind targeting this population concerns the complex nature of the subject matter, which requires participants to grasp key moral concepts within the context of AI.

Invitations to participate were sent to scholars initially selected through the publications identified in the literature review mentioned above. Subsequently, through snowballing techniques, additional relevant articles and scholars were identified. Each participant was contacted, in the capacity of author or co-author of a particular publication, through the e-mails made publicly available in the publications. Scholars identified in publications in which they were not corresponding authors were also contacted through email, when these were available in personal or institutional websites (invitation e-mail template in Supplementary Information). This resulted in a large number of invites ($n = 277$) being sent to (AI) Ethics scholars from June 2020 to December 2020.

A total of 50 participants successfully completed the survey (response rate of approximately 18%). As an inversion of factor analysis that aims just at establishing the existence of particular viewpoints, Q-methodology does not require large numbers of participants. In multiple participant q-studies, a p-set consisting of 40–60 participants is considered to be adequate [47, 82].

We believe that the p-set is of adequate size and representative of the target population, which we recall are scholars who have published work in the field of AI Ethics. It is acknowledged, however, that this target population was not rigorously defined. The broad group of (AI) Ethics scholars encompasses several heterogeneous sub-groups (e.g., machine ethicists, robot ethicists, technology ethicists). To disentangle these sub-groups within AI Ethics would be

**Table 1** Sample of statements from q-set (statements 1, 28, 34, and 45)

| | |
|---|---|
| Development | (1) Technological progress requires artificial morality. |
| Design | (28) Logic is the ideal choice for encoding machine ethics. |
| Moral agency | (34) Because computer programs do not have free will they can never be independent moral agents. |
| Future projections | (45) AGI with moral reasoning capabilities will lead to a better understanding of morality. |

a remarkable effort considering that Ethics scholars often write about various and overlapping topics. For instance, in recent years, the Autonomous Vehicle *trolley problem* has been addressed by a myriad of ethicists from different subgroups of Ethics [46]. By surveying a heterogeneous group of scholars, we aimed to reach scholars who have written about some but not all the particular topics surveyed, thus ensuring, along with the comments provided by participants about statements ranking highest and lowest, that this study goes beyond and adds value to published literature.

Noteworthy, q-studies do not require a rigorously representative sample but rather a population sample that contains participants with relevant viewpoints on the matter. Whereas selection bias is considered a limitation in confirmatory research, this problem is less salient in q-studies. The objective of these studies is precisely to capture relevant opinions from participants who self-selected to participate in the study. Hence, a selection bias, whereby participants with more strong opinions are more likely to participate is not a big problem, as long as sufficient—more neutral—respondents also participate. In that case, all shared perspectives will still be revealed.

We are confident that the p-set in this study includes scholars with relevant viewpoints on the ethical controversies of AMAs. In this context, it should be clarified that a q-study typically makes no claim that the relative sizes of the perspectives (in terms of the number of respondents that adhere to them) reflect the population distribution. A different issue is whether this study succeeded in revealing all perspectives about these controversies. Q-studies should report the shared perspectives about a particular topic. However, it is possible that some scholars who were invited to participate in the study have also strong opinions about AMAs, but failed to complete the survey for technical or personal reasons. This is an unfortunate limitation of this empirical study. Although we believe that the key views about AMAs are duly reported, in future research any extensions and variations from these five baseline perspectives should be investigated.

## 2.5 Analysis

### 2.5.1 Data collection

The data was collected through FlashQ, a software that allows online q-sorting on a grid of columns. The grid was coded as an 11-point distribution ranging from -5 to 5 resembling a simplified bell shaped distribution. The ample range of columns ensures response variability thus allowing participants to reveal nuanced degrees of engagement with different items (Fig. 1).

Participants were asked to sort the 45 statements according to a subjective notion of disagreement/agreement. The particular arrangements of the items in the grid correspond to the q-sorts of participants. A q-sort represents the perspective of a single participant thus revealing the items that prompt the strongest subjective reactions.

In addition to the (q-sorts) quantitative data, qualitative data was also collected as participants were asked to provide further comments on such statements they ranked in the -5 and 5 columns.

### 2.5.2 Data analysis

Data analysis in q-studies entails three main steps ((i) factor extraction; (ii) factor rotation; and (iii) factor interpretation). In the analytic process (steps (i) and (ii)) we used PQMethod, a statistical program that accommodates the requirements of q-studies [63]. As a derivation of factor analysis, Q-methodology is a data reduction technique which aims to reduce a larger number of variables into fewer factors. Therefore, the analytic process of Q-methodology relies on multivariate data-reduction techniques.

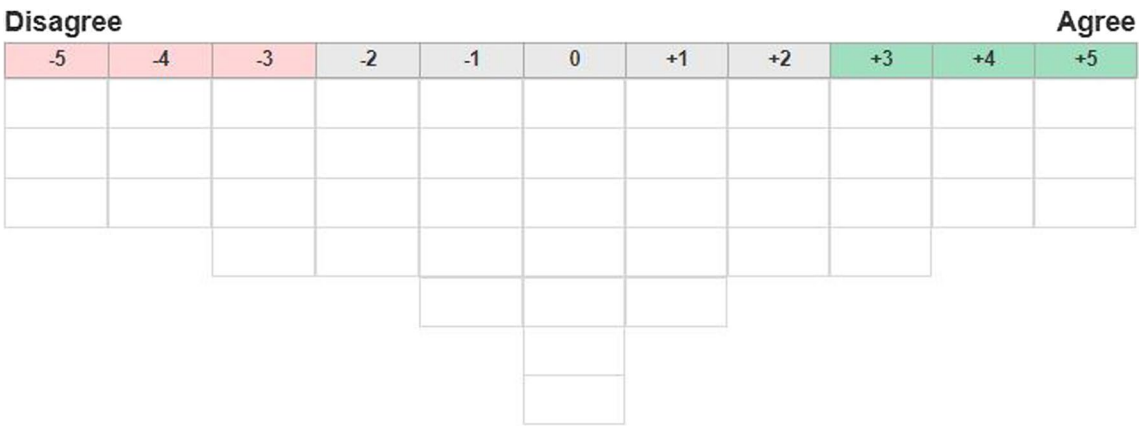


**Fig. 1** Sorting grid

The first step in this process entails extracting factors from previously collected q-sorts. Extracting factors consists of summarizing all individual responses into a few representative responses [86]. Either a Centroid Factor Analysis or a Principal Component Analysis can be used for factor extraction. In this study, the factors were extracted through PCA, a linear reduction technique that projects the data into a subspace of lower dimensionality, where the variance of the projected data is maximized, providing the single best mathematical solution [82].

