Machine Learning Ethics in the Context of Justice Intuition

Natalia Mamedova[1,*], *Arkadiy* Urintsov[1], *Nina* Komleva[1], *Olga* Staroverova[1] and *Boris* Fedorov[1]

[1]Plekhanov Russian University of Economics, 36, Stremyanny lane, 117997, Moscow, Russia

Abstract. The article considers the ethics of machine learning in connection with such categories of social philosophy as justice, conviction, value. The ethics of machine learning is presented as a special case of a mathematical model of a dilemma - whether it corresponds to the "learning" algorithm of the intuition of justice or not. It has been established that the use of machine learning for decision making has a prospect only within the limits of the intuition of justice field based on fair algorithms. It is proposed to determine the effectiveness of the decision, considering the ethical component and given ethical restrictions. The cyclical nature of the relationship between the algorithmic algorithms subprocesses in machine learning and the stages of conducting mining analysis projects using the CRISP methodology has been established. The value of ethical constraints for each of the algorithmic processes has been determined. The provisions of the Theory of System Restriction are applied to find a way to measure the effect of ethical restrictions on the "learning" algorithm

1 Introduction

The intuition of justice is regarded as a sincere, emotionally saturated belief in the justice of some position. This belief is a prism through which the refraction of any information is received by the subject and the adoption of intuitively obvious decisions.

What everyone considers to be fair is decided individually for oneself, but the process of developing the intuition of justice is completely determined by the environment in which the individual develops. His own nature and range of external influences ultimately determine whether to take each individual position on faith or rethink, accepting or denying it. Thus, the pivot points with which any concepts of justice generated by society should be compared fall into the field of the intuition of the justice of an individual [1].

Studies in the history of philosophy show that the concepts of justice, the cornerstone of which is value or benefit (individual or collective), are readily recognized by an individual as intuitively valid, as proven and reliable [2]. Supporting the concept of justice by a multitude of individuals gives it an egregorial character, creating an even more stable foundation for conviction. As a result, a conditionally permanent conviction complex (reliable, intuitively obvious, just) is formed, localized in the field of intuition of justice.

Thus, a local conviction that has value within a society is recognized as fair [3]. And yet, freedom of thought is the highest iteration of the intuition of justice, it is capable of leveling the individual's egregorial dependence, going beyond a fair local judgment. We owe freedom of thought to all the results of human activity, both the best and the worst.

Accepting the freedom of thought is authentic intuition of justice of one individual in relation to himself and others. But for artificial intelligence (hereinafter - AI) freedom of thought is not supposed - neither for a strong AI, nor for a weak AI. First, the prospect of free thinking for AI is seen as a frightening way of developing the future [4]. Second, the level of technology development is insufficient to create a strong (true, general) AI - a machine that is capable of thinking, learning new things and being aware of itself. As for the weak (narrow) AI, its freedom of thinking is limited by the framework of the machine learning algorithm, according to which the machine solves certain human problems. This happens during machine learning without explicit programming by learning from precedents and as part of a training set. This kind of "learning" algorithm does exactly what it can and can do what is provided for by the mathematical model. Further, we will consider AI only in the aspect of machine learning (weak AI), excluding the field of futurology.

The mathematical model of the "learning" algorithm is a product of human thinking, and in the process of its development ethical deviations can be made. Using the logical method of bringing to the point of absurdity (reductio ad absurdum), we define that the maximum ethical deviation should be considered the complete absence of ethical restrictions in the "learning" algorithm. Although other ethical deviations, such as "pollution", "poisoning" of the initial data of the training sample, manipulations with feedback loops and false correlations, or unethicality of the task itself, can also be applied to the ethical deviations scale.

To identify such deviations in most cases is not difficult, since the intuition of justice works flawlessly, but, as we have already mentioned, the limits of justice

are determined individually. Therefore, the parameters of the acceptable level of ethical deviations are variable, they are difficult to formalize in machine learning. However, it is necessary to do so that the "learning" algorithm itself and the algorithmization process fit into the logic of the egregorial concept of justice accepted in society.

In other words, the ethical component of the "learning" algorithm must meet the intuition of the justice of individuals, must have value within society. Therefore, the ethics of machine learning ethics, namely the ethical limitations of the "learning algorithm", come to the forefront. Two approaches are applicable to machine learning ethics. According to the first approach, it is aimed at expanding the knowledge of human ethics with the help of "learning" algorithms. In accordance with the second approach, the ethical component is used in the development of machine learning algorithms. In this study, the second approach is applied.

