Legal Construction of Algorithm Interpretation: Path of Algorithm Accountability

Luo Weiling[1] *and Liang Deng*[2]

Abstract: Nowadays the development of AI technology is not yet mature, let alone the legal definition and regulation of its type, even the type of technology itself is full of uncertain factors. Because of the rapid development of technology and the openness of theories, scientists have not yet formed a unified consensus and system on cutting-edge technical issues. Therefore, at present, governments all over the world are actively formulating the development plans of AI, but the supervision and regulation of AI are scattered and lagging behind. There is nothing wrong with encouraging the development of new technologies, but the application of technologies requires a responsible response to various ethical demands from human society. No matter what form of AI technology and its application are inseparable from the algorithm and the issue of "algorithm accountability" may probably be a focus of legal regulations on AI and the path of accountability is algorithm interpretation. It is desirable but regrettable that the EU's GDPR stipulates the non-binding "right to explanation". But the stop of GDPR is exactly the starting point of constructing the algorithm interpretation mechanism in law.

The Necessity and Approach of Algorithm Accountability

With the development of AI technology, algorithms are increasingly affecting all aspects of human life. While technology improves efficiency and convenience, it also raises concerns about being "ruled" by algorithms. In the increasingly tense man-machine relationship, human ethical demands, such as security, fairness and privacy, are raised.

[1] Ms. Luo Weiling received her Ph.D. degree from South China Normal University (majoring in Philosophy of Science and Technology and Philosophy of Morality). She is currently a lecturer in philosophy of science and technology at Guangdong Polytechnic Normal University. Email: lingfeng83@126.com. Address: Guangdong Polytechnic Normal University, 293 Zhongshan Avenue West, Tianhe District, Guangzhou, the People's Republic of China.

[2] Mr. Liang Deng is a licensed lawyer in China, partner of Kingson law firm, part-time supervisor of Master of Law in both South China University of Technology and South China Normal University. Mr. Liang received a master's degree in law (majored in jurisprudence) from South China University of Technology and an MBA (a joint project with MIT) from Sun Yat-sen University. Email: liangdeng@junxin.com. Address: Kingson Law Firm, 20/F., Guangfa Finance Centre 713 Dongfengdong Road, Guangzhou, the People's Republic of China.

These ethical demands are progressively one by one, but all point to the crisis of trust. The response to these ethical demands lies in the establishment of man-machine trust, which is based on the accountability of AI algorithm.

The above ethical demands are protected by law, and rights can be created. To be specific, the safety claims can generate civil rights such as the right to life, the right to body and the right to health. The demands for fairness and equal opportunity can generate the basic civil and political rights, such as the right to equality, the right to education, the right to be free from gender discrimination, and consumer rights and interest protection, such as the right to know. Privacy claims can generate civil rights such as right of reputation, right of honor and right of privacy, as well as inviolability of personal dignity. Therefore, it is one of the approaches of algorithm accountability to stipulate specific legal rights for the relative party[3] of algorithm behavior. However, rights that algorithm accountability refers to cover the whole field of public law, private law and social law. And the span and depth of the law involved in these rights determine that it is impossible to create a specific right called "right to counter algorithm".

Thus, it can be seen that the method of algorithm accountability through stipulating legal rights is general but not specific and clear. Since the right can only be remedied by proving that it has been infringed and damaged, and the algorithm liability party can be sued. However, it is extremely difficult to prove that the right has been infringed and the subject of the right has been damaged under the existing legal regulation mode. We understand that the existing legal regulation has its realistic considerations: the logic of AI algorithm operation is difficult to understand, especially for unsupervised learning algorithm, regardless of ordinary people, even its developers cannot predict the results of algorithmic decision-making, the algorithm itself is like a "black box", and trying to open the "black box" falls into the so-called "transparency fallacy".[4] We argue that neither "black box" nor "transparency fallacy" is the reason for avoiding

[3] Here the concept of "relative party" is similar to the concept of administrative relative party in administrative law. Because AI algorithm is not only used in commercial field, but also in public sector, even in commercial application scenarios, algorithmic decision-making in big data environment also has the public function of resource allocation, so the person affected by algorithmic decision-making may be the user of algorithmic decision-making or the party affected by algorithmic decision-making. Specifically, the driver of an autopilot with AI algorithm is affected by the algorithm decision, and belongs to the user of the algorithm instead of relative party. In the scene where the institutions or agencies use algorithmic decision-making to recruit students, loan approval and so on, the influence of algorithmic decision is the relative party of the algorithm application instead of the user.