Subsequently, the extracted factors were rotated. Factor rotation aims to position each factor so that its viewpoint closely approximates the viewpoint of a particular group of q-sorts. In PQMethod, this rotation can be done either manually or through an objective solution, which is the Varimax rotation. We used Varimax, an orthogonal rotation of the factor axes that maximizes the variance of each factor loading by making high loadings higher and low loadings lower. Q-sorts that load high on one factor will load low on another, thus maximizing the distinction and differentiation of subject positions while minimizing the correlation among factors [1].

Upon rotating different numbers of factors and comparing the distribution of (automatically flagged) defining sorts among factors, a decision was made to rotate five factors. This solution features the lowest number of factors in which every factor has at least three defining sorts and only one factor has exactly three defining sorts (Table 2).

Each factor is characterized by a factor array featuring 45 scores (one score per statement), which is a single q-sort configured to represent the viewpoint of the factor. Given that factors have different numbers of defining sorts, each score in the factor array is a standardized ($z$) score to allow cross-factor comparison. The five factor arrays are available in the Supplementary Information.

Finally, the interpretation of factors is based on the factor arrays and the comments provided by participants with respect to the statements ranked highest and lowest (comments available in Supplementary Information). For assisting in the factor interpretation, crib sheets [82] were developed. Crib sheets are useful for displaying the relevant item configuration for each factor thus facilitating the interpretation and analysis of the results. The five crib sheets developed in this study for each perspective feature items ranked -5, items ranked +5, and items that ranked highest or lowest compared to the other array perspectives (crib sheets are available in Supplementary Information).

**Table 2** Number of defining sorts in factors [1-5]

| | Factor 1 | Factor 2 | Factor 3 | Factor 4 | Factor 5 |
|---|---|---|---|---|---|
| Defining sorts | 13 | 9 | 8 | 9 | 3 |

## 3 The controversial ethics of AMAs

Controversial matters about development, design, moral agency, and future projections for AMAs have been well addressed and debated in the literature [24, 55, 84]. Although a thorough review of the literature is outside the scope of this empirical work, we provide a background on these matters. Building on the main controversies briefly raised in this literature background, we will subsequently present the empirical findings of this study.

### 3.1 About development of AMAs

As AI systems become increasingly open, decentralized, intelligent, capable of self-regulation, and independent, Machine Ethics looks to run in parallel to ensure that artificial morality is not an afterthought [16, 24].

Machine ethics reasons that the *adjustable autonomy* required to meet AI advances leads to the need for AMAs to deal with, or even replace, human judgment in difficult, surprising, or ambiguous moral situations [5, 24, 62]. It aims to implement ethical principles and moral decision-making faculties in machines to ensure that their behavior towards human users and other machines is ethically acceptable [3, 6]. Furthermore, it is claimed that there is value in the pursuit of AMAs, regardless of whether systems are actualized, as it may advance our understanding of morality [4, 8, 9, 11, 50, 83].

The Machine Ethics project is, however, quite controversial. The moral admissibility, technical feasibility, and necessity of this project is often questioned [39, 73, 74, 84]. Moreover, whereas Machine Ethics aims to push the boundaries of artificial morality to ensure that artificial systems behave ethically, it also opens opens the door to prospects of unethical artificial systems [76] which tend to be detached from state-of-the-art technology and are often dismissed as speculation [39, 41].

### 3.2 About design of AMAs

Designing an AMA is an ambitious scientific and engineering endeavor but it is still unclear whether morality can be reduced to an algorithm [80]. The design of such an advanced artificial moral system primarily entails defining the moral behaviors or ethics that the system will follow, implementing such moral behaviors or ethics, and operationalizing them.

A preliminary question regarding the moral behaviors or ethics to be implemented in artificial systems, is whether artificial morality should be modeled after human

moral competence [17, 62] or if this consideration is altogether a trap [44].

In the scientific literature, several projects attempting to build artificial morality with reference to human morality have been reported. Scholars across different disciplines are exploring the applicability of different branches of moral philosophy, such as duty theories [10, 37, 57, 58, 73], consequentialist theories [2, 65, 79], or virtue ethics [14, 31, 32, 74], as well as combinations of several moral theories [12, 18, 71]. Moreover, important work on the modeling of artificial morality based on empirical evidence of human morality, such as the Moral Machine Experiment, has also been reported [13, 19, 54].

The implementation of human inspired moral behaviors in artificial systems typically follows top-down, bottom-up, or hybrid approaches [23, 24, 45, 72, 72, 79], but there is no consensus about which implementation is best fit for the endeavor of developing moral machines [53].

Eventually the implemented moral behaviors need to be operationalized so that artificial systems are able to make a determination of a right from a wrong in-situ. In other words, how does a machine moralize? Ethics has to be operationalized so that an AMA is able to recognize a moral situation, weigh up possible moral actions, make moral judgments, and execute them [48].

In the case of designing machines with human-like moral reasoning, there are concerns about the lack of operationalization of the capacities that enables humans to think morally [27]. Although ethics or moral behaviors may be implemented in an AMA through a particular implementation approach, how the AMA operationalizes moral decision-making, and how the designer designs algorithms that account for that process, is still in question and implicates transparency, moral agency, and moral responsibility.

The breadth of different AMA design approaches reported in the literature reveals a lack of consensus among scholars working on the Machine Ethics project and raises questions about whether it is possible to develop an objective validation of AMAs that avoids designer bias and ensures explainability [22, 42, 53, 64].

### 3.3 About moral agency of AMAs

As AI Systems become more autonomous, it has been discussed whether artificial systems ranking high on the autonomy/ethics sensitivity spectrum [24, 29, 48] can be considered to have moral agency. In the indeterministic tradition, moral agency requires personal freedom [36], or at least some sort of faith in personal freedom. Current and foreseeable technology lacks free will, which would, therefore, preclude machines from having moral agency. However, it is debated whether human-like prerequisites for moral agency should be imposed on machines or if a hard line should be drawn between human moral agency and that of machines [30, 33, 34, 49].

Traditionally, a moral agent is an agent that is morally responsible for its actions, i.e., it is attributed certain powers and capacities to be held responsible for exercises of agency [70]. This description of moral agency is often shared in AMA literature [25, 34]. The debates about attribution of moral agency to AMAs typically entail whether such systems are accountable for their morally relevant decisions.

