Scientific and Social Research, 2025, Volume 7, Issue 1
https://ojs.bbwpublisher.com/index.php/SSR
Online ISSN: 2981-9946
Print ISSN: 2661-4332

The System Construction of Moral Artificial Intelligence: From the Perspective of Normative Ethical Theories

Yue Hu*

Baoji University of Arts and Sciences, College of Political Science and Law, Baoji 721000, Shanxi, China

***Corresponding author:** Yue Hu, 2045922946@qq.com

Abstract: With the rapid progress of technology, the decision-making and behaviors of artificial intelligence have begun to shift from external settings to internal development. Intelligent agents gradually possess varying degrees of adaptive, decision-making, and behavioral abilities, and their autonomous capabilities are continuously enhanced. For moral considerations, artificial intelligence with autonomous decision-making and behaviors has begun to be regarded as a moral agent. Therefore, how traditional morality can play an autonomous role in intelligent technologies has become a problem that must be faced. The three main theories of normative ethics, consequentialism, deontology, and virtue ethics all have the potential to solve this problem. This article aims to use normative ethical theories to construct an artificial intelligence system capable of making moral decisions, and it is necessary to ensure that the autonomous reasoning of artificial intelligence can be constrained by human social morality and values, remain consistent with human values, and assume the "responsibility" of decision-making.

Keywords: Moral artificial intelligence; Consequentialism; Deontology; Virtue ethics

Online publication: January 23, 2025

1. Introduction

In the early days of computerization, the cyberneticist Norbert Wiener pointed out that technology can help humans become better people and create a more just society, but to achieve this, humans must control technology [1]. Among the three major information technology revolutions in history, the latest one is artificial intelligence technology, and this technology will bring many moral (ethical) problems.

Human morality depends on the rule systems such as moral standards, values, and laws and regulations in the context of human social culture, and the "morality" of artificial intelligence is similar. Researchers have pointed out that the moral behavior of artificial intelligence in society is mainly a normative issue rather than a

descriptive one [2]. Moral norms can not only shape human moral behavior but also be internalized as the self-restraint of artificial intelligence and serve as the foundation for establishing "morality." Therefore, verifying whether the intelligent system complies with the existing legal framework is only the first step. More importantly, it is necessary to ensure that it can interpret and apply human moral values, clarify what the morality of artificial intelligence means, and how the system follows morality and adheres to certain value orientations.

2. Ethical basis of moral artificial intelligence

From the perspective of understanding moral principles and applying them to the design of artificial intelligence systems, normative ethics has special applicability. Consequentialism (utilitarianism), deontology (duty ethics), virtue ethics (virtue ethics), and so on, are typical representatives of normative ethics. They not only put forward norms for human moral behavior but also attempt to solve moral dilemmas. Therefore, it is entirely possible to apply normative ethics to the design of artificial intelligence systems by exploring the established human moral system.

2.1. Consequentialism and artificial intelligence

Consequentialism holds that the morality of an action depends on the consequences of the action [3]. A morally correct action is an action that produces "good" results. Therefore, whether an action is morally correct can be determined by examining its consequences, which are either caused by the action itself (act utilitarianism) or by the general rules that require such an action (rule utilitarianism).

In the consequentialism model, artificial intelligence must know the consequences of an action and what they mean for itself, humans, and other things, and also be able to evaluate these consequences. For humans, it is difficult to determine all the actual consequences of a certain action, let alone those of a rule. However, usually, an action (or rule) will increase or decrease general utility, which is of great significance for guiding the design of artificial intelligence. Considering various possibilities not yet occurred, the evaluation of consequences by moral artificial intelligence is mainly not for actual consequences but for expected consequences. Therefore, moral artificial intelligence based on consequentialism usually adopts heuristic search algorithms, which consist of search, stop, and decision-making strategies and can continuously develop [4]. This is very useful when rapid decision-making is required, but it is also a limitation at the same time.

2.2. Deontology and artificial intelligence

Deontology is a normative moral stance and is unrelated to the character of the actor (compared to virtue ethics). It judges the morality of an action according to rules and does not consider consequences (compared to consequentialism). Humans have the rational ability to create and abide by rules. Rules allow the emergence of duty-based moral norms, which are crucial to human existence. "Duty" is very important in Kantian ethics. Kant believed that responsible actions have moral value. In the deontology model, duty is the starting point and can be transformed into rules, which are divided into rules and meta-rules.