[4] See, Lilian Edwards, Michael Veale: Slave to the Algorithm? Why a 'Right to an Explanation' is Probably Not the Remedy You are Looking For, *Duke Law & Technology Review* 16(14), 2017, pp. 43, 65-67.

the regulation of AI algorithms. If "black box" cannot be justified by "transparency" or other methods, does it mean that the law will give up restricting a system that may be "out of control"? How do we choose between realism and a responsible attitude? Undoubtedly, the former seems more "economic" in the short term, because it does not need to consider the cost of opening the "black box" or other methods. But in the long run, AI practitioners will be at a loss due to the uncertainty of the basis of behavioral accountability and unclear regulatory boundaries, which is not conducive to the development of science and technology and industry. In our opinion, when analyzing the issue of algorithm accountability, it seems more accurate to replace "right to counter algorithm" with "algorithm obligation" from the perspective of right-holder to the perspective of responsible person. In fact, rights mentioned above are "activated" by clarifying the responsibilities and obligations of the product or service providers and the algorithm developers in the specific scenarios where the AI algorithm works. Thus, the key of algorithm accountability process is not to entitle but to assign obligation, that is, to add and clarify the interpretation obligations of the control party of AI algorithm in the relevant legislation.

The Essentials of Algorithm Interpretation

The Necessity of Algorithm Interpretation

The algorithm disclosure mechanism helps solve the problem of algorithm transparency and further implement the algorithmic responsibility. One of the most attractive and controversial issues in GDPR (General Data Protection Regulations) is the introduction of the concept of algorithm interpretation, which is a useful attempt in our opinion, but the task of GDPR to construct the "right to explanation" has not been successful. Point (71) of the preamble paragraph of GDPR deals with the "right to explanation", and this legislative idea is to protect data subjects from the negative impact of "wrong" automated decision-making. When faced with the adverse results of automated decision-making output, data subjects can correct the conclusions by requiring interpretation and manual intervention, and refuse to accept the adverse results of automated decision-making when the "right to explanation" is not effectively exercised. This is an ideal picture of algorithm responsibility legal regulation, but these provisions are made in the preamble paragraph of GDPR, which is not legally binding. And in the main body of GDPR, we find that articles 12, 13 and 14 deal with algorithmic "transparency" and algorithmic "information providing". It can be seen that the obligation to provide information stipulated in the text of GDPR is limited to making automated decisions (i.e., algorithm processing). It means that the so-called "right to explanation" here refers to the unanalyzed personal data rather than the output results after algorithm processing. The obligation of providing information stipulated in the text of GDPR is completely different from the mechanism of algorithm interpretation that the preamble paragraph attempts to construct. Therefore, the concept of preamble paragraph and text in GDPR is inconsistent with the "right to explanation" of the algorithm, and the deficiency of GDPR lies in its reluctance to express and its failure to establish an effective mechanism of algorithm interpretation. However, the idea of the preamble paragraph of GDPR is worthy of reference. It can be said that the stop of GDPR is exactly the starting point of constructing the algorithm interpretation mechanism in law.

However, some scholars have questioned that it may be futile not only to open the "black box", but also that the so-called "right to explanation" is not what the data subject needs and "in some cases transparency or explanation rights may be overrated or even irrelevant".[5] By analyzing some current cases in the EU, the questioners argue that the right to explanation is often not a remedy sought. Among them, the case of Google in Spain introduced the "right to be forgotten". Questioners believe that the case shows that remedy of the data subject's personal rights can completely replace the "right to explain" with "the right to be forgotten". In the case, the plaintiff asked Google to delete the top search result link related to his name, which pointed to an outdated page of a newspaper file that recorded that he was repaying the government's long-term debt (in fact, he had already repaid it). The plaintiff's appeal in court was to delete "inaccurate" data, and he was not interested in why Google's search algorithm continued to put obsolete data at the top of its rankings. As the case of Google in Spain shows, interpretation does not really mitigate or compensate for the emotional or financial loss suffered by the data subject, it may only serve as a warning to algorithm developers not to make the same mistakes again.[6]

The two factors which were mentioned above, "prevention" and "remedy", are the criteria for determining whether an accountability mechanism is effective. This perspective is meaningful. However, we do not agree with the conclusion that "right to be forgotten" can replace the function of "right to explanation". The questioners think that the algorithm interpretation only has the function of "prevention" but not

[5] Lilian Edwards, Michael Veale: Slave to the Algorithm? Why a 'Right to an Explanation' is Probably Not the Remedy You are Looking For, *Duke Law & Technology Review* 16(14), 2017, p. 43.