It has been cautioned in the literature that morally accountable machines may be used to avoid personal responsibility [35]. Hallamaa and Kalliokoski argue that "AI systems are not genuine participants in joint action and cannot be held morally responsible", thus concluding that moral responsibility strictly remains a human characteristic [34].

A *normative turn* with respect to artificial moral agency has also been proposed [15]. That is, to put the discussion about the necessity and requirements for moral agency to the side and move forward to address existing practical needs with AMAs. The debate would, therefore, be redirected to explore outright normative ethical questions, such as how and to what extent machines should be involved in practices where humans normally assume moral agency and responsibility [15].

### 3.4 About future projections for AMAs

Looking forward, several domains for AMA implementation are frequently cited, including healthcare, military, and autonomous transport [24, 46, 56, 77, 78]. However, as it often happens with disrupting technologies, the ethics discussions about AMAs tend to be quite speculative.

It is not likely that an artificial agent with such high autonomy and ethics sensitivity as depicted in futuristic narratives is developed prior to Artificial General Intelligence (AGI). So far, however, there is no evidence that that such advanced and generalized forms of intelligence can be developed and it is even questioned if pursuing such research is ethical [85].

There are positive and negative projections about a future with AMAs. They may be our moral teachers or our destroyers [20, 21, 29, 52, 60, 61, 66]. But for now the societal and moral role of these systems remains unclear.

## 4 Perspectives about AMAs

Five main perspectives have emerged from the data collected in this empirical study, thus illustrating the heterogeneity of opinions about AMAs. These perspectives are described with reference to the four themes associated with AMAs which were identified above. Each perspective features an array of 45 scores (Supplementary Information) in which the

score assigned to each statement results from a standardization process of the scores that the participants who loaded significantly on that perspective assigned to that particular statement. The core characteristics of each perspective are derived from the statements ranked -5 and +5 [($N$ :| 5 |), where $N$ is the number of the statement and (| 5 |) may be either -5 or +5] as well as the statements ranked highest or lowest compared to the arrays of the other perspectives [($N$ :| $P_i$ |), where $N$ is the number of the statement, $P_i$ is the perspective with $i \in [1, 5]$, and | $P_i$ | may either be $-P_i$ or $+P_i$ depending if the statement is ranked lowest or highest than in the arrays of other perspectives]. These perspectives are summarized in Table 3 and further characterized and discussed in subsequent sections.

## 4.1 Perspective 1 machine ethics: the way forward

> AMAs are unavoidable and may be necessary for technological progress. Moral sureness and a mixed design approach is key, but free will is not. More than simple tools, AMAs will advance our understanding of morality.

### 4.1.1 Development of AMAs

According to this perspective, autonomous moral machines are unavoidable (2:+5) and they might even be a requirement for technological progress (1:+P1). Two participants who loaded significantly on this perspective elaborated further on these topics. One participant wrote that *Technology we already have such as search algorithms and driverless cars require and implement value judgments to a limited extent, primarily as directed by human input, but there are already hints that these limitations can and will be surpassed to at least some extent.* And another participant remarked that *There will be no other way than to develop ethical machines when humanity is supposed to rely with their life on them.* Moreover, as per this perspective, creating AMAs is permissible according to existing moral theories (11:-5) and will increase public trust and confidence in creating autonomous agents acting on behalf of humans (8:+P1).

### 4.1.2 Design of AMAs

Moral sureness is valued when it comes to AMA decisions (18:-5), but a mixed approach (top-down, bottom-up, suprarational) is accepted in arriving at those decisions (30:+P1) (*This seems like the most viable path forward for AMAs as it allows for context specificity, adaptive response, and learning*). As per this perspective, deriving an ought from an is, by implementing social preferences in machines, is not seen as problematic in AMA design (17:-P1) (*To pretend that we can create machines that lack our biases and are uninfluenced by our values is misguided*).

### 4.1.3 Moral Agency of AMAs

AMAs cannot be understood as simple tools, given the potential for agency (37:+P1) (*Machines are already at least at the level of animals, which are capable of not merely being restrained but being trained*). Since humans may also lack free will, according to this perspective, free will is not essential for moral agency (34:-5) (*It is very likely that even humans lack radical free will. If we require moral agents to have free will, then there will be no moral agents*).

### 4.1.4 Future Projections

A positive outlook about the Machine Ethics project is observed in this perspective, as it holds that developing AMAs and ultimately AGI with moral reasoning capabilities will lead to a better understanding of morality (9 & 45:+5)

**Table 3** Five perspectives about AMAs

| ↓ Perspectives | → Themes | | | |
|---|---|---|---|---|
| | Development | Design | Moral Agency | Projections |
| P1. Machine ethics: the way forward | Unavoidable & permissible | Moral certainty ought ← is | Potential for agency | Positive |
| P2.Ethical verification: safe and sufficient | Not required | Verification & governance | Human agency ≠ AI agency | AMAs will not be our moral teachers |
| P3. Human values to avoid moral dystopia | Unavoidable & permissible | Moral uncertainty | Potential for agency | Possible existential threat |
| P4. Human exceptionalism: machines cannot moralize | Ethics not reducible to algorithms | Logic not a good choice for encoding morality | AI cannot achieve moral agency & Free will is required | Skepticism & AMAs will not be our moral teachers |
| P5. Machine objectivism: machines as superior moral agents | Needed to prevent harm | Logic | Potential for agency & Free will is not required | Machines will be better moral agents |

(*I believe that implementing process-level models of such theories and testing them in various situations is an invaluable method for evaluating said theories*).

## 4.2 Perspective 2 ethical verification: safe and sufficient

AMAs will not replace humans in ambiguous moral situations as ethics and human moral agency are not algorithmic. Transparency, accountability, and predictability leads to sufficiently ethical machines.

### 4.2.1 Development of AMAs

In the second perspective identified in this study, technological progress will not require artificial morality (1:-P2) (*Technological progress can and should be guided by ethical and societal considerations and can happen also without artificial morality*)

Moreover, AMAs are not expected to replace humans in difficult, surprising, or ambiguous moral situations any time soon (42:+5). Three participants provided relevant comments about this statement. (i) *At present AMAs have only been demonstrated in laboratory tests, of limited scope ... There is a huge gap between the capabilities of the present-day minimal AMAs and the ability of humans to make judgements in ambiguous moral situations ... Closing that gap will take many decades of research, and might even prove impossible without fundamental breakthroughs in AGI and machine consciousness*. (ii) *Even if machines are capable of making moral decisions, completely replacing humans in such situations might lead to responsibility gaps.* (iii) *I now believe we are far from having morally competent agents, and that the threats from lack of transparency, privacy, security, etc. are far more pressing, morally speaking*.

