Design Studies and Intelligence Engineering
L.C. Jain et al. (Eds.)

doi:10.3233/FAIA220738

On the Algorithmic Design of Moral Judgment Based on Artificial Intelligence

Shaohui Lu[a,1], Anqi Chen [b] and Xuewei Qian [c]
[a] *School of Marxism Studies, Tsinghua University*
[b] *Academy of Art & Design, Tsinghua University*
[c]*School of Philosophy, Rennmin University*

Abstract. The application of artificial intelligence such as autonomous driving not only brings convenience to humans, but also brings a series of ethical issues To address this issue, this paper attempts to analyze the difference between the moral judgment paths under two major forms of algorithm design, symbolism and connectionism, and the difference with traditional technology ethics from the perspective of the core of AI, i.e., algorithm design, and to provide a solution for the design of AI algorithms based on the perspective of moral judgment.

Keywords. Artificial Intelligence, Algorithmic Design, Symbolism, Connectionism, Moral Judgment

1. Introduction

The design of Algorithm is the base of artificial intelligence, and the core of intelligent behavior is the operation of algorithm. The Medical Expert Systems, Self-Driving Vehicles, Information Push, etc., are all based on certain algorithms. With the wide application of artificial intelligence with the wide application of artificial intelligence algorithm, artificial intelligence is more and more manifested as some kind of autonomy, and the algorithm is more and more replacing the human Thus, the analysis of two existing AI algorithm design approaches in the field of moral judgment is particularly important. In this paper, the thinking on the ethics of artificial intelligence is based on the characteristics of the algorithm. Only by understanding the design of algorithm, can we deeply think about the ethics of artificial intelligence.

2. Algorithm design and its characteristics

Algorithm is a clear and complete specification for solving a kind of common problems in the fields of mathematics and computer science. It represents a strategic mechanism for solving problems by systematic methods. The core is to define the goal of solving

[1] Corresponding Author, Shaohui Lu, Tsinghua University Shuangqing Building 412, Beijing; E-mail: sh-lu20@mails.tsinghua.edu.cn.

problems, create a model for solving problems, and finally, rely on the algorithm to get the required answer *output* according to the input *input* of the problem in a limited step.

Up to now, the *intelligence* of the algorithm cant be understood if it is simply understood by a mechanical and linear reasoning traditional mechanical view. However, leaving aside the essence of algorithm technology and talking in general terms will go beyond the limits of the algorithm and fall into a fantasy. Therefore, this paper will further clarify the nature of the algorithm from the aspect of its computation.

In fact, the inquiry into the computational problem of the algorithm is to ask the boundary problem of the problem domain that the algorithm can solve. Because the algorithm itself is a program used to solve problems in the field of mathematics, it is limited by both the physics of computer and the problem field of mathematics. We cant talk about the algorithm beyond the mathematical calculation itself. In principle, all problems that can be formalized and regularized can be solved by algorithms; On the contrary, the problems beyond the computability cannot be solved by the algorithm. To some extent, this is a circular definition, that is, *the algorithm can solve all the algorithmic problems*. *Whether a problem is computable or not is completely equivalent to whether the problem has a corresponding algorithm.* The reason why algorithms are regarded by many people as impossible to replace human intelligence also stems from this. Many aspects of human intelligence show a kind of anti-formulaic and anti-computational, such as free will, creativity, emotion, morality and so on. In fact, even some purely mathematical problems are not necessarily computable. Some of this non-calculability is due to the objective non-calculability caused by the huge calculation volume, and some are because there is no algorithm that can be used for calculation in theory.

Previous discussions of what constitutes artificial intelligence have tended to fall into a logical misunderstanding, where one tends to think in terms of what problems can be solved to demonstrate the intelligence of an algorithm, such as *The Turing Experiment* or *simulating the human ability to understand stories* as emphasized by Schank (1977). The logic behind this is that (1) certain abilities or performances are a form of intelligence. (2) Some algorithms make machines capable of such abilities or performances. (3) Therefore, such an algorithm is intelligent. But this logic allows us to see only the external manifestations of the algorithm and ignore the internal nature of the algorithm. The solution of many seemingly complex problems is not due to the algorithm itself possessing thinking or intelligence, but rather, it simply means that we have found a computable method for such seemingly complex problems.