[6] See, Lilian Edwards, Michael Veale: Slave to the Algorithm? Why a 'Right to an Explanation' is Probably Not the Remedy You are Looking For, *Duke Law & Technology Review* 16(14), 2017, pp. 41-43.

"remedy", which is exactly the defect of the "right to be forgotten" advocated by the questioners in the algorithm accountability. Let's extend the case of Google in Spain to see whether the plaintiff can get remedy when facing AI algorithm decision-making under the regulatory framework of the "right to be forgotten". Assuming that the outdated data related to the plaintiff's name has not been deleted, when the plaintiff applies for commercial loans, big data will include the "inaccurate" information as a factor into the algorithm, resulting in the rejection of the plaintiff's loan application by the AI algorithm decision system. In this case, the plaintiff claims to delete the relevant information of personal data according to the "right to be forgotten". The result of "remedy" is that the outdated information related to his name is deleted in the relevant web pages, but it will be impossible for the plaintiff to analyze and process the data related to the outdated and "inaccurate" information in the subsequent algorithmic decision-making, so as to avoid the similar situation in the future. It actually reflects the "prevention" function, but the "remedy" function proved by the questioners cannot be satisfied in this situation. Because the object of the "right to be forgotten" is privacy, in the case of Google in Spain exemplified by the questioners, as long as the relevant information is deleted, the remedy function will be realized; but in the scenario extended here by this paper, the plaintiff's claim is to overturn the result of improper algorithmic decision-making, and at this moment resort to the "right to be forgotten" cannot change the result of commercial loan rejection and achieve the goal of "remedy". In addition, due to the lack of algorithm interpretation mechanism, the data subject can only claim the "right to be forgotten" or other data protection rights in the face of adverse results of algorithm decision-making, and such rights only relate to the data itself, which is the input information of algorithm decision-making, and have nothing to do with the algorithm process or algorithm results. In this right- accountability mode, it inevitably leads to the question of the legitimacy of the data source of the data controller, which intensifies the antagonism and contradiction between human and algorithm (or data controller), and damages the shaky foundation of man-machine trust.

The Possibility of Algorithm Interpretation

If algorithm interpretation is necessary, how to solve the obstacles it faces? Faced with the technical and legal constraints of the algorithm interpretation proposed by the questioners, we argue that it is necessary to clarify that such constraints and obstacles only deny the possibility of algorithm transparency or opening the "black box" of the algorithm, while algorithm interpretation is not the same as algorithm transparency or opening the "black box" of the algorithm. In other words, the difficulty of opening the “black box” of the algorithm does not necessarily mean that the feasibility of the algorithm interpretation is low.

Several legal scholars, computer scientists and cognitive science experts from Harvard University published a paper in 2017 to demonstrate the possibility of using algorithmic interpretation for legal accountability.[7] Before discussing the operation of algorithm interpretation, they first make it clear that “explanation does not require knowing the flow of bits through an AI system”. Subsequently, they introduced two technical ideas that make interpretation possible. One is “local explanation”, which refers to the interpretation of specific decisions in the field of AI, rather than the interpretation of the overall behavior of the system. Algorithm interpretation is often done by systematically exploring (external) inputs to determine what factors have the greatest impact on decision results. This explanation is local because the important factors may vary from case to case. For example, for one person, a repayment record may be the reason his loan was rejected, and for another it may be the reason his income was not up to par. In fact, technology has developed a tool called “Local Interpretable Model-Agnostic Explanations” (LIME), which can be used to interpret the predictions of any machine learning classifier.[8] The second technical idea is “Counterfactual Faithfulness”, which helps us to answer this question: Is it a factor that determines the output? And related question: What factors lead to the difference in results? “For example, if a person was told that their income was the determining factor for their loan denial, and then their income increases, they might reasonably expect that the system would now deem them worthy of getting the loan”. “Counterfactual Faithfulness” actually draws on the theoretical resources of philosophy of science and logic on “counterfactual conditionals”. “Counterfactual conditionals are also called ‘virtual implication propositions’. They have the form of ‘if P then Q’, and their preconditions express a situation that does not conform to reality. For example, ‘If the sun does not rise today, there will be no day today’ is a counterfactual conditional sentence.”[9] At the same time, other scholars have made use of this theory to prove the possibility of “Counterfactual explanation” in AI

[7] The analysis of this paragraph, unless otherwise marked, is mainly referred to as: Finale Doshi-Velez, F, Mason Kortz: Accountability of AI Under the Law: The Role of Explanation, Berkman Klein Center Working Group on Explanation and the Law, Berkman Klein Center for Internet & Society working paper, 2017. Website: http://nrs.harvard.edu/urn-3:HUL.InstRepos:34372584, visited on 21 February, 2019.