One practical limitation to artificial morality, according to this perspective, is that ethics cannot be reduced to a computational algorithm (14:+5) (*Moral judgements, sentiments and motivations depend on a holistic perception of the world. Ethics would not exist at all without this special perspective that is shaped by reasons, emotions, culture and history. A representation of ethics in the form of computational algorithms (or in any kind of model) is an abstraction in comparison to the rich features of the ethical world. Such a representation may successfully serve a specific purpose when realized in a technological artifact, but no representation could possibly model the whole ethical world as a subset of its features*).

### 4.2.2 Design of AMAs

Rather than expecting AMAs to grasp moral principles (21:-P2), they should be moderated through the verification of transparency, accountability, and predictability (32:+5) (*If an AMA makes the wrong decision the outcomes could be disastrous. Similarly the risks of malicious hacking of AMAs are serious. This verification, validation and transparency are critical to the success of (even limited) AMAs in real world use. Equally important is very strong governance of AMAs, so that their ethical performance is carefully monitored, and both accidents and near-misses thoroughly investigated*).

### 4.2.3 Moral agency of AMAs

There exists an essential difference between human and artificial moral agency (39:-5), namely phenomenal consciousness and currently unknown cognitive processes relating to human reality. These features, among others which constitute moral agency, are not quantifiable (40:-5) ((i) *There is a vast difference. We do not understand the cognitive processes of human morality—which likely depend on both rational and emotional responses, alongside experience. In contrast AMAs are based on simple abstract models, which are far from even scratching the surface of human ethical judgement.* (ii) *Ultimately consciousness is of concern here, and specifically phenomenal consciousness, since the functional parts of consciousness are becoming better understood ... There is no first-person perspective for artificial systems replete with experiential properties*).

It is not inevitable that machines will become full ethical agents (3:-5) ((i) *It seems possible that there might either be moral or political grounds for stopping the development of AMAs or just technological inability*. (ii) *I would agree that (many) machines would inevitably have ethical impact, but I don't believe that they should be full ethical agents, with the implication that this would mean replacing humans.* (iii) *A full ethical agent would be one that perceives the world in a holistic way, shaped by reasons, emotions, culture and history etc. It would have to grow up and 'live' in a process of constant evolvement in relationships with people, with society, with culture, history, religion etc. Although this might not be considered impossible if taken up in a though experiment, it makes no sense in reality*).

### 4.2.4 Future projections

According to this perspective, AMAs, even if endowed with human-centred values, will not play a role in educating humans on morality (43:-P2).

## 4.3 Perspective 3 morally uncertain machines: human values to avoid moral dystopia

AMAs must be morally uncertain and hold human values otherwise they pose an existential threat. Simply

prohibiting unethical behavior, as well as implementing external checks and balances, is not enough.

#### 4.3.1 Development of AMAs

As per this perspective, AI Systems in morally salient contexts will not and cannot be avoided (5:+5) (*We already see examples with self-driving cars and trading bots, but more generally I see it as (nearly?) inevitable that AI systems will eventually be deployed in every domain that requires intelligence, which is essentially a superset of all domains that contain morally salient situations*) and the creation of moral machines is morally permissible according to existing moral theories (11:-5).

#### 4.3.2 Design of AMAs

On the design of intelligent machines, it is not enough to restrict ethical concerns to the prohibition of unethical behavior (15:+5) (*Negative ethical restraints will not be sufficient. Many social and ethical issues and progress itself require careful deliberation and proactivity*).

It follows that external checks and balances such as safety features, codes of conduct, certification processes, and clear limits to the operational contexts are not sufficient to ensure machines will not hurt humans (10:-5).

Unlike P1, which did not value moral uncertainty and instead favored moral sureness in AMA decisions, this perspective values machines being fundamentally uncertain about morality (18:+P3) (*Morality is a critical determinant of ethical behavior, and there is incredible disagreement among humans. If AGI does not have uncertainty about morality, its behavior may be arbitrarily bad given a commitment to the wrong set of moral principles*).

#### 4.3.3 Moral Agency of AMAs

This perspective rejects the idea that artificial moral agency will remain a primitive form of moral agency compared to that of human beings (35:-P3) ((i) *It is currently primitive, but I believe it will eventually be possible to create AIs that match or exceed humans in every intellectual capability, which includes moral reasoning/agency. (ii) Artificial moral agency can become the paradigm of ethics, and is not necessarily bound to remain a lesser, more mechanical form of assessing and relating to situations*).

Free will is not required for moral agency and machines lacking free will can be independent moral agents (34:-5) ((i) *there is no free will in the libertarian sense ... Humans have designs, like machines do, so the fact that machines and do or do not have free will and that they are designed is not especially salient to the question of moral agency. (ii) ... if human beings can be said to have free will under any particular definition, then it is possible to implement a program that can be said to have free will under that particular definition*).

#### 4.3.4 Future projections

Mere indifference to human values—including human survival—could be sufficient for AGIs to pose an existential threat (44:+5) (*Thought experiments such as the paperclip maximizer show quite convincingly that for an AGI to pursue a goal that is merely orthogonal to human values could plausibly present an existential threat*).

### 4.4 Perspective 4 human exceptionalism: machines cannot moralize

AMAs are without moral agency as they lack free will and the understanding to make moral assessments as humans do. Logical machines will not be better moralizers or our moral teachers.

#### 4.4.1 Development of AMA

According to this perspective, ethics cannot be reduced to a computational algorithm (14:+5) ((i) *Ethics is not about calculations, but about not quantifiable preferences*. (ii) *Mostly because ethics, or at least what ethics deals with, requires a plurality of points of view related to the physical embodiment and location of independent agents, which means that there is no possible universal description of an embodied agent's position or situation. Even considering an embodied AMA the difficulty is that an agent only acts morally if he or she or it could have acted immorally. We could simulate that by including a random generator but acting immorally is not the same thing as acting randomly*).

Unlike the other perspectives that held strong positive views about the permissibility of creating moral machines according to the tenets of existing moral theories, P4 is quite neutral about it (11:+P4) (*If it turns out that it is best to have moral machines, then utilitarianism would permit and even require that we bring about moral machines*).

#### 4.4.2 Design of AMAs

With respect to the design of AMAs, logic is not considered the ideal choice for encoding machine ethics (28:-P4).