2 Materials and methods

Machine learning is a section (subset) of AI. On the one hand, this section examines the algorithms and statistical models used by computer systems to effectively perform a specific task [5]. A special feature is that instead of instructions, the computer system relies on patterns and established patterns (dependencies). On the other hand, this section of AI is studying methods for constructing algorithms capable of learning [6]. This involves developing computer programs that can access and use data for training.

The learning process begins with observations or baseline data, however, the search for patterns and decision making based on precedents are carried out using predetermined algorithms. Considering that the ethical component of machine learning is formed directly by algorithmization, we will focus on the concept of algorithm in more detail.

It is generally accepted to use the concept of an algorithm as a computational procedure, which, according to a training sample, sets up a model [7]. The result of the procedure is a function that establishes (approximates) dependence. This function is called a hypothesis or concept [8]. We also agree with the opinion that the function itself is also an algorithm [9], since, like the computational procedure itself, it determines the efficiency of computer-aided solution of the problem. We consider it expedient to determine the effectiveness of the decision taking into account the ethical component and given ethical restrictions. Thus, machine learning will develop in the field of the intuition of justice.

Studies in machine learning are carried out through experiments based on model or real data. The heuristic approach compensates for the difference between theoretical assumptions and the conditions of real problems. And the ultimate goal of machine learning is not even confirmation of the efficiency of the algorithm, and not automating the process of learning a computer without human help. It is obvious to us that the search for patterns in the data carries the potential for making the best decision in the future based on the examples (precedents) set today. However, the best solution should take into account ethical restrictions; otherwise the solution proposed by the machine may be contrary to the intuition of justice. From here we can conclude that an ethical restriction is formulated by imposing a categorical ban on the commission of an operation (action, choice).

This characteristic of the ethical constraint allows us to go further and determine that the ethics of machine learning is a special case of a mathematical model of a dilemma, since it involves only two answers - yes or no. There are questions at what stage or at what stages of the algorithmization process it is necessary to establish ethical restrictions, what focal points need to be determined to measure ethical restrictions.

Answering the first question asked, it is necessary to initially present the general process of machine learning and focus on the process of algorithmization in machine learning. The projection of the ethical component on the algorithmization process will help answer the question - at what stage or stages are ethical restrictions necessary or appropriate.

Objectively, the core of machine learning is the construction of a model of general dependence (patterns, relationships) of data. Unlike formalized expert knowledge, transformed by deductive learning [10], the model of general dependence is formed by learning from precedents - inductive learning. Learning from precedents is synonymous with an earlier notion - recovering dependencies from empirical data [11]. This concept is related to computational learning theory (COLT), which studies mathematical dependencies and quantitative restrictions on the parameters of the maximum complexity of a model and the reliability of data.

The process begins with the selection of cases. The precedent is considered relevant if it corresponds to some private data that describe the precedent. A set of descriptions of precedents is a training sample. Next, a learning algorithm is formed that reveals a general dependence (pattern, relationship) of data on all the precedents of the training set, as well as on all precedents for which the same description exists. The algorithm is designed to solve two types of problems - the regression problem (real value) and the classification problem (discrete value). Considering the method of minimizing the empirical risk [12], the operation of the algorithm is limited by fixing the functional quality of the algorithm. The functional quality of the algorithm shows how well the model describes the collected data. Quality is determined by the nature of the general dependence (pattern, relationship) of data. Thus, it is easy to imagine the process of machine learning.

Finding an adaptation process within the limits of normal distribution allows you to manage this process. The control function of the control is performed by the operator - the subject who monitors the process of adaptation of students. In the interval of normal distribution, the operator gets the opportunity to gradually increase the pace of training and expand the

list of operations performed on arrays of information. The position outside the normal distribution interval means that the adaptation occurs beyond the expected learning path. This implies the use of special measures to support and adjust the learning process based on the principles of the individualization of learning.

3 Results and discussion

We know that the "learning" algorithm searches for such a set of model parameters in which the quality functional takes an optimal value on a given training sample [13]. In this case, the problem is considered solved. But if the machine learning process is decomposed into subprocesses and ethical restrictions are applied to each of them, we obtain the following data.

To separate the algorithmization process into subprocesses, the classification of the stages of conducting CRISP-DM data mining projects (CRoss Industry Standard Process for Data Mining), as the most common and popular project management methodology, was used [14]. The process of algorithmization of machine learning in it corresponds to 4 of 6 stages, implemented sequentially - understanding of business, understanding data, data preparation, modeling. The consistency of algorithmic algorithms subprocesses in machine learning and the stages of conducting data mining projects is presented in Table I. Conformity criteria are established by highlighting the common feature of the actions performed.