[8] Talk about LIME, see Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin, “Why Should I Trust You?”: Explaining the Predictions of Any Classifier, 2017. Website: http://nrs.harvard.edu/urn-3:HUL.InstRepos:34372584, visited on 21 February, 2019.

[9] Chen Xiaoping: Scientific Laws and Counterfactual Conditionals on the New Riddle of Induction, *Journal of Sun Yat-sen University (Social Science Edition)* Supplement, 2003, p. 67.

and machine learning.[10] The two technical ideas have one thing in common: neither operation requires an understanding of how an algorithm system makes decisions, or opens an algorithm "black box". In addition, numerous algorithm interpretation practices have also proved the possibility of algorithm interpretation. For example, a team of technology business and medical experts published an article in the *Harvard Business Review* in August 2018 about the potential for preventing discrimination of machine learning algorithm from a product design perspective.[11]

The Hermeneutic Characteristics of Algorithm Interpretation

We find that the essence of "interpretation" elucidated by philosophical hermeneutics has many merits in improving and enriching the connotation of AI algorithm interpretation. According to the position of philosophical hermeneutics, "interpretation and understanding can never discover and completely reproduce the 'author's original intention', and furthermore, the purpose of understanding is not to discover the original intention".[12] Therefore, the "meaning" of interpretation is not a definite existence prior to interpretation and understanding. "Interpretation is recreation in a specific sense, but this recreation is based not on a preceding creative act, but on the interpreter's expression of the image according to the meaning he found in it."[13] Because "prejudice" is an open rather than a closed existence, the interpreting parties can enter the other side's horizon through "understanding" to form a consensus network of "meanings", from which the interpretation can be completed jointly by the parties.

The AI algorithm interpretation is not to seek a definite and priori "solution" but point to the understanding and meaning consensus of participants (including product or service providers, algorithm technology developers and users). We argue that in the specific context of AI algorithm, it refers to the man-machine trust. As Roland Barthes, a French literary critic of the 19th century, said, the author died when the work was born.[14] When AI algorithm was born and put into application, even the developers

[10] See, Sandra Wachter, Brent Mittelstadt, Chris Russell: Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR, *Harvard Journal of Law & Technology* 31(2), 2017, pp. 23-43.

[11] See, Ahmed Abbasi, Li Jingjing, Clifford Gari, Herman Taylor: Make "Fairness by Design" Part *of Machine* Learning, website: https://hbr.org/2018/08/make-fairness-by-design-part-of-machine-learning, visited on 21 February, 2019.

[12] Yin Ding: Fate of Understanding, Beijing: SDX Joint Publishing Company, 1988, p. 50.

[13] Hans-Georg Gadamer: Truth and Method: Basic Characteristics of Philosophical Hermeneutics, Chinese Version translated by Hong Handing, Shanghai: Shanghai Translation Publishing House, 1999, p. 155.

[14] See, Roland Barthes: Death of the Author, Selected Essays by Roland Barthes, Chinese Version translated by Huaiyu, Tianjin: Baihua Literature and Art Publishing House, 2005.

of algorithm technology could not fully and accurately interpret the principle, goal, ethical value and potential risks of algorithm application in the way of "restoration" and "reproduction". The interpretation of these factors is more from the inner conviction of users. Does this mean the shift of interpretation from "author-centered" to "reader-centered"? Does "There are a thousand Hamlets in a thousand people's eyes" lead to AI algorithm interpretation falling into obscure subjective cognition? The answer of this paper is no, because even if the phenomenon of "a thousand hamlets" cannot be avoided at the beginning of the interpretation, "Hamlet" is ultimately "Hamlet", which is the basic category and starting point of the interpretation, and also the image that needs to be reshaped by all parties in the final consensus of meaning. According to the approach of philosophical hermeneutics, the interpretation of AI algorithm is not an arbitrary and one-way interpretation, but a dialogue and multiple interpretations. To be specific, interpretation first means that the service providers or algorithm technology developers re-understand and convey the meaning of the algorithm to the users. Meanwhile, the users start from their "prejudice" to engage in confrontation and integration with the meaning conveyed by the service providers or developers, and finally form the users' inner conviction and reach the consensus of the meaning of all parties.