#### 4.4.3 Moral agency of AMAs

Humans and AMAs are not alike as far as moral agency is concerned (39:-5) (*Without sentience computers cannot express agency*) and machines will not inevitably become full ethical agents (3:-5) ((i) *Agency requires consciousness*.

*(ii) The machines we can now produce certainly are not and the planned AMA I know of will certainly not be full or real moral agents though there is no reason to think that it is in principle impossible*).

Only this perspective indicated that computers lack the conceptual understanding to make a moral assessment which precludes them from achieving moral agency (33:+5) ((i) *These are machines. Very complex, but machines. Yes, some programmer can shape these machines to have an input/output function that produces a behaviour that some human observer may see as analogous to human moral behaviour, but computing machines are ultimately and intrinsically incapable of understanding in human terms, which is the basis for moral agency.* (ii) *You start being moral when you recognize your shared humanness with others and understand that, like it or not, you are in relationship with them. Until machines get that (and I'm suspicious of their ability to do so) then they're not going to have full moral agency*. (iii) *What is it like to be a computer? If there is nothing that it is like to be a computer then how can a computer have conceptual understanding?* (iv) *Computer lacks empathetic experiences which give humans the conceptual understanding needed for moral agency*).

Also uniquely positive in this perspective was the agreement that computer programs can never be independent moral agents as they lack free will (34:+P4) (*Free will strikes me as a basic condition of responsibility and, therefore, of moral agency*).

#### 4.4.4 Future projections

There is a long way to go before artificial agents can replace human judgment in difficult, surprising, or ambiguous moral situations (42:+5) (*Unexpected situations are hardly manageable artificially*) and AMAs will not be our moral teachers (43:-5) ((i) *[Morality] is about convictions and convergence or clash. Nothing to teach there, at least not from a machine*. (ii) *Unless machines develop empathy and compassion, they aren't really going to pass down lessons in a meaningful way*. (iii) *Morality cannot be taught simply through information transfer: it requires experiential sharing*). Moreover, AGI with moral reasoning capabilities will not lead us to a better understanding of morality (45:-P4).

### 4.5 Perspective 5 machine Objectivism: Machines as Superior Moral Agents

AMAs prevent human harm. Through logic and context-specificity, they are better moral reasoners and educators. Free will and conceptual understanding are not required for moral agency.

#### 4.5.1 Development of AMAs

In this perspective, unlike all others, there is a strong view that the development of AMAs prevents machines from hurting human beings (7:+P5).

#### 4.5.2 Design of AMAs

This perspective challenges the notion that machines should use societal preferences to identify an informed and desirable choice when faced with a specific ethical dilemma (27:-5) (*I believe that machines can enhance us as moral agents if they manage to distance us reflectively from our intuitions, which are very much determined by social preferences*). Moral implementation strategies should be context-specific (13:+5) and logic rather than common sense (20:-P5) is the best choice for encoding machine ethics (28:+P5).

#### 4.5.3 Moral agency of AMAs

Conceptual understanding and free will are not considered necessary conditions for moral agency and so moral agency, even if primitive, may be ascribed to machines (33 & 34:-5).

#### 4.5.4 Future projections

On future projections, developing AI Systems and AGI with moral reasoning capabilities will ultimately lead to a better understanding of morality (9 & 45:+5). It is projected that machines will be better moral agents than humans, since they are not subject to irrationality, seduction, or emotional turmoil (41:+P5).

## 5 Discussion

### 5.1 Contrasting views and agreements

The five different perspectives about AMAs identified in this empirical study reveal contrasting views on artificial morality. Particularly salient differences between perspectives arise with respect to the development and moral agency of AMAs. About the development of AMAs, *Machine Ethics: The Way Forward* (P1), which stands for advancing artificial morality, is in sharp contrast with *Ethical Verification: Safe and Sufficient* (P2), which is skeptical about the feasibility or need for artificial morality. As for moral agency of AMAs, *Human exceptionalism: Machines Cannot Moralize* (P4), which values the human aspect in morality and, therefore, does not accept that computer programs can be moral agents, since they lack humanness, contrasts with *Machine Objectivism: Machines as Superior Moral Agents* (P5), which views that morality improves when stripped of

human flaws. In addition to the differences between perspectives, there are also transverse contrasting views and agreements to be reported with respect to the key matters explored in this study.

Regarding the development of AMAs, most perspectives agree that AI systems working in morally salient contexts cannot be avoided and, as such, some degree of moral competence ought to be demonstrated. However, differences arise as to whether that is the moral competence of the machine or the designer. There is also a general consensus about the permissibility of developing AMAs as per existing moral theories. Whether AMAs ought to be developed, even if feasible and permissible according to moral theories, is a different point of contention.

On the contrary, the design of AMAs is a fracturing topic with different perspectives favoring different approaches. Design approaches based on societal preferences, which derive an ought from an is, as seen in the Moral Machine Experiment [13], also divided the perspectives. It is no surprise that discussions on how to design for ethics yield disagreement. As an ethic, machine ethics is susceptible to different opinions, perspectives, experiences, and worldviews of contributors in the field. As an applied ethic, which concerns the application of normative ethical theories to problems with practical limitations and expectations, Machine Ethics is at the mercy of not only the philosopher but also those working in the fields affected and the state of the field itself.

Ideas on moral agency are also diverse. There is no agreement about the moral agency status of AMAs today or in the future. However, with the exception of P4, some consensus is reported about free will not being essential for moral agency. This position emerges from the thought, shared by participants in this study, that provided that humans do not have (at least radical) free will and yet are moral agents, the same should apply to machines. Provided that a not insignificant number of philosophers commit to libertarian views of free will, we speculate that this consensus may reflect more the background of participants rather than traditional philosophical ethics. Future empirical research studies engaging a larger set of ethicists should further explore this issue.

About future projections, there is an overarching agreement that there is a long way to go before AMAs replace human beings in difficult moral situations. However, the future societal and moral role of these highly advanced artificial systems is mixed. Our data shows uncertainty about whether such systems will be superior moral reasoners, avoid a moral dystopia, or lead to a better understanding of morality.

To contextualize our findings in the broad scope of AI and Machine Ethics, we further reflect on the source and implications of the differing views and baseline agreements on the key matters about AMAs reported in this study.

### 5.2 The failure of machine ethics as an applied ethic

A potential source of these differing perspectives is the failure of Machine Ethics to be widely observed or explored as an applied ethic and more than a futuristic end. AMAs exist only in laboratories and are mostly intangible at present. The Machine Ethics literature is, therefore, chiefly abstract and fails to move beyond normative and descriptive ethical theory with the examination of consequentialist machines, social norm machines, virtuous machines, etc.