Before algorithm substantially subverts computability, we should think more about which problems can be algorithmically calculated. The logic that people are surprised by Alphago's achievements in the field of Go is that we presuppose that Go is a problem that transcends mechanics and needs abstract wisdom to solve. Therefore, according to this logic, the algorithm of artificial intelligence solves the problem of Go, so it is naturally elevated to a wisdom level. But the fact is, we should show that we have found a way to regularize Go and turn it into a digital computable method. Googles *Face Recognition* and Silicon *Valleys Open AI* ... their difficulty is not that they break through the limits of computer algorithms and create a new kind of intelligence. Their difficulty lies more in how to find an algorithm that can be calculated from a problem that seems difficult to be regularized and formalized. Here we would like to quote a quote from Nick Bostrom in *Super Intelligence*. In other words, so far AI is a kind of computation, but we tend to artificially elevate this computation to a higher level of abstract thinking, but this is not true.

The definition of *computability* here is strictly in accordance with a mathematical and computer standard, rather than a generalized computational theory. According to the famous Church-Turing thesis, a computable function is a general recursive function, which simply means that it can be changed in a finite number of steps, starting from a known symbol and continuing to change it step by step according to a set of rules, and eventually arriving at a symbol that meets a pre-condition. Some narrow cognitive psychology separates cognition from emotion, society, etc., and considers cognition as a computation of symbols by the brain. Cognitive scientist Gardner (1985) suggested that *we can understand the human mind through computers because our cognition is a series of computations.* Such a view is actually a generalized computational view, reducing everything to a computational view, simply comparing the brain to a computer, which does not really help us understand the essence of computation. It does not really help us to understand the essence of computation. But such a narrow cognitivism has its own significance, because it notes the existence of a rule of our thinking, and sees that some human mental activities are processual. Many philosophers in the field of epistemology actually have a more or less narrow cognitivist view, such as Locke in his Theory of Human Reason, which points out that some simple ideas provided to us by experience, through the minds self-operation, such as the identification of similarities and differences, give rise to complex ideas. This self-operation has been interpreted by many as a calculation of mental symbols by the brain. To some extent, this pan-computism view facilitates a more abstract and essential understanding of computation, but it is still quite far from the definition and status quo in the context of computing. At present, algorithms should be able to solve problems that are computable in principle, and the physical limits of computation should be considered from a practical point of view, taking into account the storage capacity and operating speed of computers.

3. Value orientation of algorithm design

Algorithm is a clear and complete specification for solving a kind of common problems in the fields of mathematics and computer science. It represents a strategic mechanism for solving problems by systematic methods. The core is to define the goal of solving problems, create a model for solving problems, and finally, rely on the algorithm to get the required answer *output* according to the input of the problem in a limited step.
Does the design of algorithm have ethical value? Or is the operation of the algorithm involved in the field of moral judgment?

Generally, when we talk about the value of algorithms, we usually classify the words *morality* and *ethics* into the ethical value of algorithms without breaking them down too much. In fact, when we discuss whether an algorithm is good, we often evaluate it from two aspects: First, the internal structure of the algorithm. As an embedded computing program, is an algorithm ethical at the beginning of its design? Secondly, the corresponding consequences of algorithm operation, do they bring some ethical problems?

From a technically neutral perspective, algorithms are just digital representations of computers, not about value judgments, but about people using the results of algorithms to give them a value of good or bad. In this way, the algorithm issue is naturally not about morality, and only humans themselves are really involved in moral issues. The algorithm is thus seen as a program based on a symbolic system of operations, a symbolic, formal, computational system of symbols, and the real problems raised by the algorithm can be attributed to the absence of an operational norm, which is not the same as ethics, because the word *norm* is still used for machines. This so-called norm is not the same as ethics, because the term norm is still a control over formal systems such as machines, numbers, and programs. Such a view is superficially reasonable, because algorithms, after all, are still part of technical tools and do not achieve true *general intelligence* or *super intelligence*. The algorithm only gives the results according to the corresponding rules and data, and the algorithm itself is not related to value. However, this not only generalizes algorithms to general technical tools, ignoring their specific technical characteristics, but also allows us to ignore the value orientation load hidden in the computation process before the result.

According to Han Yu, the so-called ethical issues can be summarized in three dimensions: First, Consequentialist ethics, which emphasizes that the results produced have moral influence; The second is Deontological ethics, which emphasizes that the actors should act according to some social moral laws. The third is Virtue ethics, which emphasizes thinking about certain ethical values in action. According to these three dimensions, the ethics of the algorithm can be preliminarily analyzed.