The Focus of Legal Construction of Algorithm Interpretation

The Relation between Algorithm Interpretation and Legal Interpretation

We believe that algorithm interpretation is an effective path for algorithm accountability. But if the task of algorithm interpretation only aims at algorithm accountability in law, can we use the mature technology of legal interpretation to replace the algorithm which is facing many problems and challenges? To answer this question, it is necessary to sort out the relation between algorithm interpretation and legal interpretation. This section will be carried out from the following two aspects.

Firstly, algorithm interpretation and legal interpretation are different approaches but lead to similar satisfactory results. Comparing algorithm interpretation with the legal interpretation, it can be found that they have a lot in common. Since the extension of the two is too broad, it is necessary to limit the scope of the two before discussing them. Algorithm interpretation refers only to the legally constructed algorithm interpretation for algorithm accountability. And legal interpretation only refers to the authorized interpretation about the algorithm or the algorithm accountability, not including the theoretical interpretation and other informal interpretation. As mentioned above, the algorithm interpretation is multi-dimensional interpretation, while the legal interpretation seems "arbitrary"[15] to some extent. Because the legal interpretation is a "power" rather than a "right", the power of interpretation can only be exercised by the legislator or the judiciary. Yet these differences between algorithm interpretation and legal interpretation do not prevent them from achieving the same effect: both algorithmic interpretation and legal interpretation point to accountability. The purpose of algorithm interpretation is to respond to ethical demands from the perspective of overall social utility. In specific cases, it is to provide remedial measures for users or relatives parties of algorithmic products or services when facing erroneous, harmful and biased algorithmic output. And the other side of the remedy is the responsibility of the product or service providers or algorithm technology developers. Therefore, no matter from the perspective of overall utility or specific case, the ultimate direction of algorithm interpretation is accountability. The purpose of legal interpretation is generally regarded as seeking certainty to ensure the stability and consistency of law application. According to Dworkin, legal interpretation can make the best interpretation of the overall legal practice according to various existing legal materials, from which a consistent or integral principle system can be explained, and then, on this basis, make the best judgment of the practice in law. This best judgment is the "One Right Answer".[16] The pursuit of "One Right Answer" in legal interpretation is also for the purpose of accountability. Hart believed that there was an "open texture" in the law that needs to be interpreted. For example, "No vehicles in the park" plainly means an automobile is forbidden, but what about bicycles, roller skates, toy automobiles and airplanes?[17] The key point of the interpretation of this specification is the responsibility of the owners or users of the "vehicles". Therefore, the purpose of both legal interpretation and algorithm interpretation is to solve the problem of accountability.

Secondly, algorithm interpretation and legal interpretation cannot replace each other. On the one hand, algorithm interpretation cannot replace legal interpretation. As mentioned above, legal norms have an "open texture", so that limitations of language (including everyday language and legal language) make it impossible for legal norms to contain all facts. Legal interpretation will always play an important role. It is because in practice "most cases (*Citation note*: *here should be references to difficult cases*) cannot obtain certainty",[18] so whether to obtain the "One Right Answer" is still a theoretical

[15] Chen Jinzhao: Rule of Law and Legal Methods, Jinan: Shandong People's Publishing House, 2003, p. 216.

[16] See, Ronald Myles Dworkin: A Matter of Principle, Harvard University Press, 1985, pp. 136-137.

[17] See, H. L. A. Hart: The Concept of Law, 2nd ed., Oxford: Clarendon Press, 1994, p. 128.

[18] Richard A. Posner: The Problems of Jurisprudence, Chinese Version translated by Su Li, Beijing: China University of Political Science and Law Press, 2001, p. 265.

problem and there is no "correct answer" in itself. This means that even though the algorithm interpretation mechanism will be perfected in law in the future, the legal interpretation will still play a role in the process of law application at that time. On the other hand, legal interpretation cannot replace algorithm interpretation. At present, legal interpretation is playing a role in the legal regulation of algorithm responsibility. For example, the FTC (the Federal Trade Commission) through the case processing of consent orders to form a common law-like "precedent" rule, which reflects the application of common law reasoning techniques. The fundamental difference between common law and statute law is that "one is a conceptual system and the other a textual system... As far as common law is concerned, interpretation is marginal or irrelevant... Interpretation... probably means to function within a tradition... It could mean less".[19] Therefore, the conceptual system reasoning in common law can be understood as a broad interpretation of law. Another example is the application and interpretation of GDPR in France. French data protection authority (CNIL) ruled on January 22, 2019 that Google provided insufficient information to users, distributed information on multiple pages, and did not obtain valid permission on the issue of personalized advertising. Therefore, Google was fined 50 million euros according to GDPR.[20] A key issue in the case is whether Google's behavior of obtaining user's consent to collect data by ticking options in advance meets the requirement of GDPR on consent. CNIL held that Google's approach was inconsistent with the definitions of "specific", "unambiguous" and "statement", "action" and "signify" defined in terms of GDPR.[21] Therefore, the data collection for Google personalized advertising in this case is illegitimate. This is a typical restrictive interpretation of "restrictive legal meaning, limited to the core",[22] which excludes implied consent from consent. As defined above, the legal interpretation of the above two examples is about the legal interpretation of algorithm responsibility, rather than the algorithm interpretation constructed in law. The objects of legal interpretation are normative texts, legal facts and value factors of conceptual system (the former two correspond to the textual system of enactment law, the latter to the conceptual system of common law), and the objects of algorithm interpretation are the process or output results of algorithm decision-making. Algorithm interpretation, said by its opponents, is something that involves the "black box" or at least the external data of the "black box". In addition, if the facts, acts, conditions and