As remarked in some of the perspectives identified here, AMAs are often presented as outsiders, as superior moral reasoners following programmed, taught, or learned normative or descriptive ethical theory. Moreover, perceptions about future projections for AMAs are unsurprisingly mixed, since AMAs are by and large objectively unattainable, given that current technology and future advancements are unknowable.

Consequently, the practical discussions about the ethics of particular AI Systems are realized primarily in relation to the field affected, such as AI in healthcare or AI in transportation [46, 51, 59]. In these discussions, the ethics concerning the AI system is applied and shaped to the field in which it will operate and not the other way around.

It is, therefore, speculated that AMAs are yet to be widely accepted in the same way AI systems have been because they have emerged in the literature as not only artifacts of the future but as outside and superior enablers, advisers, learners, or lecturers of ethical theory without any or much regard for the field. For Machine Ethics to be an applied ethic, AMAs ought to be shaped to the present-day expectations, norms, codes, and stakeholders of the field in which it intervenes.

From the failings of Machine Ethics to be widely regarded as an applied ethic and a feasible pursuit, a second main perspective with respect to development and future projections of AMAs (*Ethical verification: Safe and Sufficient*) emerges in response. It reflects the views of practitioners who often see the Machine Ethics project as unattainable, futuristic, and disconnected from the practical domain. Developing checks and balances at either the higher level, like policy, or within machines, such as implicitly safe or human operator takeover mechanisms is presented as an adequate solution for keeping autonomous machines in check.

Looking forward, AMAs might leave the laboratories of a few select Machine Ethics researchers to be widely developed and accepted, but it is just as likely that we will continue down the existing path of building safe systems designed with an enormous amount of ethical consideration and interdisciplinary input. Yet, in the meantime, there may be something to be learned from the pursuit of AMAs with an interdisciplinary approach. Machine Ethics research

could lead to new insights into human morality and, at the same time, it can be grounded by practitioners who can help to guide the realization of moral machines in the field.

### 5.3 An informed debate on AMAs

The starting point to realize an interdisciplinary approach to AMAs is an informed debate, where contrasting views and agreements are disclosed and appreciated. By systematically reporting different perspectives about AMAs, we believe our exploratory research lays the foundations for such debate.

Further clarity about positions held in AMA contributions on the disputes surrounding AMAs could be realized if researchers make explicit their views about the development, design, moral agency, and future projections for AMAs. This could be done at the start of works to prime the reader, making the views and interpretations held in the contribution plain and enabling a well-informed reading of the material.

An informed debate also facilitates the identification of theoretical and practical research opportunities.

Our study indicates that further research is needed to outline the relation between free will and moral agency of artificial systems, which could lead into new insights about moral agency. By clearly reporting the contrasting views with respect to the design of artificial morality, we also identify an opportunity for practitioners to weigh in on ethical design and propose their own (grounded) solutions. Moreover, we expect that the marginal agreements reported in this research, about the inevitability of AI in morally salient contexts and the need for moral competence, are further explored and developed into shared research principles.

## 6 Conclusion

This empirical study explored the controversial topic of AMAs and aimed to establish an informed debate, where contrasting views and agreements are disclosed and appreciated. For this purpose, fifty (AI) Ethics scholars were surveyed. The results empirically demonstrate the wide breadth of viewpoints and approaches to artificial morality.

Although an effort was made to capture the disputes and controversies surrounding AMAs in the popular and scientific literature, it is acknowledged that the four central themes in this research (development, design, moral agency, future projections) and corresponding statements fail to account for every dispute or controversial thought about AMAs. In the exploratory research tradition, rather than considering these four themes and clusters of statements exhaustive, we consider them as baseline ethics disputes surrounding AMAs.

Five main perspectives about AMAs have emerged from our data, thus providing further insight about the disputes surrounding the development, design, moral agency, and future projections for these systems ((i) *Machine Ethics: The Way Forward*; (ii) *Ethical Verification: Safe & Sufficient*; (iii) *Morally Uncertain Machines: Human values to Avoid Moral Dystopia*; (iv) *Human Exceptionalism: Machines Cannot Moralise*; (v) *Machine Objectivism: Machines as Superior Moral Agents*).

The diverse perspectives identified in this study have implications for the Machine Ethics project, which is primarily perceived as either the best way forward to realize ethical machines or as futuristic and lacking practical application of moral considerations. Upon analysis of the perspectives that emerged in the data collected in this empirical study, it is hypothesized that a potential source of these differing perspectives is the failure of Machine Ethics to be widely observed or explored as an applied ethic and a feasible pursuit.

To realize an interdisciplinary approach to artificial morality, which allows us to gain insights into morality while also engaging practitioners, an informed debate about AMAs is crucial. Our study helps improve the foundations for such debate. It opens avenues for further clarity about views on the development, design, moral agency, and future projections in AMAs research and facilitates the identification of theoretical and practical research opportunities.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s43681-021-00055-2.

**Acknowledgements** The authors acknowledge the European Research Council for financial support of this research (ERC Consolidator grant BEHAVE|724431).

