Cambridge Quarterly of Healthcare Ethics (2024), 33: 3, 346–359
doi:10.1017/S0963180123000555

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

Machine Ethics in Care: Could a Moral Avatar Enhance the Autonomy of Care-Dependent Persons?

Catrin Misselhorn

Philosophisches Seminar, Georg-August-Universität Göttingen, Göttingen, Germany
Email: catrin.misselhorn@uni-goettingen.de

Abstract

It is a common view that artificial systems could play an important role in dealing with the shortage of caregivers due to demographic change. One argument to show that this is also in the interest of care-dependent persons is that artificial systems might significantly enhance user autonomy since they might stay longer in their homes. This argument presupposes that the artificial systems in question do not require permanent supervision and control by human caregivers. For this reason, they need the capacity for some degree of moral decision-making and agency to cope with morally relevant situations (artificial morality). Machine ethics provides the theoretical and ethical framework for artificial morality. This article scrutinizes the question how artificial moral agents that enhance user autonomy could look like. It discusses, in particular, the suggestion that they should be designed as moral avatars of their users to enhance user autonomy in a substantial sense.

Keywords: artificial moral agent; artificial morality; capabilities approach; care robot; machine ethics; moral avatar; shortage of caregivers; user autonomy

Introduction

Whereas artificial intelligence aims to model or simulate certain cognitive abilities, artificial morality explores whether and how artificial systems can be furnished with moral capacities.[1] These questions become more and more pressing since the development of increasingly intelligent and autonomous technologies will eventually lead to these systems having to face morally relevant situations. Therefore, it seems almost inevitable to develop machines that have some degree of independent moral decision-making capacity.[2] This is at least one of the central arguments of the proponents of artificial morality.[3] Machine ethics provides the theoretical and ethical framework for thinking about the possibility of artificial morality as well as its desirability for individuals and society from a moral point of view.

Artificial morality presupposes that there can be artificial moral agents (AMAs).[4] One can distinguish between three types of AMAs[5]: moral impact agents, implicit and explicit moral agents. *Moral impact agents* generate moral consequences, but this is not part of their design. *Implicit moral* agents are technical devices whose construction (in terms of hardware) reflects certain moral values, for instance, security. In contrast to moral impact agents and implicit moral agents, *explicit moral agents* are able to explicitly recognize and process morally relevant information and come to moral decisions.

They are agents in the same sense, as a chess program is an agent, which recognizes the information relevant to chess, processes it, and takes decisions with the goal to win the game. Whereas a chess program represents the current positions of the pieces on the chessboard and is able to discern the permissible and most promising moves for reaching its goal, an explicit moral agent represents the options for action in a situation and is able to choose among them based on moral criteria.

It is apparent that explicit moral agents are programmed, and one might wonder whether it is not the designers that really take the moral decisions. However, even in the case of a rather simple chess program, the idea that the designers have supplied the system with a complete set of instructions that explicitly includes every possible outcome is inadequate. This is even more obvious with respect to systems like AlphaGo (the first program that defeated a professional human Go champion), AlphaZero (the first program that did not learn from human games), or MuZero (the first program that even learned the rules of different board games from scratch). The programmers do not prescribe these systems each move in a chess or Go match. They could not even possibly do so because these programs play far better chess or Go than their designers, who could certainly not compete with the world champions in these games.

We can learn from these cases that explicit moral agents must have a certain capacity for autonomous domain-oriented moral reasoning and execution in specific situations.[6] The lack of foreseeability and control of their behavior is one reason to call them agents in their own right. They are, however, moral agents only in a functional sense.[7] That is, their moral reasoning is just a matter of information processing. They are not full moral agents, who are general moral problem solvers and possess consciousness, intentionality, the capacity for reflexive reasoning and moral justification, free will, and moral responsibility. For this reason, we have to deal with a very special type of moral agent that cannot bear responsibility.[8]

The moral psychologist Lawrence Kohlberg suggested that there are three fundamental levels of moral development[9]: the preconventional, the conventional, and the postconventional level. One might use this classification as a heuristic for describing different types of moral agents without sticking to the details of the schema and the strong empirical claim that these levels mark a fixed order of stages in human moral development. The relevant point in our context is that functional moral agents as described here are situated at the level of conventional moral reasoning in Kohlberg's schema. That is, they are following given moral norms but do not reflect on them or even challenge them.

One also has to keep in mind that functional moral agents are moral agents only relationally and not intrinsically. That is, their status as moral agents depends on certain human purposes and interpretations, whereas full moral agents possess this status intrinsically.[10] Take a chatbot like the various versions of ChatGPT in comparison. The inscriptions of such a system are only meaningful in a parasitic way that depends on established human linguistic practices. Thus, the bot has never encountered an apple and does not really know or understand any meanings, but we can interpret its output "apples are healthy" as meaning that apples are healthy. In the same way, an artificial agent will only count as moral, given the background of human moral practice. Just as we call ChatGPT a dialogue partner, we can refer to an artificial agent with the relevant features as a moral agent, even if it is only a functional linguistic or moral agent.

A much-discussed area of application for machine ethics is the use of artificial systems in care, particularly in elder care. Due to demographic change, the number of elderly people will strongly increase in many societies in the coming decades. This will lead to a severe shortage of caregivers. One possibility to meet this shortage is to use artificial systems in geriatric care. As is argued, care systems could alleviate the shortage of caregivers and help to mitigate the expected cost explosion.[11] Another point in favor is that care systems that operate autonomously could enable care-dependent people to live independently in their own homes for longer.[12]

However, the deployment of autonomous artificial systems in care also gives rise to ethical worries. Even if we take for granted that it is undesirable to substitute machines entirely for human caregivers, there will be many situations in which the behavior of an artificial system that is in contact with care-dependent persons will somehow have to be morally regulated by itself. This is particularly true in domestic care, which is the focus here.