[19] Richard A. Posner: The Problems of Jurisprudence, Chinese Version translated by Su Li, Beijing: China University of Political Science and Law Press, 2001, pp. 311-312.

[20] See SINA News, website: http://news.sina.com.cn/sf/2019-01-22/doc-ihrfqziz9975281.shtml, visited on 21 February, 2019.

[21] See point (11) of Article 4 of GDPR.

[22] Yang Renshou: Methodology of Law, Beijing: China University of Political Science and Law Press, 1999, p. 148.

standards have been understood outside the legal system before they are interpreted, and a consensus of meaning has been formed, then the non-legal factors have been clarified and the starting point of legal interpretation will be improved, which will help to improve the accuracy of legal interpretation. For this reason, we can say that algorithm interpretation promotes legal interpretation to be closer to the "One Right Answer" in algorithm accountability.

Substantive Composition of Algorithm Interpretation

As discussed above, the legal construction of algorithm interpretation is mainly completed by adding the obligation of algorithm interpretation to algorithm controllers. The obligation of algorithm interpretation in this paper should include at least two aspects:

The first is the obligation of general interpretation. It mainly refers to the pre-regulation of algorithm product/service providers or algorithm technology developers before the algorithm decision is made. Specifically speaking, it is required to respond to the possible ethical appeals of the algorithm one by one in the research and development of algorithm products, namely the so-called Safety by Design, Fairness by Design, Privacy by Design, and others. For example, in the application scenario of autopilot cars, should autonomous driving algorithm distinguish products of different styles, such as positive and decisive type or prudent and steady type, according to people's driving habits? Common sense is that prudent and steady type is safer. But on the road with dense traffic, will the conservative style of autopilot affect the overall efficiency of the road? In addition, as for the ethical appeal of safety, autopilot algorithm also has to "face the great value question of 'whose safety'. If collision cannot be avoided, should this 'safety' be the car's first or related to the possible collision party? Is the relevant algorithm centered on the 'self' of the car, or on the other side, to protect others?"[23] Such ethical conflicts as "Trolley Problem" may be debated endlessly in academic circles, but it will not affect the development and commercial deployment of AI products relying on algorithmic decision-making such as autopilot. Instead, discussions of ethical and value issues prompt AI products developers to think more about these issues during the products blueprint design phase. More importantly, the ethics and value embedding of each product need to be clearly defined in the legislation that it must be disclosed in the product specification, so that consumers, users and the

[23] See Gao Zhaoming, Gao Hao: Information Security Risk Prevention and the Value Principle of Algorithm Law - Two Practical Philosophical Problems of R&D of Autopilot vehicle, *Philosophical Dynamics* 9, 2017, p. 81.

relative parties can know the different "personalities" and value orientations contained in the products of different manufacturers.

Secondly, the obligation of algorithm interpretation includes the obligation of concrete interpretation. It mainly refers to the obligation of the product/service providers or the algorithm technology developers to disclose the factors related to the algorithm or the output result of the algorithm at the request of the specific user or the other party after the algorithm decision is made. In addition to the main obligation to disclose the factors related to the algorithm, the concrete interpretation obligation should also include the collateral obligation to perform in order to remedy the damage suffered by the users or the relative party of the algorithm due to the "improper" decision-making of the algorithm, otherwise the simple disclosure algorithm will lose its significance. The collateral obligations of specific interpretation should include: (a) The obligation to correct, that is, the relevant factors in the algorithm should be corrected and adjusted by the algorithm application service provider or the algorithm technology developer after the damage made by algorithm decision-making has been identified by the relevant authority; (b) Re-decision obligation, that is, the algorithm application service provider or algorithm technology developer shall re-make decisions according to the adjusted and corrected algorithm for the relative party affected by the "wrong" and "improper" algorithm decisions; (c) Stop infringement and compensation obligation, that is, if it is unnecessary or impossible to correct the algorithm or make a new decision, the algorithm application service provider or algorithm technology developer shall assume the obligation to stop infringement and compensate for the loss of the affected relative party of algorithm decision. The above-mentioned collateral obligations are applicable to different situations, and legislation should be adopted to define the joint responsibility, separate responsibility, or supplementary responsibility of the algorithm product/service provider and the algorithm technology developer according to the scenario of exercising specific rights and the principles conducive to remedy.