Table 1. Conformity data.

The name of the stage in the CRISP-DM methodology	**General sign of action**	**The name of the subprocess in ML**
Understanding business	Definition of the goal, assessment of the situation	Building a model of general data dependency
Understanding the data	Data collection and description	Choosing a finite set of use cases
Data preparation	Selection, cleaning, casting data	Determination of algorithm quality
Modeling	Testing, calibration model	Building algorithms of the restored dependencies

A subprocess of building a model of general data dependency. In essence, this subprocess is a selection of parameters for a predictive model. The need for a model arises when there is reliable data in the required volume, but due to dynamically changing conditions it is difficult to formulate the rules by which the forecast is made. That is, within the framework of this subprocess, the problem of machine learning is set. The application of the ethical component of algorithmization in machine learning to this subprocess consists in studying the meaning of the problem and assessing its conformity with the intuition of justice. That is, the result of applying the ethical component will be the answer to the question of whether the task is ethical or not.

A subprocess of choosing a finite set of precedents. According to the description of the precedent, in which some data is collected (measured), the decision is made whether or not to include the precedent in the training set. As a result, an educational sample is formed for machine learning. It is undesirable to apply ethical restrictions to the totality of empirical data, since these restrictions will distort the result of evaluating the practical performance of the algorithm. After all, the heuristic problem of machine learning consists in the maximum approximation of the conditions of the experiment to the conditions of real problems. This conclusion is valid subject to the absence of restrictions on access to data or subject to existing restrictions. However, the ethical component can have a positive impact in assessing the reliability of the data included in the description of precedents. In this case, the data of the "Garbage In, Garbage Out" format, characterizing the weak accuracy, will not fall into the training sample.

Subprocess of determining the algorithm quality functional. The quality functional shows how adequate the model is to the observed data. The degree of adequacy takes quite clear outlines in the computational learning theory (COLT) [15], in particular, by controlling the generalizing ability of the algorithm and by assessing the reliability of the algorithm. The functional quality of the algorithm generates the result of training on precedents, therefore, the method of minimizing empirical risk is relevant for the quality functional. From the standpoint of machine learning ethics, the risks of generalization by the algorithm of empirical data are manifested in the occurrence of a false correlation or feedback loop - when the algorithm decides on a pattern that does not correspond to the logic and ethics of the process. The ethical component for this subprocess contributes to limiting the ability of the algorithm to generalize, thereby enhancing protection against such undesirable manifestations of the operation of the algorithm, such as retraining and under-training. But there is another side. Ethical limitations can significantly complicate the model, and since retraining is associated with the excessive complexity of the model used, this risk should be optimized by empirically measuring the probability of retraining (Monte Carlo method) [16].

A subprocess of building algorithms of the recovered dependency. The subprocess includes the choice of algorithmic methods, the testing of the algorithm and the subsequent calibration, taking into account the changes made in the totality of precedents and requirements for the functionality of the algorithm. As you can see, at this stage there is a return to the previous subprocesses. The choice of a method or algorithmic methods is determined by the fundamental possibility and speed of interpretation of its work. The logic of the method chosen should not only explain the prediction obtained, but also demonstrate the internal dependencies in the data. Interpretation of the results of the algorithm is the most important parameter for this subprocess. It gives an understanding of the significance of the attributes of dependency in the data and becomes the basis for managing the traits in this or a more complex model.

The application of the ethical component in this subprocess can be considered as insurance against the inexpedient use of machine learning. For example, the intuition of justice can help to establish that the restored dependence is explained not by the "learning" algorithm, but by the intuitively fair rules. Or another example, the intuition of justice contributes to the refinement of parameters for the predictive model, identifying data that are poorly reliable and poorly related to the predicted value. The ethical component embedded in the subprocess influences the calibration parameters of the recoverable dependence.

The revealed interrelation between the algorithmization process in machine learning and the stages of conducting mining analysis projects is complemented by the cyclic interconnection of algorithmic subprocesses (Fig. 1).