Domestic care systems are supposed to fulfill a variety of tasks. They should provide people in need of care with medication, food, and drink and remind them to take them on time and regularly. They should assist care-dependent persons in lying down, sitting up, transferring them from bed to living room, and provide support in their hygienic routines. They should also be able to react if care recipients fall or stop moving. Apart from fulfilling basic needs, the machines are also supposed to be usable for therapeutic purposes and not least for entertainment.

To perform these tasks, they need a body and motor skills. They must have auditory and visual interfaces with the people in need of care and human caregivers. Since most care-dependent persons are presumably people with low technical skills, the machines should be able to communicate in a natural language to make interaction with them easy and intuitive. Given the variety of tasks that care systems are intended to fulfill, the systems must be either extremely specialized or highly complex. Care systems are often imagined as butlers or servants in robotic form that help elderly people to manage everyday life as independently as possible.[13] So far, such machines are not yet available on the market, but there are a number of prototypes.

The aim of this article is to scrutinize the suggestion that a care system that functions as a moral avatar of the user has advantages over other types of care systems. We will first look at how to bring the morals in the machine and survey the main approaches to moral implementation (top-down, bottom-up, hybrid). Then, we will discuss some paradigmatic examples for developing an ethical software module for artificial systems in care. To begin, we will examine attempts to implement a utilitarian ethics. The second example will be a deontological care system. These examples serve as the contrasting foil from which to draw lessons for the design of an AMA in care.

One proposal for designing an AMA implementing the lessons learned will be developed in the fourth part of the article. The elements of this approach are (1) a hybrid approach to moral implementation; (2) a moral theory that fits this hybrid approach particularly well. This theory draws on the capabilities approach by Martha Nussbaum and Amartya Sen; and (3) a design that regards the care system as a moral avatar of the user, that is, a flexible representation or extension of the user's moral character.

One of the main concerns that drives research on AMAs in care is to enhance user autonomy. This is the most important argument to show that artificial systems are not just instrumentally useful to deal with the shortage of caregivers but have a positive moral impact. This concern is also at the core of this article. It will be argued that the proposed AMA can contribute particularly well to enhancing user autonomy. However, the focus on autonomy as the core value of care has been challenged, and we will have to address the question whether strengthening autonomy is a sensible thing to strive for in the context of care at all.

Approaches to moral implementation

Moral implementation is required for designing a functional AMA, that is, an artificial systems that can take moral decisions.[14] One standardly distinguishes between top-down, bottom-up, and hybrid approaches.[15] All three types of moral implementation combine a certain ethical view with a certain approach to software design. The ethical approach provides a description of a certain way to think about moral capacities, whereas the corresponding approach to software design is a method for bringing about the relevant type of capacities in terms of information processing.[16]

Ethical theories and approaches to software design should not be identified; however, the relationship between the two is rather that certain approaches to software design lend themselves more naturally to implementing certain ethical views than others. This has to do with the fact that ethical theories generally have two aims, a theoretical and a practical one.[17] Theoretically, they are supposed to specify the underlying features of actions, persons, and other items of moral evaluation that make them morally good or bad, right or wrong. Practically, they aim at providing decision procedures that will reliably lead moral agents to correct moral judgments. Approaches to software design try to make these two aims computationally tractable.[18]

Top-down approaches bring together an ethical view that conceives of moral capacities as an application of moral principles to particular cases with a top-down approach to software design. Moral principles represent the moral criteria that make actions right or wrong (intrinsically good or bad) in terms of principles. Candidates are, for instance, the utilitarian principle of maximizing utility, Kant's categorical imperative, or Isaac Asimov's three laws of robotics. The decision procedure consists in an inference of the form.[19]

Moral principle: An action is right (wrong), if and only if it has feature x.

Factual claim: This action has feature x.
Conclusion: This action is right (wrong).

Top-down approaches try to implement moral principles in terms of rules in a software, which is then supposed to derive how the system should act in a specific situation. The conclusion will immediately result in an action since AMAs do not have free will and, hence, do not suffer from weakness of the will. They also lack incentives not to act according to their moral judgment since they are neither troubled by emotional distress nor by self-interest.

A major challenge that such a software is facing is how to get from abstract moral principles to particular cases. It seems that context plays a role here, which is notoriously difficult to take into account in designing an AMA. Particularly with respect to utilitarian systems, the question arises as to how much information they should take into account since "the consequences of an action are essentially unbounded in space and time."[20]

Deontological approaches might instead require types of logical inference, which may lead to problems with decidability.[21] There is also a controversy with regard to the question whether moral rules are fully systematizable at all.[22] There is, however, reason to assume that at least a partial systematizability is possible and sufficient for moral implementation. That is, it may be impossible to build a general moral problem solver, but it may not be out of reach to build a domain-specific artificial moral reasoning system.[23] That is to say that the principles in question need not necessarily be general, but they can also be domain-specific.

The alternative to top-down are bottom-up approaches, which do not understand morality as rule-based. This view corresponds well with moral particularism, a meta-ethical position, which rejects the claim that there are strict moral principles and that moral decision-making consist in the application of moral principles to particular cases.[24] Particularism comes in various brands that have different understandings of what principles are and why they ought to be rejected.[25]

The type of moral particularism that is relevant in the context of moral implementation uses to think of moral capacities in terms of case-based reasoning by attending to morally relevant features (or values) that a particular situation instantiates. From these judgments, one can then extrapolate inductively to new cases. To bring about this capacity in machines, they have to be capable of moral learning, for example, by using connectionist approaches.[26] A neural network can be trained to recognize corresponding patterns in the data (i.e., human moral judgments) that allow new cases to be morally decided.[27]

Although machine learning has made great progress in the last decades, there are reasons to doubt the suitability of purely bottom-up approaches for implementing moral capacities in artificial systems used in care. Because the learning processes lack transparency, they pose problems of operationalization, safety, and acceptance. It is difficult to evaluate when precisely a system possesses the capacity for *moral* learning and how it will, in effect, evolve. Since the behavior of such systems is hard to predict and explain, bottom-up approaches might put potential users at risk. It is, moreover, important that autonomous artificial systems do not just behave morally, as a matter of fact, but that the moral basis of their decisions is transparent.