Procedural Composition of Algorithm Interpretation

We consider that there are at least three main points to consider in the procedural composition of algorithm interpretation.

First, the enforceability or justiciability of algorithm interpretation. The algorithm interpretation is implemented by adding the interpretation obligation, which requires the intervention of public power as a guarantee of coercive force. There are two forms of public power intervention: One is the mode of administrative power. The unified administrative law enforcement department is responsible for filing and investigating

AI algorithm problems, and through similar process of hearings, it is expounded and debated by the algorithm product/service provider, the algorithm technology developer and the algorithm application user or the relative party, on the basis of which, the administrative law enforcement department determines the algorithm responsibility and supervise the implementation of the algorithm interpretation obligation. Another is the judicial mode, that is, the users or the relative party of the algorithm apply to initiate litigation or arbitration on the basis of the specific right of claim, and the disputes of algorithm interpretation are settled by the judicial organ, and the oblige applies for compulsory enforcement, requiring the specific body responsible to fulfil the obligation of protecting the relative party.

Second, the trigger condition of algorithm interpretation. The previous analysis shows that algorithm interpretation is necessary and feasible. However, it cannot be ignored that algorithm interpretation is costly and expensive. It not only consumes technical resources, but also occupies administrative and judicial resources. Therefore, in order to avoid the abuse of algorithm interpretation mechanism by obligee, legislation should set the trigger condition of algorithm interpretation, so as to ensure the efficiency of algorithm interpretation. On the one hand, since the purpose of algorithm interpretation is to assign responsibility and remedy the party who suffers from the loss of right, the obligee who makes the request for algorithm interpretation should be the party who suffers from the adverse impact of algorithm decision. Taking the automated decision-making of loan approval algorithm as an example again, if the applicant's loan has been approved, and the law should not support his request, which is merely to satisfy his curiosity, for algorithm interpretation. On the other hand, it is suggested to introduce GDPR regulations on automated decision-making.[24] In this paper, the authors think that the requester of algorithm interpretation should suffer direct and legal effects due to the algorithm decision. Specifically, it is necessary to eliminate the uncontrollable indirect losses. If the loan apply based on automated decision is rejected and the house purchase is affected, the loss of deposit, penalty and commission fee caused by the failure to continue the house purchase transaction can be counted as the loss directly affected, and the increase of house re-purchase cost caused by house price rise is not directly affected. The principle of "legal effects" specifically refers to the influence of "improper" automatic decisions is based on the violation of specific rights. If the objects of damage are not specific rights but non-statutory interests, the request for algorithm interpretation should not be supported. For example, in the blind

[24] According to Article 22 of GDPR, "data subjects have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her".

date matching based on automatic decision screening, in the case the relative party is not satisfied with the object of machine matching or fails in the blind date caused by the wrong matching of the machine, the relative party suffers no specific right of legal protection, and the relative party shall not get legal support when he asks for algorithm explanation.

Third, the proof standard of algorithm interpretation. Coercive force is the guarantee to enforceability or justiciability of the algorithm interpretation, and the prerequisite for coercive force is the legitimacy and rationality of the implementation of the algorithm accountability, which is based on "due process".[25] It means that the user or the relative party of the algorithm application can resist the validity of the algorithm decision-making by inquiring and questioning the relevant factors of the algorithm decision-making within the framework of due process, while the provider of the algorithm product/service or the developer of the algorithm technology can respond to the heckling by interpreting the relevant factors to prove the validity of the algorithm decision-making. In due process, algorithm product/service providers or algorithm technology developers bear more burden of proof. They need to prove that algorithm decision-making conforms to the interpretation standards recognized by law, otherwise they should provide remedy measures to algorithm users or the relative parties. Since algorithm interpretation is not equal to algorithm transparency or open algorithm "black box", it is unnecessary to require algorithm product/service provider or algorithm technology developer to interpret all input data, interpret the whole algorithm logic, open source code, etc. With regard to the interpretation criteria that should be stipulated in legislation, this paper basically agrees with the following views: "For the purpose of remedy, the content of the interpretation should meet two criteria: first, the relevance, that is, it must be related to the specific automated decision-making of the relative person; second, the relative person can understand. The ultimate goal is to prove that automated decision-making can be trusted. In addition to comprehensibility and relevance, different interpretation criteria should be formulated for different automated decision-making contents..."[26] Relevance criterion embodies the principle of local interpretation and ensures the feasibility of algorithm interpretation; and comprehensibility criterion embodies the non-arbitrary pluralism advocated by philosophical hermeneutics, that is, algorithm product/service providers, algorithmic