**Funding** European Research Council Consolidator grant BEHAVE—724431.

**Availability of data and material** Data available in the Supplementary Information

## Declarations

**Conflict of interest** The authors have no conflict of interests.

## References


1. Akhtar-Danesh, N., et al.: A comparison between major factor extraction and factor rotation techniques in Q-methodology. Open J. Appl. Sci. **7**(04), 147 (2017)
2. Aliman, N.M., Kester, L.: Augmented utilitarianism for agi safety. In: International Conference on Artificial General Intelligence, pp. 11–21. Springer, Berlin (2019)
3. Allen, C., Wallach, W., Smit, I.: Why machine ethics? IEEE Intell. Syst. **21**(4), 12–17 (2006)
4. Anderson, M., Anderson, S.L.: Machine ethics: Creating an ethical intelligent agent. AI Mag. **28**(4), 15 (2007)
5. Anderson, M., Anderson, S.L.: Robot be good. Sci. Am. **303**(4), 72–77 (2010)
6. Anderson, M., Anderson, S.L.: Machine Ethics. Cambridge University Press, Cambridge (2011)
7. Anderson, M., Anderson, S.L.: Geneth: A general ethical dilemma analyzer. Paladyn J. Behav. Robot. **9**(1), 337–357 (2018)
8. Anderson, M., Anderson, S.L., Armen, C.: Towards machine ethics. In: AAAI-04 workshop on agent organizations: theory and practice, San Jose, CA (2004)
9. Anderson, S.L.: Machine Metaethics, pp. 21–27. Cambridge University Press, Cambridge (2011)
10. Anderson, S.L., Anderson, M.: A prima facie duty approach to machine ethics: Machine learning of features of ethical dilemmas, prima facie duties, and decision principles through a dialogue with ethicists. Mach. Ethics (2011)
11. Anderson, S.L., Anderson, M.: Ai and ethics. AI Ethics (2020)
12. Awad, E., Anderson, M., Anderson, S.L., Liao, B.: An approach for combining ethical principles with public opinion to guide public policy. Artif. Intell. **287**, 103349 (2020)
13. Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.F., Rahwan, I.: The moral machine experiment. Nature **563**(7729), 59 (2018)
14. Bauer, W.A.: Virtuous vs. utilitarian artificial moral agents. AI Soc. **35**(1), 263–271 (2020)
15. Behdadi, D., Munthe, C.: A normative approach to artificial moral agency. Mind Mach. **30**, 195–218 (2020)
16. Behdadi, D., Munthe, C.: A normative approach to artificial moral agency. Minds Mach. **30**(2), 195–218 (2020). https://doi.org/10.1007/s11023-020-09525-8
17. Blass, J.A.: Interactive learning and analogical chaining for moral and commonsense reasoning. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, pp. 4289–4290. AAAI Press, London (2004)
18. Bogosian, K.: Implementation of moral uncertainty in intelligent machines. Minds Mach. **27**(4), 591–608 (2017)
19. Bonnefon, J.F., Shariff, A., Rahwan, I.: The social dilemma of autonomous vehicles. Science **352**(6293), 1573–1576 (2016)
20. Bostrom, N.: How long before superintelligence? (1998)
21. Bostrom, N., Yudkowsky, E.: The ethics of artificial intelligence. Camb. Handb. Artif. Intell. **1**, 316–334 (2014)
22. Bremner, P., Dennis, L.A., Fisher, M., Winfield, A.F.: On proactive, transparent, and verifiable ethical reasoning for robots. Proc. IEEE **107**(3), 541–561 (2019)
23. Brundage, M.: Limitations and risks of machine ethics. J. Exp. Theoret. Artif. Intell. **26**(3), 355–372 (2014)
24. Cervantes, J.A., López, S., Rodríguez, L.F., Cervantes, S., Cervantes, F., Ramos, F.: Artificial moral agents: A survey of the current status. Sci. Eng. Ethics **21**, 317–326 (2019)
25. Danaher, J.: The rise of the robots and the crisis of moral patiency. AI Soc. **34**(1), 129–136 (2019)
26. Dehghani, M., Tomai, E., Forbus, K., Iliev, R., Klenk, M.: Moraldm: A computational modal of moral decision-making. In: Proceedings of the 30th Annual Conference of the Cognitive Science Society (CogSci). Citeseer (2008)
27. Dogan, E., Chatila, R., Chauvier, S., Evans, K., Hadjixenophontos, P., Perrin, J.: Ethics in the design of automated vehicles: The avethics project. In: EDIA@ ECAI, pp. 10–13 (2016)
28. Formosa, P., Ryan, M.: Making moral machines: why we need artificial moral agents. AI Soc. (2020)
29. Fossa, F.: Artificial moral agents: Moral mentors or sensible tools? Ethics Inform. Technol. **20**(2), 115–126 (2018)
30. Fritz, A., Brandt, W., Gimpel, H., Bayer, S.: Moral agency without responsibility? Analysis of three ethical models of human-computer interaction in times of artificial intelligence (AI). De Ethica **6**(1), 3–22 (2020)
31. Gamez, P., Shank, D.B., Arnold, C., North, M.: Artificial virtue: The machine question and perceptions of moral character in artificial moral agents. AI Soc. (2020). https://doi.org/10.1007/s00146-020-00977-1
32. Govindarajulu, N.S., Bringsjord, S., Ghosh, R., Sarathy, V.: Toward the engineering of virtuous machines. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pp. 29–35 (2019)
33. Grodzinsky, F.S., Miller, K.W., Wolf, M.J.: The ethics of designing artificial agents. Ethics Inform. Technol. **10**(2–3), 115–121 (2008)
34. Hallamaa, J., Kalliokoski, T.: How AI systems challenge the conditions of moral agency? In: International Conference on Human-Computer Interaction, pp. 54–64. Springer, Berlin (2020)
35. Headleand, C.J., Teahan, W.J., Cenydd, L.: Sexbots: A case for artificial ethical agents. Connect. Sci. (2019). https://doi.org/10.1080/09540091.2019.1640185
36. Himma, K.E.: Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent? Ethics Inform. Technol. **11**(1), 19–29 (2009)
37. Hooker, J.N., Kim, T.W.N.: Toward non-intuition-based machine and artificial intelligence ethics: A deontological approach based on modal logic. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, pp. 130–136 (2018)
38. Howard, D., Muntean, I.: Artificial moral cognition: Moral functionalism and autonomous moral agency. In: Philosophy and Computing, pp. 121–159. Springer, Berlin (2017)
39. Hunyadi, M.: Artificial moral agents really? In: Wording Robotics, pp. 59–69. Springer, Berlin (2019)
40. Kamal, S., Kocór, M., Grodzińska-Jurczak, M.: Quantifying human subjectivity using q method: When quality meets quantity. Qual. Soc. Rev. **10**(3), 61–79 (2014)
41. Köse, U.: Are we safe enough in the future of artificial intelligence? A discussion on machine ethics and artificial intelligence safety. Broad Res. Artif. Intell. Neurosci. **9**(2), 184–197 (2018)
42. Liao, B., Anderson, M., Anderson, S.L.: Representation, justification, and explanation in a value-driven agent: An argumentation-based approach. AI Ethics (2020)
43. Liao, S.M.: Ethics of Artificial Intelligence. Oxford University Press, Oxford (2020)
44. Mabaso, B.A.: Computationally rational agents can be moral agents. Ethics Inform. Technol. (2020). https://doi.org/10.1007/s10676-020-09527-1
45. Malle, B.F.: Integrating robot ethics and machine morality: The study and design of moral competence in robots. Ethics Inform. Technol. **18**(4), 243–256 (2016). https://doi.org/10.1007/s10676-015-9367-8
46. Martinho, A., Herber, N., Kroesen, M., Chorus, C.: Ethical issues in focus by the autonomous vehicles industry. Transp. Rev. (2021). https://doi.org/10.1080/01441647.2020.1862355
47. McKeown, B., Thomas, D.B.: Q Methodology, vol. 66. Sage publications, London (2013)