Top-down and bottom-up are the most common ways to think about the implementation of moral capacities in artificial systems. It is, however, also possible to combine both types of approaches. The resulting strategy is called *hybrid approach*.[28] Hybrid approaches operate on a predefined framework of moral values, which is then adapted to specific moral contexts by learning processes. Which values are given depends on the area of deployment of an artificial system and its moral characteristics. There is not really a general answer to the question, which approach to moral implementation is best. It rather depends on the area of application, the context, and purpose of AMAs.

Utilitarian care systems

The first system considered goes back to Michael Anderson and Susan Leigh Anderson, who are among the pioneers of machine ethics.[29] They were working together with Chris Armen on a utilitarian AMA named "Jeremy," in reminiscence of Jeremy Bentham, one of the fathers of utilitarianism. This system

functions as a utilitarian ethics advisor. That is, it does not act fully autonomously but gives moral advice what to do in a certain situation.

It has an input screen that asks the user to describe what they want to do and which people will be affected. Then, the user must enter an estimate of how much pleasure or suffering the action will create for each of these people (values range from 2 = very pleasant, 1 = pleasant, 0 = neither pleasant nor unpleasant, −1 = unpleasant, and − 2 = very unpleasant). Furthermore, the probability of the consequences occurring must be indicated: 0.2 stands for a low probability, 0.5 for a medium probability, and 0.8 for a high probability. The user must provide this information for each person and for each possible course of action.

When the data are complete, Jeremy uses it to calculate which action has the best balance of pleasure and suffering. Although the developers emphasize impartiality as a virtue of Jeremy, the input may be biased. Biases can occur with regard to the user's assessment of how much pleasure or suffering a planned action will produce in others and the likelihood of those consequences occurring. One may be ready to underestimate negative consequences for others if one expects a benefit for oneself.

One way to deal with this problem is to focus only on one user. This can be justified, especially for a domestic care system, to the extent that such a system cares only for one person in need of care, whose wellbeing it is supposed to assess. Although other people may be affected by its actions, many of these people are usually capable of taking care of their own wellbeing. Thus, a care system does not have to consider the impact of an option on the wellbeing of the family or friends of the care-dependent person. This would be different if a care system were responsible for many persons in need of care; that might be a reason to give each user their own personal care robot, which is then not used by other care-dependent people.

The Utilibot Project developed by Christopher Cloos is in line with such considerations based on user-centered utilitarianism, the goal of which is an autonomous mobile robot that can act according to utilitarian principles.[30] This utilitarian system is supposed to guarantee "safety through morality" because the robot avoids all behaviors that lead to injury, harm, or death to the user and directs its behavior toward promoting one user's life, health, and wellbeing.

To cope with the complexity that makes the implementation of utilitarian ethics so difficult, Cloos proposes a decision-theoretic procedure that can model decisions under uncertainty. The effects of an action on the user are determined biometrically using parameters such as pulse, blood pressure, ECG, body temperature, and blood oxygen saturation. When the system is initialized, the user fills out a detailed questionnaire. This includes personal information, medical history, including family history, information on current health status and drug treatment, and lifestyle-related risk factors (such as smoking). This enables the Utilibot to create a detailed medical user profile and make corresponding probability assumptions.

Subsequently, the system measures the vital values and takes them as a yardstick for possible changes in the user's state of health. Permanent biometric monitoring happens, for example, via some kind of smart watch, while a smart home allows the environment to be monitored and related to the measured vital signs. In this way, the system can learn over time, which changes in the environment have which effects on the user.

A particular benefit of the system is its ability to detect the physiological signs of an impending heart attack or stroke at an early stage, taking into account possible risk factors such as high blood pressure. The Utilibot can also make or deepen its diagnosis by observing external symptoms and questioning users. For example, the system can find out if the person has severe chest pain, nausea, or feelings of weakness, which would reinforce a suspicion of a heart attack.

On this basis, the system should arrive at a reliable assessment of the person's state of health. Possible courses of action in a medical emergency include making an emergency call in the event of a stroke, administering aspirin, or providing an oxygen mask, if necessary. Vital signs can also be recorded for later treatment. In addition, the system should be able to deal with the main accident risks that occur in domestic environments, especially falls, poisoning, fire or burns.

The architecture of the system consists of four modules: first, a user network that represents the user's health status, and second, an environmental network that represents the influence of the environment on

user health. The third module is a decision network that integrates the input from the user network and the environmental network and provides a model of possible actions and their utility. These data serve as input to the fourth module, the wellness planner. This module selects which course of action is likely to provide the greatest benefit in an uncertain environment.

This brief description already indicates that data protection is one of the major problems for the Utilibot. Any amount of highly sensitive data about the users (including their families), especially about their state of health, is recorded and passed on. The question of how this is compatible with user privacy is not even raised. Yet, privacy is arguably constitutive for autonomy. One of the reasons why the Utilibot cannot take it adequately into account is its overly narrow focus on the users' wellbeing. There are other moral values, which care-dependent persons might want to pursue exercising their autonomy. Some actions might improve the user's wellbeing at the cost of interfering with other moral rights (as in the case of privacy) or those of others. If, for instance, offending or injuring other people or animals contributes to the wellbeing of the person in need of care, then the system should not take these preferences into account. Such considerations point beyond utilitarianism.