[25] See, Danielle Keats Citron: Technological Due Process, *Washington University Law Review* 85(6), 2008, p. 1249; Danielle Keats Citron, Frank Pasquale: The Scored Society: Due Process for Automated Predictions, *Washington Law Review* 89(1), 2014.

[26] Zhang Linghan: Research on Right to Explanation of Auto Decision-Making, *Law Science (Journal of Northwest University of Political Science and Law)* 3, 2018, p. 72.

technology developers, algorithm application users and the relative parties are all interpreters and have the right to understand a specific algorithm decision-making. The ideal scenario of algorithm interpretation should be through dialogue, so that all parties of interpretation can blend horizons and reach a consensus of meaning: algorithm interpretation can be understood and accepted by all parties, and man-machine trust can be established.

References

Abbasi, A., Li J., Clifford, G. & Taylor, H.: Make "Fairness by Design" Part of Machine Learning, *Harvard Business Review*, 2018. Website: https://hbr.org/2018/08/make-fairness-by-design-part-of-machine-learning.

Barthes, Roland: *Death of the Author, Selected Essays by Roland Barthes*, Chinese Version translated by Huaiyu, Tianjin: Baihua Literature and Art Publishing House, 2005.

Chen Jinzhao: *Rule of Law and Legal Methods*, Jinan: Shandong People's Publishing House, 2003.

Chen Xiaoping: Scientific Laws and Counterfactual Conditionals on the New Riddle of Induction, *Journal of Sun Yat-sen University* (Social Science Edition), 2003 Supplement.

Citron, Danielle Keats: Technological Due Process, *Washington University Law Review* 85(6), 2008.

Citron, D. K. & Pasquale, F.: The Scored Society: Due Process for Automated Predictions, *Washington Law Review* 89(1), 2014.

Doshi-Velez, F. & Kortz, M. A.: Accountability of AI Under the Law: The Role of Explanation, Berkman Klein Center Working Group on Explanation and the Law, *Berkman Klein Center for Internet & Society working paper*, 2017. Website: http://nrs.harvard.edu/urn-3:HUL.InstRepos:34372584.

Dworkin, Ronald Myles: *A Matter of Principle*, Harvard University Press, 1985.

Edwards, L.& Veale, M.: Slave to the Algorithm? Why a 'Right to an Explanation' is Probably Not the Remedy You are Looking For, *Duke Law & Technology Review* 16(14), 2017.

Gadamer, Hans-Georg: *Truth and Method: Basic Characteristics of Philosophical Hermeneutics* (Volume 1), Chinese Version translated by Hong Handing, Shanghai: Shanghai Translation Publishing House, 1999.

Gao Z. & Gao H.: Information Security Risk Prevention and the Value Principle of Algorithm Law - Two Practical Philosophical Problems of R&D of Autopilot vehicle, *Philosophical Dynamics* 9, 2017.

Hart, H. L. A.: *The Concept of Law,* 2nd ed., Oxford: Clarendon Press, 1994.

Posner, Richard A.: The Problems of Jurisprudence, Chinese Version translated by Su Li, Beijing: *China University of Political Science and Law Press*, 2001.

Ribeiro, M. T., Singh, S. & Guestrin, C.: *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*, 2017. Website: http://nrs.harvard.edu/urn-3:HUL.InstRepos:34372584.

Wachter, S., Mittelstadt, B. & Russell, C.: Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR, *Harvard Journal of Law & Technology* 31(2), 2017.

Yang Renshou: Methodology of Law, Beijing: *China University of Political Science and Law Press*, 1999.

Yin Ding: *Fate of Understanding,* Beijing: SDX Joint Publishing Company, 1988.

Zhang Linghan, Research on Right to Explanation of Auto Decision-Making, *Law Science (Journal of Northwest University of Political Science and Law)* 3, 2018.