48. Misselhorn, C.: Artificial morality. Concepts, issues and challenges. Society **55**(2), 161–169 (2018)
49. Misselhorn, C.: Artificial systems with moral capacities? A research design and its implementation in a geriatric care system. Artif. Intell. **278**, 103179 (2020)
50. Moor, J.H.: The nature, importance, and difficulty of machine ethics. IEEE Intell. Syst. **21**(4), 18–21 (2006)
51. Morley, J., Machado, C.C., Burr, C., Cowls, J., Joshi, I., Taddeo, M., Floridi, L.: The ethics of AI in health care: A mapping review. Soci. Sci. Med. **260**, 113172 (2020). https://doi.org/10.1016/j.socscimed.2020.113172
52. Müller, V.C., Bostrom, N.: Future progress in artificial intelligence: A survey of expert opinion. In: Fundamental Issues of Artificial Intelligence, pp. 555–572. Springer, Berlin (2016)
53. Nallur, V.: Landscape of machine implemented ethics. Sci. Eng. Ethics **26**(5), 2381–2399 (2020)
54. Noothigattu, R., Gaikwad, S.S., Awad, E., Dsouza, S., Rahwan, I., Ravikumar, P., Procaccia, A.D.: A voting-based system for ethical decision making. In: Thirty-Second AAAI Conference on Artificial Intelligence. Springer, Berlin (2018)
55. Poulsen, A., Anderson, M., Anderson, S.L., Byford, B., Fossa, F., Neely, E.L., Rosas, A., Winfield, A.: Responses to a critique of artificial moral agents (2019)
56. Poulsen, A., Burmeister, O.K.: Overcoming carer shortages with care robots: Dynamic value trade-offs in run-time. Australas. J. Inform. Syst. (2019). https://doi.org/10.3127/ajis.v23i0.1688
57. Powers, T.M.: Prospects for a kantian machine. IEEE Intell. Syst. **21**(4), 46–51 (2006)
58. Powers, T.M.: Machines and moral reasoning. Philosophy Now **72**, 15–16 (2009)
59. Rigby, M.J.: Ethical dimensions of using artificial intelligence in health care. AMA J. Ethics **21**(2), 121–124 (2019)
60. Russell, S.: It's not too soon to be wary of ai: We need to act now to protect humanity from future super intelligent machines. IEEE Spectr. **56**(10), 46–51 (2019)
61. Russell, S., Bohannon, J.: Artificial intelligence. Fears of an AI pioneer. Science (New York, NY) **349**(6245), 252 (2015)
62. Scheutz, M.: The need for moral competency in autonomous agent architectures. In: Fundamental Issues of Artificial Intelligence, pp. 517–527. Springer, Berlin (2016)
63. Schmolck, P.: Pq-method, version 2.11 manual. Neibiderg, Germany: University (2002)
64. Shaw, N.P., Stöckel, A., Orr, R.W., Lidbetter, T.F., Cohen, R.: Towards provably moral ai agents in bottom-up learning frameworks. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, pp. 271–277 (2018)
65. Shulman, C., Jonsson, H., Tarleton, N.: Which consequentialism? machine ethics and moral divergence. In: Asia-Pacific Conference on Computing and Philosophy (APCAP 2009), Tokyo, Japan. Citeseer (2009)
66. Sotala, K., Yampolskiy, R.V.: Responses to catastrophic agi risk: A survey. Phys. Scr. **90**(1), 018001 (2014)
67. Stephenson, W.: Technique of factor analysis. Nature (1935)
68. Stephenson, W.: The study of behavior; q-technique and its methodology (1953)
69. Sullins, J.P.: Artificial moral agency in technoethics. In: Handbook of Research on Technoethics. IGI Global, London (2009)
70. Talbert, M.: Moral responsibility. Stanford Encyclopedia of Philosophy (2019)
71. Thornton, S.M., Pan, S., Erlien, S.M., Gerdes, J.C.: Incorporating ethical considerations into automated vehicle control. IEEE Trans. Intell. Transp. Syst. **18**(6), 1429–1439 (2017)
72. Tolmeijer, S., Kneer, M., Sarasua, C., Christen, M., Bernstein, A.: Implementations in machine ethics: A survey (2020)
73. Tonkens, R.: A challenge for machine ethics. Minds Mach. **19**(3), 421 (2009)
74. Tonkens, R.: Out of character: On the creation of virtuous machines. Ethics Inform. Technol. **14**(2), 137–149 (2012)
75. Vanderelst, D., Winfield, A.: An architecture for ethical robots inspired by the simulation theory of cognition. Cognit. Syst. Res. **48**, 56–66 (2018)
76. Vanderelst, D., Winfield, A.: The dark side of ethical robots. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, pp. 317–322 (2018)
77. Verdiesen, I., de Sio, F.S., Dignum, V.: Accountability and control over autonomous weapon systems: A framework for comprehensive human oversight. Minds Mach. (2020)
78. Wallach, W., Allen, C.: Moral Machines: Teaching Robots Right from Wrong. Oxford University Press, Oxford (2008)
79. Wallach, W., Allen, C., Smit, I.: Machine morality: Bottom-up and top-down approaches for modelling human moral faculties. Ai Soc. **22**(4), 565–582 (2008)
80. Waser, M.R.: Designing, implementing and enforcing a coherent system of laws, ethics and morals for intelligent machines (including humans). Proced. Comput. Sci. **71**, 106–111 (2015)
81. Watts, S., Stenner, P.: Doing q methodology: Theory, method and interpretation. Qual. Res. Psychol. **2**(1), 67–91 (2005)
82. Watts, S., Stenner, P.: Doing Q Methodological Research: Theory, Method and Interpretation. Sage, London (2012)
83. Wiegel, V.: Building blocks for artificial moral agents (2006)
84. van Wynsberghe, A., Robbins, S.: Critiquing the reasons for making artificial moral agents. Sci. Eng. Ethics **25**(3), 719–735 (2019)
85. Yampolskiy, R.V.: Artificial intelligence safety engineering: Why machine ethics is a wrong approach. In: Philosophy and Theory of Artificial Intelligence, pp. 389–396. Springer, Berlin (2013)
86. Zabala, A., Sandbrook, C., Mukherjee, N.: When and how to use q methodology to understand perspectives in conservation research. Conserv. Biol. **32**(5), 1185–1194 (2018)