Implementing deontological ethics

The Andersons' themselves have expressed doubts whether utilitarianism is convincing as an ethical approach to machine ethics in care, particularly because of fairness considerations.[31] This led them to turn to the intuitionistic ethics of W. D. Ross.[32] As Kant's moral philosophy, Ross' approach is a deontological ethics based on the concept of duty. In contrast to Kant, Ross distinguishes between different duties that are prima facie binding but may conflict with each other in individual cases, so that a balancing becomes necessary.

There is, for instance, a prima facie duty to keep a given promise, but if harm results, a conflict arises with the prima facie duty not to cause harm. All things considered, it may be our duty to break the promise in favor of avoiding harm. Humans decide such conflicts intuitively, according to Ross. In the case of an AMA, this option is not available; instead, a decision principle is needed that can be invoked when different duties conflict.

Such a principle is supposed to be derived with the help of reflective equilibrium, a method that goes back to John Rawls.[33] This methodological approach consists of arriving at a general moral principle starting from considered moral judgments about particular instances and the rules that are supposed to govern then. Principles and considered moral judgments become more and more aligned in this repetitive process, resulting in an equilibrium. The result should be a principle that prescribes how to proceed when there is a conflict between different prima facie duties.

With respect to care, one suggestion is to implement three principles that are common in bioethics in an artificial system. These principles are supposed to have the status of prima facie duties[34]:

1) the principle of respect for the care-dependent persons' autonomy, that is, of accepting and supporting their autonomous decisions,
2) the principle of nonmaleficence, that is, of causing no harm to the care-dependent person, and
3) the principle of beneficence, that is, of relieving, lessening, or preventing harm and providing benefits to the care-dependent person.

A conflict between these principles might occur, for example, when a competent adult person in need of care is prescribed a medication by a physician but refuses to take it. The conflict arises between the principle of autonomy and the principles of maleficence or beneficence[35]: Do the caregivers have to accept this decision or should they continue to influence the patient?

To represent such conflicts, each possible action is assigned an ordered set of values indicating whether a prima facie duty is fulfilled or violated and to what extent this is the case. A strong violation of duty is represented by −2, a less severe one by −1. 0 means that the duty is not affected, + 1 stands for a moderate fulfillment of duty, and + 2 for a strong one. The decisive factor for the recommendation of an

action is the resulting overall balance, which, however, allows different weightings of the individual parameters.

This claim can be illustrated by a variation of the example given: Suppose a care-dependent person refuses, for religious reasons, to take an antibiotic that the doctor has prescribed to prevent complications in the course of a disease. Suppose further that the complications are not too serious and the patient understands the consequences of their attitude. If the nurse accepts this attitude, the duty to respect the autonomy of the care-dependent person is scored +2. Since the person may suffer some harm and loses beneficent effects as a result, this course of action slightly violates the duties of nonmaleficence and beneficence. Each of them is, therefore, assigned −1.

By further insisting that the care-dependent person should take the medicine, the person's autonomy is compromised, but not as much as if they were forced to take the medication. Therefore, the duty to respect their autonomy is given a value of −1, while the duties to avoid harm and to act for the person's benefit are each given a value of +1. The respective values of the two courses of action are thus: accept: +2/−1/−1 and go on to persuade: −1/+1/+1. Which of these case profiles should be chosen is a matter of the intuitions of the ethics experts.

There are 18 case profiles in total. Based on only four given case profiles, an artificial system was able to derive a principle through inductive logic programming that led to a correct assessment of the other 14 case profiles.[36] This principle, which is also supposed to correspond to the intuitions of experts, states that caregivers should override a patient's decision if this means a slight violation of the patient's autonomy and he or she suffers some harm as a result or misses out on a great benefit. In the Andersons' view, the derivation of this principle is relevant beyond machine ethics because it has not yet been explicitly articulated. This fuels the hope that machines are not only capable of following ethical principles but can even help us discover them.

The principle has been implemented as an example in a few applications related to care. The first is MedEthEx, an expert system that advises users on ethical conflict situations.[37] It requests the morally relevant aspects of an individual case via a user interface, reshapes them in such a way that they can serve as input to a decision-making procedure, gives out the response of this procedure, and justifies it. The second system is called EthEl, and its purpose is to remind elderly people to take medication and, if necessary, to notify relatives or a medical service if this is not done.[38]

EthEl receives information from the doctor or nurse about the prescribed medication, including the time it was taken and the greatest harm that can occur if it is not taken and after what time it is predicted to occur. The greatest benefit resulting from taking the medication is given in relation to the time elapsed. From this input, the system calculates the degree of duty fulfillment or violation of the principles over time, which results from the maximum harm or benefit and the time it takes for these effects to occur. A reminder is issued if this is on balance morally preferable according to the underlying principles. The same is true for notifying relatives or the medical service. The advantage of such an ethically sensitive system over other more rigid devices that are reminding the care-dependent person each time or always call the physician after three refusals to take the medication is that a reminder or notification is given neither too infrequently nor too frequently.

This decision procedure was implemented in a Nao robot, which is thus intended to be an example of a morally acting robot.[39] The robotic platform, manufactured by the company Aldebaran is approximately 58 cm high and weighs 4.3 kg. It has the ability to move around, recognize and produce basic linguistic utterances, grasp objects, and recognize faces and markings. Nao is touch-sensitive and has a wireless Internet interface, infrared sensors, sound localization, and telepresence. These features allow the robot to visit patients to remind them to take their medications or bring them their medications. Nao can communicate in natural language and notify a doctor or nurse if necessary.