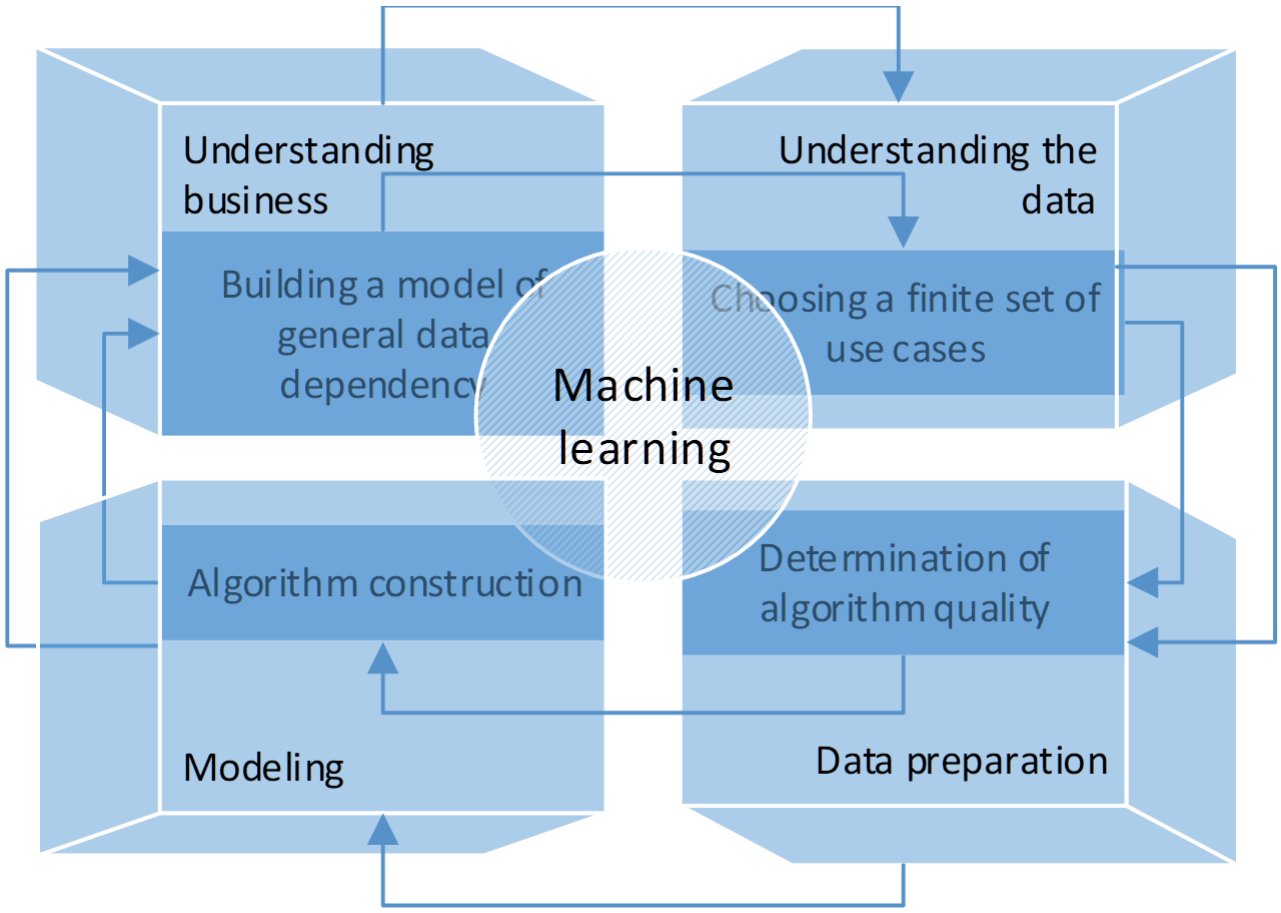

Fig. 1. Cyclic interconnection of algorithmic processes.

The cyclical nature allows you to optimize the subprocess of building algorithms by returning to the previous subprocesses and making the transition to models that are more complex. As a result of the projection of the influence of the ethical component on the algorithmicizing subprocesses, it was determined that the ethical component has the greatest influence on the input and output subprocesses of the main process - when building a data dependency model and constructing algorithms for the restored dependency.

Thus, the ethical component of the algorithmization process in machine learning is aimed at limiting the useless, potentially harmful or dangerous results of machine learning. Conventional benchmarks for such restrictions are various iterations of the laws of responsible application of AI - from the laws of robotics by A. Azimov [17] to the principles of ethics of AI from Google [18] or the BS8611 Standard of the British Standards Institute [19].

Answering the first question posed in the study on the application of the ethical component in the process of algorithmization in machine learning, you should go to the second question and determine the method of measuring ethical restrictions.