First of all, from the perspective of consequentialist ethics, the practice of algorithms will always produce some kind of results and ethical problems. For example, when we browse the web, the artificial intelligence algorithm will recommend the content that suits us. Its principle is to use the correlation of our personal information, behavioral characteristics, interpersonal relationships and other information to recommend push that meets our personal interests. This trend has become more and more obvious today. This seemingly humanized service makes us more and more closed in an era of more and more open information. We are labeled with a fixed label, and this label will be further confirmed and solidified by the algorithm. Luck Dormehl described the algorithmic discrimination of the U.S. government. They judged which potential terrorists were based on the personal data of passengers, they replace causality with the correlation of data, and bind the personalized information features with some higher-level personality judgment. Therefore, even though many algorithms were designed to facilitate peoples life at the beginning, their impact touched on moral issues.

Secondly, from the ethical point of view of deontology, the algorithm is always a certain rule, and its internal design is loaded with certain value judgments, resulting in ethical problems. The intentional value load is reflected in the fact that our moral judgments, in order to be realized by the algorithm design, must be translated into clear logic and laws in the algorithm, and these pre-defined logic and laws themselves inherently carry certain value presuppositions. Take the *trolley problem* as an example. Suppose one day we achieve autonomous driving and the AI chooses to sacrifice two people to save five, its value presupposition is that when faced with the choice of life, we should start from a utilitarian perspective and save as many lives as possible. Artificial intelligence may also choose to save the scientists or the rich and powerful people, and choose to sacrifice the ordinary people, then its value presumption seems to be more inclined to the value of the lives of the elite than the value of the lives of ordinary people, these value judgments seem to be the algorithms own ruling, but in fact, it is still the designers thinking. The algorithm also has a more hidden value load, which has been separated from the designers original intention, which may originate from the algorithms self-learning and evolution, from a black box principle of the algorithms computation process, and from the emergence of a complexity of the algorithm from the bottom to the top.

Thirdly, from the perspective of value ethics, algorithms as an arithmetic act are always loaded with values. Even though the original intention of algorithm design is value-neutral, it will inevitably bring some value judgment in its development process. Even though the original intention of algorithm design is value-neutral, it will inevitably bring some value judgment in its development process. For example, the algorithm is based on objective data, trying to replace the decision-making of things with cold mathematical equations instead of emotional impulses in order to achieve a value neutrality of results. But in fact, data doesnt mean justice, and data itself may be discriminatory. On the one hand, many data itself are useless data and wrong data, which makes the learning direction of the algorithm wrong at the beginning; On the other hand, small probability events in data learning are often ignored, but the interests or voices of a few people are often hidden in a small amount of data, which makes the algorithm only consider the opinions of the vast majority of people. In addition, the distributed connection makes it more and more difficult to define the subject of moral judgment, and the black box principle makes it difficult to accurately consider the value evaluation, which makes it difficult to identify the subject of moral responsibility and accountability of moral behavior in the algorithm, and further complicates the value load of the algorithm.

Thus, this paper attempts to point out that both the design idea of the algorithm itself and its operation process and results actually involve the field of morality.

4. Moral judgment dilemma in algorithm design

If the moral judgment of the algorithm is necessary and embedded in the algorithm design, then we face a number of ethical dilemmas. The first one is the moral ontological dilemma: how can moral judgments, which are usually regarded as contingent, beyond causal laws, and reflecting the free will of the subject, be given a real, even somewhat mechanical, algorithmic operation? We should think about the possible dilemmas of algorithmic moral judgments in the light of the moral essence and in the light of the technical nature of algorithms.

First, the abstract nature of morality itself makes it difficult to define the goodness of algorithm. The fact judgment is different from the value judgment. The fact judgment points to the actual question *what is this*, while the value judgment points to the ought-to-be question *how should this be*. David Hume once pointed out that any group of claims that are only descriptive cannot necessarily lead to a certain normative claim. The answer to the contingency question relies not on facts, but on a metaphysical understanding of morality. The teleological theory, for example, it attributes all moral behaviors to the pursuit of goodness, so moral problems will eventually point to an ultimate problem, that is, what is good. This is a Socratic inquiry, that is, instead of asking a specific good person or deed, we are looking for what is good itself.