In a more recent modified version, the system is no longer provided with a set of given duties, which have to be weighted, but it should itself be able to extract both the duties and the principles from the specification of the morally relevant characteristics of a situation.[40] The goal is to develop a machine that takes moral decisions in a given domain completely autonomously. The starting point is a dialogue with an ethics expert via a graphical user interface with the goal of capturing the morally relevant features of a

situation and the duties resulting from them, as well as generating principles for dealing with morally conflicting situations. The moral principle derived in this way is (after computational adjustment):

An action α is morally preferred to an action β,

if the harm difference between the two is greater than or equal to 1

or

if the difference in benefits is greater than or equal to 3

or

if the harm difference between the two is greater than or equal to -1, the

benefit difference is greater than or equal to -3, and the autonomy difference is greater

or equal to -1.

Following this principle, the machine informs the responsible physician or nursing staff if the patient suffers any harm as a result of not taking a medication at the prescribed time. Notification is also given if the person misses out on a significant benefit by not taking the medication on time. The system does not notify if no harm is expected but only a small benefit is forfeited. In this case, the patient's autonomous decision not to take the medication at the prescribed time prevails.

There remain, however, a number of questions regarding the identification of the morally relevant features, the resulting obligations, and the principle of their weighting.[41] The three principles are clearly relevant in the context of care. Yet, they are still too undifferentiated and do not fully exhaust the spectrum of the relevant moral values.[42] One has to distinguish, for instance, between physical health and psychological wellbeing. Other important moral dimensions are the dignity and self-esteem of the persons in need of care, the value of social interaction, but also enjoyment or play. Issues of data protection, privacy, and intimacy are other morally relevant aspects: It is not at all clear, whether these aspects can simply be subsumed under the principles and whether they can be quantified in this way and offset against each. Moreover, the scope of application of such a system (medication reminders) is very narrow, and it is not easy to imagine how such a system could serve other tasks in care. The more complex care situations become, the more difficult it is to account for possible balances.

Another fundamental objection concerns the role of ethics experts. It remains open whether it is sufficient for only one ethics expert to go through the dialogue with the system or whether a large number of experts should be involved. It is also not clear what distinguishes ethical experts from ordinary people. It is questionable whether there are experts in ethics who can claim the authority to make binding decisions for other persons. The burdens of judgment in the field of ethics render this doubtful for they show the limits that even the most careful rational moral considerations face. The formulation of the burdens of judgment goes back to John Rawls who distinguishes six categories[43]:

1) The empirical evidence relevant to solving a moral problem is complex and often conflicting.
2) Even when there is agreement on the relevant considerations, they may be weighed differently.
3) Moral concepts are sometimes vague and require interpretation.
4) The weighing of evidence and moral values depends on the entire context of experience of the person making the judgment.
5) Conflicts may arise between incompatible normative considerations.
6) In many cases, certain moral values can only be realized at the expense of others.

Despite these limitations, Rawls does not advocate an ethical skepticism that would claim that there is no right or wrong in moral questions or that moral judgments cannot be justified. However, he believes that there is a reasonable pluralism in ethics even among perfectly rational agents. One consequence that one can draw from the fact of reasonable pluralism is that we have a moral right to taking our own moral decisions.

This right has a special weight in care when it comes to situations in which primarily the person in need of care is affected and the moral claims of others are only of secondary importance. For this reason, those in need of care have a special position with regard to the moral evaluation of their situation. It would be a form of paternalism if a group of experts alone were to determine the morally relevant characteristics of care situations and the resulting obligations to those in need of care, without involving them in the decision-making process. Although the Andersons have refined their work in machine ethics, they still adhere to the view that moral implementation should be driven by explicit principles based on the consensus of experts.[44]

This contributes to the third shortcoming of the Andersons approach. While their system is able to generate an abstract rule, which can then be applied to new situations, the rule itself is applied rigidly and regardless of context. Although the system has some ability to learn, their approach is blind to those cases in which the particular needs and moral values of the individuals in need of care become relevant. Recently, they tried to give public opinion a certain weight via crowd-sourced data.[45] Yet, this would only render paternalism in care a matter of expert opinion plus majority view. The step toward a context-sensitive care system, which adapts to the ethical value profiles of the individuals in need of care is denied to the Andersons approach for methodological reasons.

Designing moral avatar

The deficits of the approaches considered so far arguably have to do with their strongly top-down character. An alternative is to opt for a hybrid approach in contrast. Such an approach has a top-down element since it operates within a predefined framework of moral values that are relevant in care. At the same time, it has a bottom-up element because it is capable of moral learning by adapting to the way in which individual users weigh these values. Such a hybrid approach might be ideally suited to incorporate the perspective of those concerned and to provide the level of differentiation and flexibility that is desirable in such a system without collapsing into a mere preference maximizer.[46]

A good starting point in ethics for a hybrid system is Martha Nussbaum's and Amartiya Sen's capability approach.[47] The approach determines capabilities that human beings need for flourishing and to which they are morally entitled, for example, life, bodily health, and integrity, but also the exercise of one's cognitive and sensual capacities and the possibility to form a conception of the good. For the sake of simplicity, one can subsume all these different categories under the term "moral value." These values would represent the top-down element of the suggested hybrid approach. Since the capabilities approach includes a variety of moral values, it can avoid the one-dimensionality of utilitarianism and the difficulty of the deontological approach to take into account the differences between distinct values.