Deontology adheres to the view that moral law is a rational framework for the moral evaluation of the subject's behavior, so it is considered that it can be more easily formalized to produce "responsible" artificial intelligence [5–6]. Its algorithm for rule-based moral judgment is very suitable for the construction of moral artificial intelligence. The application of Kant's categorical imperative to artificial intelligence is regarded as a "top-down"

construction. This method defines the morality of an action based on a set of predetermined rules, and artificial intelligence can only take a certain action when it is allowed by the established rules. Therefore, the expected behavior can be placed in the traditional deontological categories (prohibited, permitted, and mandatory) by simply conducting a consistency test on the action rules. Here, moral judgment is the result of the consistency test, the test is the method of constructing rules, and moral behavior is to establish a set of rules.

2.3. Virtue ethics and artificial intelligence

Producing acceptable moral behavior in a sufficiently small and predictable system like deontology, or moral reasoning when the problem definition is clear enough and information is complete as in consequentialism, is usually not fully satisfied in real-life scenarios because there is a large amount of incomplete information in real-life scenarios. Virtue ethics is rooted in classical moral philosophy and is very useful in evaluating, judging, and taking actions that are in line with character. Virtue ethics focuses on the internal characteristics of people (such as temperance, justice, courage, and wisdom), and unlike deontology and consequentialism, it is a subject-based view.

Aristotle's teleology provides ideas for artificial intelligence research based on virtue ethics. It not only includes the overall goal orientation based on moral behavior but also pays special attention to value orientation. Therefore, the key to constructing moral artificial intelligence based on virtue ethics lies in making values consistent with humans and selecting goals according to the complex values of humans. There is a high similarity between machine learning and virtue ethics. Goal orientation is a core part of modern artificial intelligence, especially advanced robotics. Therefore, virtue ethics is more suitable for the bottom-up moral learning design method based on machine learning. Aristotle also believed that virtues must be discovered and learned through practice, and machine learning also improves the ability of machines to perform tasks through experience. Therefore, machines cannot possess practical wisdom and implement moral behaviors before learning from real data. If virtues can be well combined with functions and task execution, it is entirely possible to develop artificial intelligence based on virtue ethics.

Artificial intelligence based on virtue ethics also has deficiencies in interpretability. It is difficult to explain or prove how its virtues are formed through experience, and virtues are the basis of its actions. If artificial neural networks are used to realize the ability of artificial intelligence to learn virtues, it will bring greater problems because it is almost impossible to extract intuitively understandable reasons from numerous network weights. Therefore, virtue ethics requires more judgment calls and needs to introduce a new interpretive reasoning mechanism to evaluate probabilities and risks, which may not be very reliable in itself.

In short, the following deontology is the simplest way to achieve moral artificial intelligence. Although it is only the direct application of rules, it requires higher-level rules to reason about the actions themselves. Artificial intelligence must know the logical relationship between its actions and the rules. Consequentialism can be achieved through heuristic search, but when information is limited and the impact of actions cascades in continuous interactions, it is necessary to determine the degree of moral reasoning, ignore irrelevant information, and adopt heuristic algorithms with limited search. Virtue ethics can use machine learning techniques, but it requires a new mechanism to reason about motives and consider the behaviors and results caused by motives, which is a more complex model and needs to use algorithms such as the expected utility function to handle "regret" and create new solutions to dilemmas.

3. Responsibility undertaking of moral artificial intelligence

In human society, whether it conforms to moral norms is usually judged according to how moral agents choose behaviors and their consequences. As expected, the behaviors of moral agents will produce good results morally. However, there is still great uncertainty when artificial intelligence takes action. Sometimes the actions will not achieve the expected results or even make wrong choices. When errors occur or laws are violated, it means a problem of responsibility. Therefore, moral artificial intelligence must be able to provide explanations for its decisions and behaviors. If it cannot explain its moral reasoning, it not only means the opacity of the system but also means that it cannot be responsible.

If artificial intelligence lacks a certain form of responsibility, it will not have autonomous ability. Without an accountability system, interactions will not have transparency. Therefore, the construction of moral artificial intelligence should be based on the principles of accountability, responsibility, and transparency (i.e., ART)[7].