When we trace the causes of moral judgments back to their roots, we find that the understanding of goodness becomes more and more abstract and internalized. For example, the logic behind the value judgment that people should not smoke is *dont smoke to good health to good*, which is a rise from *what you are* to *what you should be*. The metaphysical property of goodness makes it difficult to define it in physical terms, and it is difficult to use a quantified calculation to introduce this property. We can further reflect on the difficulties of moral judgment from the ontological and epistemological perspectives: in the ontological sense, the relevant concepts involved in morality lack the precise meaning and point of reference as the logical positivists say, and the logical symbols of the algorithm are an extrapolationist presupposition that cannot really touch the connotation of the relevant concepts, for goodness and justification. For concepts such as goodness and justification, we cannot find an objective object like Googles recognition of cats pictures; naturally, from an epistemological point of view, all understandings of goodness are at the level of Naturally, from an epistemological point of view, all understandings of goodness are at the level of oughtness, and we cannot be sure that our tracing of moral origins is true and reliable, and both empiricism and rationalism face their own problems in perceiving morality.

Secondly, the diversity of theories and principles behind moral judgment leads to the uncertainty of the standard of goodness. For the understanding of morality, *value hedonism* would associate morality with subjective emotions; while value perfection prefers to look for something that is good in itself apart from the subjective mind. While teleology emphasizes what we should do from the perspective of the good, deontology has a more coercive character. When considering what morality is, virtue ethics emphasizes quality first, while behaviorism emphasizes behavior first... Not only that, the moral principles on which we make moral judgments are sometimes contradictory, which may conflict in concrete reality. The conflict between the moral principles is more and more obvious as the problems become more complicated and the objects of moral judgment become more and more diverse, and it is difficult to be consistent between the principles.

The pluralistic character of moral judgment is also reflected in the cultural relativism property of morality. If moral judgments are considered to be dependent on specific social and cultural contexts, then cultural diversity necessarily leads to moral pluralism, and then we have to consider the issue of compatibility between moral judgments of different cultures. This is different from scientific knowledge and the rules of Go, where the rules of the latter two are objective and universal, while the rules of moral judgment, on the contrary, are full of uncertainty and diversity. Different moral theories, moral principles, and cultural backgrounds put high demands on our algorithmic rules, and, combined with the limitations of algorithms themselves, more often than not, our moral judgments are required to "accommodate" the algorithms, rather than the algorithms to satisfy the moral judgments. For example, under the current technical conditions of algorithms, pleasureist moralism, which emphasizes emotions, or virtue ethics, which emphasizes intrinsic qualities, has to be eliminated (although logically these two may be the most consistent with moral theory, objectively, the technical requirements of algorithms exclude them). Thus, morality under the algorithm is a more narrowly defined moral theory, both conceptually and in scope, and may, for example, tend more toward conformity to moral principles than toward an understanding of the good, and more toward a deontological imperative than a teleological self-improvement.

Finally, even if we put aside the problems of ontology and epistemology, we will understand moral judgment as a series of moral behaviors according to clear rules, such rules themselves remain problematic. Because neither algorithms nor moral judgments are the methods to solve a specific problem, but the solutions to a certain kind of problems, their principles will inevitably become more basic and broad with the increase of the scope involved. For example, when it comes to global ethics, maybe we will emphasize the basic principle of *dont kill people* more often. When it comes to solving the problem of Go, we dont need and cant make precise rules for every step of the algorithm. According to Asimovs law, machines must not harm human beings. However, the more basic this principle is, the richer its connotation becomes. For example, the word *harm* may be a better relief for an extremely painful person… There are actually endless conceptual extensions behind the basic principles. Moreover, the specific choices under the same principle are also different. As for how to get to this end, the methods are very different, and even different cognition based on facts will lead to different moral judgments under the same principle. So, under the same principle, moral behavior itself may be completely different.

In the face of these difficulties, it is not necessary to abandon thinking about the goodness of algorithms, but rather to explore new possibilities for their theory and practice. There seems to be an irreconcilable contradiction between the *ought to be* and *metaphysical natur*e of morality, the *reality of algorithm* and *the naturalistic tendency*. However, morality is by no means the *autonomous field* as the American political philosopher Larry Arnhart said, especially when we make moral judgments, it is bound to be connected with the objective physical world, and we cannot fully understand morality as an internalized quality, which theoretically gives the algorithmic nature of morality. Xu Yingjin pointed out that the emphasis on a moral norm can be digested within the metaphysical framework of materialism. *What is* provides the necessary logical support for *ought to be*, because even legal moral laws cant happen without the possibility of physics. Recognizing the physical attachment of morality, that is, moral spirit and moral behavior are accompanied by each other, will not challenge Kants moral *freedom*. Naturalism on the metaphysical level can be completely compatible with the *congenital* factors on the normative defense level The algorithmic structure of physicalism can be used as the material carrier of morality, thus closely linking moral judgment with physicalism. For some concepts that are difficult to grasp, the algorithm should not be entangled in them too much when making moral judgments, because our grasp of many concepts is relatively vague when making moral judgments. At this time, we can make use of machine learning to make the machine understand what is moral through some specific cases, although it still cant understand the concept of moral .