The idea to apply the capability approach to care robots is not new.[48] Yet, an aspect that has so far not been sufficiently taken into account is the context sensitivity of the capabilities, specifically in dependence of the individual life span.[49] People evaluate different kinds of capabilities in different phases of their life differently. Since it is difficult to anticipate the change of perspective for people who have not yet reached a certain age, it is particularly important to take into account the moral perspective of the individuals concerned in geriatric care. Moreover, not all elderly people will weigh moral values in the same way. Autonomy, physical health, and privacy are certainly among the moral values that are relevant in geriatric care. Yet, elderly people may differ in how they specify these values and how they weigh them absolutely and relatively in comparison to each other.

To take into account the perspective of the people who are in need of care when designing an AMA, one has to first specify the capabilities that they regard as important in the context of care more fine-grained. One can find this out with the help of qualitative interviews or focus groups. The results of these interviews or focus groups would be a specification of the relevant moral values of the capabilities approach with respect to the individuals in need of care.

The next step is to operationalize these values such that an artificial system is able to recognize them and weigh them according to the moral value profile of the user. This is the bottom-up element of the approach. The operationalization can be done with the help of scenarios. These scenarios will have to deal with the moral aspects of daily routines in geriatric care, particularly situations in which different

moral values get into conflict. How often and how obtrusively is a geriatric care robot supposed to intervene if a care-recipient is not moving? Individuals who are concerned with health risks might welcome such interventions. In contrast, persons who set great value on autonomy might get rather annoyed. Should the system monitor and register the medical status of the person who is in need of care constantly and report it to a hospital or care facility? Those who are again very worried about health risks and less so about issues of privacy might favor this option, whereas people who are more concerned about privacy might react sensitive to such attempts. They might even decide that their care system should not get connected to the Internet.

The results then have to be implemented. The system might ask people to choose different options with regard to the presented scenarios and infers on this basis the moral value profile of the users, that is, which moral values the user cherishes particularly and how they are ranked in comparison to each other. Based on this information, the system tries to adapt its behavior in new cases to the moral value profile of the user. That is, if it finds out, for instance, that a user ranks autonomy high and is not very worried about health risks, it will intervene less obtrusively than in the case of user profiles that are structured the other way round. The value profile is then constantly refined in the interaction with the user.

Yet, the system's capacity of moral learning must not be restricted to the training phase. The system should constantly adjust its model of the user's ethical value profile also to changes by interacting with the user and giving him or her the possibility to evaluate the adequacy of the system's decisions. This must involve the possibility to override the system's decisions at any point if they are not in accordance with the user's value profile.

In addition to the capacity for moral learning, the system should also have a self-monitoring function and regularly provide status reports whether it is functioning correctly. If this is not the case, the system should inform the user and even turn itself off if there are problems of safety. This is important in order to put the user not at risk in case the system is not functioning properly.

Hence, the system learns by training and interacting with the user to recognize and weigh moral values in alignment with the user's moral value profile and adapts its behavior accordingly. It is not just able to learn what is morally good or bad, but it can treat persons in need of care according to moral standards that these people endorse. Ideally, the system functions as a moral avatar of its user, that is, a representation or extension of the user's moral character.

Such a system might ensure the users' autonomy even better than human caregivers or relatives who might be tempted to impose their own moral views on the care-dependent persons. However, even if one were to agree that such a system could be helpful in domestic care, it would certainly not be suitable for all care-dependent persons. The target group consists of people who are not cognitively impaired and can still take fundamental decisions regarding their life but are physically frail such that they cannot live alone at home without support. They usually also lack the technical expertise that would be necessary to set up an assistance system completely on their own in accordance with their moral values.

Moral avatars and user autonomy

Autonomy is a value that drives much of the research on care robots because they are supposed to enable people in need of care to live more autonomously in their domestic environment. Particularly elderly people could choose to stay at home when they need care instead of moving to a residence. Hence, care robots could contribute to their users' *personal autonomy*, that is, their capacity to determine their life by their own "most cherished" moral values.[50]

Granting this, care robots might still also threaten their users' personal autonomy in different ways. One of them is, as we have seen, paternalism. There is a danger that care robots treat persons in need of care like children and impose upon them, for instance, a healthy lifestyle against their will. The idea of designing an artificial system with moral capacities as a moral avatar was strongly motivated by the desire to avoid the threat to personal autonomy by paternalism, and it presumably fares better in this respect than the other approaches that we have discussed.

Worries might also arise from a Kantian understanding of *moral autonomy*, that is, the capacity to choose and act on moral principles. For Kant, moral agency is the manifestation of autonomy par

excellence and the source of human dignity. If we delegate moral decisions to machines, we seem to renounce in part our autonomy and dignity. One has, however, to keep in mind that the moral avatar in question is not taking moral decisions that the user would have otherwise taken. It rather decides in situations in which the users cannot do this on their own.

Moreover, the setup and training of the system require an exercise of the user's moral capacities. From this perspective, explicitly guiding the user through several scenarios in the training phase has its advantages despite being somewhat cumbersome. Another positive aspect is that a system that would not use such a procedure but infers the user's value profile from their behavior might already violate some of the moral values that the users might cherish when collecting these data, for instance, privacy.

Since the users always have the option to interfere with the system's decisions, their power for moral decision-making is not taken away from them. This is particularly important to protect the users against compromising their personal and moral autonomy, which might result from the lack of transparency of such a system. It is clear that even a system that tries to adapt to the users' moral value profile can get it wrong. Hence, it must be in the user's power to correct the system's decisions at any time.

One might object that this argument speaks for using teleoperated systems in care instead of AMAs in order to allow the user to exercise as much autonomy as possible. Yet, a teleoperated system could not fulfill the same functions as an AMA, for instance, reminding the users of taking their medicine when they are about to forget it. If one considers these tasks as essential for allowing the persons in need of care to live in their domestic environment, AMAs might better support them in exercising their personal autonomy.