3.1. Accountability

Accountability is the primary condition for responsible artificial intelligence, which refers to the system's ability to explain and justify its decision-making mechanism. On the one hand, accountability means that the system can explain. Explanation is to base abstract principles (such as fairness or privacy) on specific system functions. John Langshaw Austin believed that the study of explanation can clarify moral norms in many ways. Human society requires artificial intelligence to prove its moral reasoning ability or at least guarantee the scope of decision-making. Explanation can reduce system opacity and support the understanding of system behaviors and limitations. On the other hand, accountability means that the system's decision-making mechanism must be proven from algorithms and data. The value-sensitive design method has been widely used in the fields of engineering and design and has great potential in ensuring accountability.

3.2. Responsibility

When artificial intelligence has control over actions, it needs to bear responsibility [8]. Consequentialism can play an important role in this regard. However, even if the artificial intelligence system is the direct cause of action, the chain of responsibility must be clear enough, and it is necessary to clarify the relationship between the decision-making behavior of artificial intelligence and stakeholders. For example, when artificial intelligence works as expected, the responsibility lies with the user, which is due to its tool attribute; or when immoral behaviors occur due to errors or accidents, in this case, the designer should bear the responsibility. Although learning and adaptive abilities are the expected features of most artificial intelligence systems, they are ultimately caused by algorithms. Moreover, the consequences of behavior based on learning are usually difficult to fully predict and guarantee, so continuous evaluation is needed, which is the key to the moral learning design method. The responsibility issue of artificial intelligence is very complex and also belongs to a legislative issue.

3.3. Transparency

The explanation of behaviors needs to maintain transparency in the selection and decision-making of algorithms, data sources, and stakeholders. That is to say, it must be possible to review the design and working mode of algorithms. Deontology has outstanding advantages in this regard. The goal of transparency is to provide sufficient information to ensure the safety and controllability of artificial intelligence. If openness can be achieved in all aspects related to the system (i.e., data, design process, algorithms, stakeholders, etc.), transparency in the system

can be guaranteed. The transparency design method is an important method for the design of moral artificial intelligence. The opacity in machine learning, the so-called "black box", is often regarded as one of the main obstacles to transparency. Therefore, it is necessary to reconsider the algorithm design of machine learning, or even go beyond the deep learning model, innovate in-depth research models, and open up new algorithmic frontiers.

4. "Value" construction of moral artificial intelligence

Human morality is universal values and behavioral norms, and values are the basis for explaining attitudes and behavioral motives. The interaction between technology and humans is fundamentally a value interaction, and the value orientation of technology is the result of human application. Artificial intelligence capable of autonomous decision-making, whether it can self-improve or not, inherently needs a "value library", which is its code of conduct. Therefore, values are at the core of moral artificial intelligence.

As artificial intelligence has more and more autonomy in decision-making and environmental operations, it must be designed to learn, adapt, and follow the moral norms and values of the target groups. Research shows that values in different cultures are quite consistent [9]. This indicates that human motives have a similar structure. Of course, even if the types and structures of human motives expressed by values are universal, individuals and groups also have different value "priorities" or "hierarchies." Differences in the order of consideration lead to differences in decisions and behaviors.

Shalom H. Schwartz's theory of basic human values is important in the field of cross-cultural research, which clarifies the common characteristics and differences of values [10]. The core of the theory of basic human values is that values form a circular structure, reflecting the motives expressed by each value. This circular structure encompasses the conflicts and compatibilities among ten universal values recognized by major cultures, and there is a dynamic relationship of conflict and consistency among values. These values constitute four higher-level dimensions: openness, self-enhancement, conservation, and self-transcendence. Values can be slightly or strongly opposed to each other, which causes values to change in a circular structure along two poles.