To sum up, although moral judgment is a value judgment on *oughtness,* it can still be solved by physicalist algorithm design. We need to think more about which moral theory is more realistic under algorithm design. Now it seems that Normative Ethics is stronger than Metaethics; Deontology Ethics may be stronger than Virtue Ethics; Behaviorism and Utilitarianism are also easier to be regulated and algorithmic.

5. Two traditions of intelligent algorithm design

There are two traditions in artificial algorithm design, namely, symbolism and connectionism. In this paper, the technical paths of different algorithms correspond to different moral judgment paths.

There are two traditions of artificial intelligence algorithm designing, namely, symbolism and connectionism. The former is called computer school or logicalism, which represents a kind of deterministic knowledge reasoning. The latter is called bionic school, which is closely related to neural network and deep learning, and represents a kind of uncertain knowledge induction. Technically, on the one hand, the algorithm is still programmed to make the machine run automatically, so its still mechanical; But on the other hand, the emergence of neural network, big data, deep learning and other technologies makes the algorithm more inclined to be an agent without biology in the form and structure, so it is intelligent. The technical paths of different algorithms correspond to different moral judgment paths.

The moral judgment under the semiotics is a top-down top-level design, and the algorithm realizes the so-called morality according to the human setting and clear and accurate moral command law. At this time, morality is artificially transformed into clear and accurate logic rules, which are pre-embedded into the machine program by using symbolic reasoning of algorithm. Such moral commands must be formalizable and algorithmic, or they cannot be translated into a language that machines can understand, beyond the problem domain that algorithms can solve under symbolism. Due to the limitation of algorithm technology, moral judgment at this time is more inclined to deontology, utilitarianism and categorical command, and inevitably gives up the discussion of moral ontology, the affirmation of moral subject intentionality and moral theories such as value hedonism. Therefore, we also bear the corresponding ethical consequences, such as the choice of presupposition ethical value, or the flexibility of the transformation of moral language into concrete actions under logical reasoning.

Moral judgment under connectionism is bottom up self-learning. Through self-learning and self-evolution, the algorithm independently studies what is the right behavior according to the purpose of "good" and forms a kind of morality that is beyond our preset or even beyond our understanding. At this time, the understanding of "good" does not need to be translated into symbolic language, and "good" is internalized into specific behaviors and cases. Through the analysis of its elements and through repeated learning and modification, the algorithm makes its own behavior become more and more towards the goal of "good". At this time, the algorithm is distributed, multi-level and decentralized. Through the modification of the simple rules at the bottom, the intelligent emergence at the high level is achieved, which is similar to the complex intelligent behavior spontaneously emerging at the high level through the interweaving of a large number of simple "activation units" at the bottom of the brain. Such technical characteristics also determine that moral learning is always in an acquired learning process, and with the deepening of algorithm complexity, the calculation process of algorithm tends to be an unknown black box for us.

This paper concludes that the technical characteristics of the algorithm determine what kind of moral judgment the algorithm can make. Thinking about the moral problem of artificial intelligence must return to its own technical characteristics, which determines the boundary of its moral behavior and reveals the hidden moral problem to us. The ethical issues of artificial intelligence are also multi-layered. We cannot deny its status as the subject of moral judgment because it does not reach the high degree of autonomy of strong artificial intelligence, nor can we simply equate the ethics of artificial intelligence with the ethics of technology. The study of morality cannot be confined to the realm of human reason. The reflection on machine morality, especially on a kind of human moral subject, can broaden our cognition of morality.

References

[1] Hao Ningxiang,Computability and unsolvability and their philosophical implications, Research on Dialectics of Nature, 1996,8.

[2] Gardner, H., & Peng, C..(1985), The minds new science: a history of the cognitive revolution. Isis, 69(3), 430.

[3] Xu Yingjin,Even Kants ethics can be algorithmic, Journal of Southwest University, 2017,10.

[4] Yu, H. , Shen, Z. , Miao, C. , Leung, C. , & Lesser, V. R. . (2018). Building ethics into artificial intelligence.

[5] Dennett, D. C., 1997. When HAL kills, Who’s to blame?. In: HAL’s Legacy: Legacy: 2001’s Computer as Dream and Reality, ed. by A.C. Clarke (MIT Press, Cambridge 1997), Google Scholar.