A more fundamental objection challenges the claim that autonomy should be a guiding value in the ethics of care at all because it does not acknowledge the dependence and vulnerability that is constitutive of any care relationship. People in need of care who are "very keen to keep what they see as their autonomy and independence" are, hence, accused of not understanding what really matters in care. They even accused of risking their humanity and dignity when they strive at enhancing their autonomy by using care robots.[51]

There is, however, the danger that such a view is patronizing those in need of care, whose desire for autonomy and self-determination is not accepted as a genuine moral claim in care. As these considerations show, it is important that no one should be forced to use a care system. At the same time, one should not morally dismiss people in need of care who would like to enhance their autonomy by using a care system that implements a moral avatar.

It is important, though, that this must not lead to social isolation of the care-dependent persons. This is a danger in societies in which the only social contacts of elderly persons are often their caregivers. Elder care, hence, is a multi-facetted challenge for society as a whole, which can certainly not be solved by technology alone. We have to arrive at an idea of the way in which we want to live in our society with and as elderly people in need of care.

There is also good reason to question the common narrative that the automation of care is necessary to deal with the shortage of caregivers; instead, we have to take into account alternative approaches to the problem on a global level, as well. One of them is economic migration. Why not let people move to places where their workforce is needed, particularly in jobs that involve manual and nonroutine tasks, as it is the case in care? This would help the world's poorest and lead to economic and humanitarian gains.[52] We should take the slogan "making the world a better place," which is so rigorously put forward by many tech-proponents seriously. In the best case, technology could be part of a comprehensive solution, which takes into view the bigger picture.

Acknowledgments. I am thanking two anonymous referees for helpful comments on the first draft of this article.

Competing interest. Author declares that there are no competing interests.

Notes

1. Misselhorn C. Artificial morality. Concepts, issues and challenges. *Society* 2018;**55**:161–9.
2. Allen C, Wallach W. Moral machines: Contradiction in terms of abdication of human responsibility? In: Lin P, Abney K, Bekey GA, eds. *Robot Ethics: The Ethical and Social Implications of Robotics*.

Cambridge: MIT Press; 2011:55–68; Anderson M, Anderson SL. Robot be good: A call for ethical autonomous machines. *Scientific American* 2010;**303**:15–24; Scheutz M. The need for moral competency in autonomous agent architectures. In: Müller VC, ed. *Fundamental Issues of Artificial Intelligence*. Cham: Springer; 2016:515–25; Wallach W. Robot minds and human ethics: The need for a comprehensive model of moral decision making. *Ethics and Information Technology* 2010;**12**:243–50.

3. For a critical assessment of this claim, see note 1, Misselhorn 2018, at 161–9; Van Wynsberghe A, Robbins S. Critizing the reasons for making artificial moral agents. *Science and Engineering Ethics* 2019;**25**:719–35; Misselhorn C. Artificial moral agents. In: Voeneky S, Kellmeyer P, Mueller O, Burgard W, eds. *The Cambridge Handbook of Responsible Artificial Intelligence: Interdisciplinary Perspectives*. Cambridge, NY: Cambridge University Press; 2022:31–49.
4. See note 3, Misselhorn 2022.
5. Moor JH. The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems* 2006;**21**:18–21.
6. Misselhorn C. Collective agency and cooperation in natural and artificial systems. *Philosophical Studies Series* 2015;**122**(4):3–25.
7. See note 3, Misselhorn 2022.
8. Misselhorn C. Digitale Rechtssubjekte, Handlungsfähigkeit und Verantwortung aus philosophischer Sicht. *Verfassungsblog* 2019 Oct 02; available at https://verfassungsblog.de/digitale-rechtssubjekte-handlungsfaehigkeit-und-verantwortung-aus-philosophischer-sicht/ (last accessed 5 June 2023).
9. Kohlberg L. *Essays in Moral Development. Vol. I: The Philosophy of Moral Development*. New York: Harper and Row; 1981; Kohlberg L. *Vol. II: The Psychology of Moral Development*. New York: Harper and Row; 1983.
10. Misselhorn C. Artificial systems with moral capacities? A research design and its implementation in a geriatric care system. *Artificial Intelligence* 2020;**278**:103179, at 37 f.
11. Veruggio G, Operto F. Roboethics – Social and ethical implications of robotics. In: Bruno Siciliano B, Khatib O, eds. *Springer Handbook of Robotics*. Berlin/Heidelberg: Springer; 2008:1499–524.
12. Sorell T, Draper H. Robot carers, ethics, and older people. *Ethics and Information Technology* 2014;**16**:183–95.
13. Dautenhahn K, Woods S, Kaouri C, Walters ML, Koay K, Werry I. What is a robot companion – Friend, assistant or butler? In: *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*. Edmonton, AB: IEEE; 2005:1192–7.
14. See note 3, Misselhorn 2022.
15. Wallach W, Allen C. *Moral Machines – Teaching Robots Right from Wrong*. Oxford: Oxford University Press; 2009.
16. See note 10, Misselhorn 2020 in more detail.
17. Timmons M. *Moral Theory. An Introduction*. 2nd ed. Lanham, MD: Rowman and Littlefield Publishers; 2013, at 3 f.
18. See note 10, Misselhorn 2020 in more detail.
19. See note 17, Timmons 2013, at 5.
20. See note 15, Wallach, Allen 2009, at 86.
21. Powers TM. Prospects for a Kantian machine. In: Anderson M, Anderson SL, eds. *Machine Ethics*. Cambridge: Cambridge University Press; 2011:464–75.
22. Horgan T, Timmons M. What does the frame problem tell us about moral normativity? *Ethical Theory and Moral Practice* 2009;**12**:25–51.
23. See note 10, Misselhorn 2020.
24. Dancy J. *Ethics Without Principles*. Oxford: Oxford University Press; 2004:7.
25. Ridge M, McKeever S. Moral particularism and moral generalism. *Stanford Encyclopedia of Philosophy* 2016; available at https://plato.stanford.edu/entries/moral-particularism-generalism/ (last accessed 5 June 2023).
26. Churchland P. *The Engine of Reason, the Seat of the Soul – A Philosophical Journey into the Brain*. Cambridge: Bradford Books; 1995; Casebeer W. *Natural Ethical Facts – Evolution, Connectionism,*