The structure of the dynamic relationship among values indicates that any behavior that pursues a certain value will conflict with some values but be consistent with others. Schwartz's value structure model provides ideas for the value setting of artificial intelligence. First, the importance rating of values. The Schwartz Value Survey (SVS) directly measures values and scores and ranks the importance of values. Therefore, weights can be used to evaluate and balance values [11]. Second, the direct similarity judgment task. The Portrait Values Questionnaire (PVQ) indirectly measures values and scores the similarity of values. Based on the neural basis of human perceptual similarity judgment, computational models can be established, such as the feature representations generated by the Deep Convolutional Neural Network (DCNN). Third, group classification. The value theory can predict class behaviors with consistent value expressions. Fourth, spatial arrangement. Based on Multi Dimensional Scaling (MDS), the value structure model divides the multidimensional space into different regions containing each value item. MDS is a very useful visual technology in artificial intelligence for evaluating things in multiple dimensions [12].

However, there is still a long way to go for the "value" setting of artificial intelligence for developers, and several problems need to be solved. First, it is necessary to establish a value library for specific groups or individuals affected by artificial intelligence and determine specific norms and attributes. Second, norms have the property of dynamic change, which requires artificial intelligence to have the ability to update and self-improve,

and the process should be transparent [13]. Third, after being integrated into the artificial intelligence system, the system may have algorithm biases and be affected by conflicts among multiple values. Solving such conflicts requires quantitative weighting among values, so the algorithm also needs to be transparent [14].

The realization of "moral" artificial intelligence is a complex systematic project. For artificial intelligence to become a "trustworthy" and "responsible" human companion, it must maintain extensive consistency with humans in terms of moral theoretical basis and expected values [15].

Many moral theories have the potential in this regard. By integrating multiple moral theories, artificial intelligence may generate a better moral system than any individual through machine learning and multi-moral intelligent agents. At present, moral artificial intelligence is still in its infancy and is mostly used in specific fields. It is entirely foreseeable and hopeful that in the future, moral reasoning will be transferred from programmers to intelligent systems for autonomous operation, thereby creating a general moral artificial intelligence with a human-level moral system [16].

Disclosure statement

The author declares no conflict of interest.

References

[1] Cavalier RJ, 2005, Impact of the Internet on Our Moral Lives. State University of New York Press, New York, 19–20.

[2] Bryson JJ, Patiency is not a Virtue: The Design of Intelligent Systems and Systems of Ethics. Ethics and Information Technology, 2018(1): 15–26.

[3] Miller DE, 2003, Actual Consequence Act Utilitarianism and the Best Possible Humans. Ratio, 2003(1): 49–62.

[4] Gigerenzer G, 2008, Why Heuristics Work. Perspectives on Psychological Science, 2008(1): 20–29.

[5] Spahn A, 2020, Digital Objects, Digital Subjects and Digital Societies: Deontology in the Age of Digitalization. Information, 2020(4): 228–242.

[6] Powers T, 2005, Deontological Machine Ethics. Working Papers of the AAAI Fall Symposium on Machine Ethics, 79–80.

[7] Cranefield S, Oren N, Vasconcelos WW, 2018, Accountability for Practical Reasoning Agents. Lujak M. Agreement Technologies. Springer, Douai, 36.

[8] McKenna M, 2008, Putting the Lie on the Control Condition for Moral Responsibility. Philosophical Studies, 2008(1): 29–37.

[9] Schwartz SH, 2006, Basic Human Values: Theory, Measurement, and Applications. Revue Française De Sociologie, 2006(4): 929–968.

[10] Schwartz SH, 2012, An Overview of the Schwartz Theory of Basic Values. Online Readings in Psychology and Culture, 2012(1): 1–20.

[11] Huang G, 2022, Moral Enhancement of Artificial Intelligence: Dynamic Qualification, Normative Position and Application Prospect. Journal of the University of the Chinese Academy of Social Sciences, 2022 (5): 18–30.

[12] Meng LY, 2022, On the Three Research Paths of the Moral Status of Artificial Intelligence. Research on Dialectics of Nature, 2022(2): 30–35.

[13] Zhang JJ, 2022, Discussion on the Moral Subject Status of Artificial Intelligence Body. Quest, 2022(1): 58–65.

[14] Yan KR, 2021, Analysis on the Moral Implication of Artificial Intelligence Design. Social Science in Yunnan

Province, 2021(5): 28–35.

[15] Cheng P, Guss Y, 2021, Philosophical Reflection on the Moral Status of General Artificial Intelligence. Research on Dialectics of Nature, 2021(7): 46–51.

[16] Wu TL, 2021, Is AI Qualified to be a Moral Subject. Philosophy Dynamics, 2021(6): 104–116.