and Moral Cognition. Cambridge: Bradford Books; 2003; see note 24, Dancy 2004; Gips J. Towards the ethical robot. In: Ford K, Glymour C, Hayes P, eds. *Android Epistemology*. Cambridge: MIT Press; 1995.

27. Guarini M. Computational neural modelling and the philosophy of ethics – Reflections on the particularism-generalism debate. In: Anderson M, Anderson SL, eds. *Machine Ethics*. Cambridge: Cambridge University Press; 2011:316–34.
28. See note 15, Wallach, Allen 2009, at 117ff.
29. Anderson M, Anderson SL, Armen C. Toward machine ethics – Implementing two action-based ethical theories. In: *Proceedings of the AAAI 2005 Fall Symposium on Machine Ethics*. Menlo Park, CA: AAAI; 2005:1–7; For the following, see also Misselhorn C. *Grundfragen der Maschinenethik*. 5th ed. (original 2018). Ditzingen: Reclamverlag; 2022.
30. Cloos C. The utilibot project – An autonomous mobile robot based on utilitarianism. In: *Papers from the AAAI Fall Symposium*. Menlo Park, CA: AAAI; 2005:38–45.
31. See note 29, Anderson et al. 2005.
32. Ross WD. *The Right and the Good*. (original 1930). Oxford: Oxford University Press; 2002.
33. Rawls J. Outline of a decision procedure for ethics. *Philosophical Review* 1951;**60**:177–97.
34. Beauchamp TL, Childress JF. *Principles of Biomedical Ethics*. 7th ed. (original 1979). Oxford: Oxford University Press; 2013:13.
35. At first glance, the principle of maleficence and of beneficence do not appear as categorically different, but rather as a continuum of the same duty in the sense of utilitarianism. There are, however, cases in which this difference might become more apparent, for example, when act α is to amputate the patient's leg (harm), but the patient survives as a result (benefit). The act would then slightly violate the principle of nonmaleficence, but strongly promote the principle of beneficence.
36. Inductive Logic Programming is a type of machine learning for deriving computer programs from data that can take into account background knowledge and examples and resembles human inductive reasoning in certain respects (Muggleton S, De Raeth L. Inductive logic programming. *Theory and methods*. Journal of Logic Programming 1994;**19**,20:629–79.
37. Anderson M, Anderson SL, Armen C. MedEthEx – A prototype medical ethics advisor. In: *Proceedings of the Eighteenth Innovative Applications of Artificial Intelligence Conference*. Boston, MA: AAAI Press; 2006.
38. Anderson M, Anderson SL. EthEl – Toward a principled ethical eldercare robot. In: *AI in Eldercare – New Solutions to Old Problems – Papers from the AAAI Fall Symposium*. Arlington, VA: AAAI; 2008.
39. See note 2, Anderson, Anderson 2010; see note 34, Beauchamp, Childress 1979.
40. Anderson M, Anderson SL. A prima facie duty approach to machine ethics and its application to elder care. In: *Proceedings of the AAAI Workshop on Human–Robot Interaction in Elder Care*. San Francisco, CA: AAAI; 2011.
41. See note 10, Misselhorn 2020, at 278.
42. Misselhorn C, Pompe U, Stapleton M. Ethical considerations regarding the use of social robots in the fourth age. *GeroPsych – The Journal of Gerontopsychology and Geriatric Psychiatry* 2013;**26**:121–33.
43. Rawls J. *Political liberalism*. New York: Columbia University Press;1993:56 f.
44. Anderson M, Anderson SL, Berenz V. A value-driven eldercare robot: Virtual and physical instantiations of a case-supported principle-based behavior paradigm. *Proceeding of the IEEE* 2019;**107** (3):526–40.
45. Awad E, Anderson M, Anderson S, Liao B. An approach for combining ethical principles with public opinion to guide public policy. *Artificial Intelligence* 2020;**287**:103349.
46. See note 10, Misselhorn 2020.
47. Nussbaum M. *Frontiers of Justice: Disability, Nationality, Species Membership*. Cambridge: The Belknap Press of Harvard University Press; 2006; Nussbaum M, Sen A. *The Quality of Life*. Oxford: Clarendon Press; 1993.
48. Coeckelbergh M. How I learned to love the robot. In: Oosterlaken I, an den Hoven J, eds. *The Capability Approach, Technology and Design*. Dordrecht: Springer; 2012:77–86.

49. See note 42, Misselhorn et al. 2013; see note 10, Misselhorn 2020.
50. Darwall S. The value of autonomy and autonomy of the will. *Ethics* 2007;**116**:263–84.
51. Coeckelbergh M. Artificial agents, good care, and modernity. *Theoretical Medicine and Bioethics* 2015;**36**:265–77.
52. Pritchett L. People over robots. The global economy needs immigration before automation. *Foreign Affairs* 2023;**102**:53–64.

Cite this article: Misselhorn C (2024). Machine Ethics in Care: Could a Moral Avatar Enhance the Autonomy of Care-Dependent Persons? *Cambridge Quarterly of Healthcare Ethics* 33: 346–359, doi:10.1017/S0963180123000555