We proceed from the fact that ethical restrictions in machine learning are one of the conditions for solving a specific problem for a "learning" algorithm. This condition formalizes the mathematical model of the dilemma, creating the necessary connection between the algorithmic logic of the machine and the intuition of human justice. Algorithmic logic restricts solution options to predefined paths defined by programmers [20]. While the intuition of justice is characterized by the entire wealth of choice for making intuitively obvious decisions.

We will also proceed from the fact that the purpose of using the "learning" algorithm is to make the best decision in the future based on the examples (precedents) set today. Therefore, when determining how to measure

ethical constraints, it is advisable to proceed from the fact that both the decision-making process itself and the decisions themselves must be intuitively fair. In this position, the connection between the intuition of justice and the previously mentioned representation of the algorithm is captured, both as a computational procedure and as a function that establishes a dependency.

Summarizing the introductory for solving the second problem, we determine that the ethical constraints are ultimately aimed at building a "learning" algorithm for making decisions. However, the solution is also to build such an algorithm, and to establish ethical restrictions. Only the decision to build will be the main, and the decision on restrictions - local. Here it is necessary to refer to the provisions of the theory of restriction systems (TOC) E. Goldratt [21]. This will help us find a way to measure the significance of the ethical constraint, based on an assessment of its local influence on the system as a whole. The role of such a system is performed by a "learning algorithm".

E. Goldratt developed a five-step sequence to provide a controlled transformation of some system. The goal of this transformation is to remove restrictions for updating and maintaining the effectiveness (and effectiveness) of a previously made decision. The most interesting is working with constraints in the process of transforming the system. After the constraint is found (step 1), it is necessary to decide how best to weaken the action of the stopper (step 2). Further, the operation of the entire system is configured so that the limiting element works with maximum efficiency (step 3). Then we analyze the results of our actions: we find out whether this restriction still holds up the work of the entire system? If not, we got rid of it and proceed to the definition of a new element that restricts the operation of the system as a whole (step 5). If so, then the restriction still exists, and the bandwidth of the weak link should be increased, and the restriction completely removed (step 4).

In our case, we do not seek to get rid of the constraints. However, defining constraints as local solutions, we are looking for the answer to the question - how to measure the influence of local decisions on the system as a whole? The answer to this question is the empirical meaning of TOC. The method is to focus attention on the limiter and ignore non-limiting elements. An analogy is the scientific approach, in which the effect is measured, caused by a change in one variable with other conditionally constant variables. The specificity of the measurement method in TOC is to obtain the maximum from the limiting element by isolating the limiter and the system. Since the influence of the ethical limiter on the "learning" algorithm can also be considered in isolation, the specifics of the measurement method in TOC is fully consistent with our task.

Following the example of E. Goldratt, we will determine the parameters by which the effectiveness of each local decision (ethical limitation) is evaluated in terms of achieving the goal of the entire system (the "learning" algorithm). Thus, each ethical constraint is evaluated according to the degree of influence on a number of parameters. As parameters we considered a causal relationship, a technical requirement, a fair algorithm, a management decision (Table 2).

Table 2. Estimation of ethical restrictions.

The name of the subprocess in ML	**The name of the parameter to assess ethical restrictions**	**Security Question**	**Subprocess target**
Building a model of general data dependency	Causal relationship	Is causal relationship significant?	Identifying patterns of empirical data
Choosing a finite set of use cases	Functional requirement	Are there the functional requirements?	Intelligent Data Analysis
Determination of algorithm quality	Fair algorithm	Is it intuitively fair?	Training on many similar tasks
Building algorithms of the restored dependencies	Management decision	Is it possible to implement the solution?	Building a "learning" algorithm

Considering the logic of cyclic communication of algorithmic processes subprocesses, the specified parameters are also interrelated in accordance with the logic of the CRISP-DM methodology. The evaluation of each ethical constraint occurs consistently within each of the subprocesses. Measuring the impact of ethical constraints on a "learning" algorithm, you need to ask a security question. A positive answer to the control question confirms the materiality of the ethical limitation and the expediency of its establishment to achieve the goal of each subprocess. A positive answer is a condition for the transition of the ethical constraint from one algorithmization subprocess to another. A negative

answer for at least one subprocess does not allow one to accept an ethical constraint for the "learning" algorithm from the point of view of achieving the goal of the system as a whole. In this case, the ethical limitation should be ignored or modified by decomposition.

In order to be able to apply these parameters to measure the effect of ethical restrictions on the "learning" algorithm, it is necessary to verify the content of each of the parameters. Figure 2 presents a method for verifying the parameters for assessing ethical constraints.

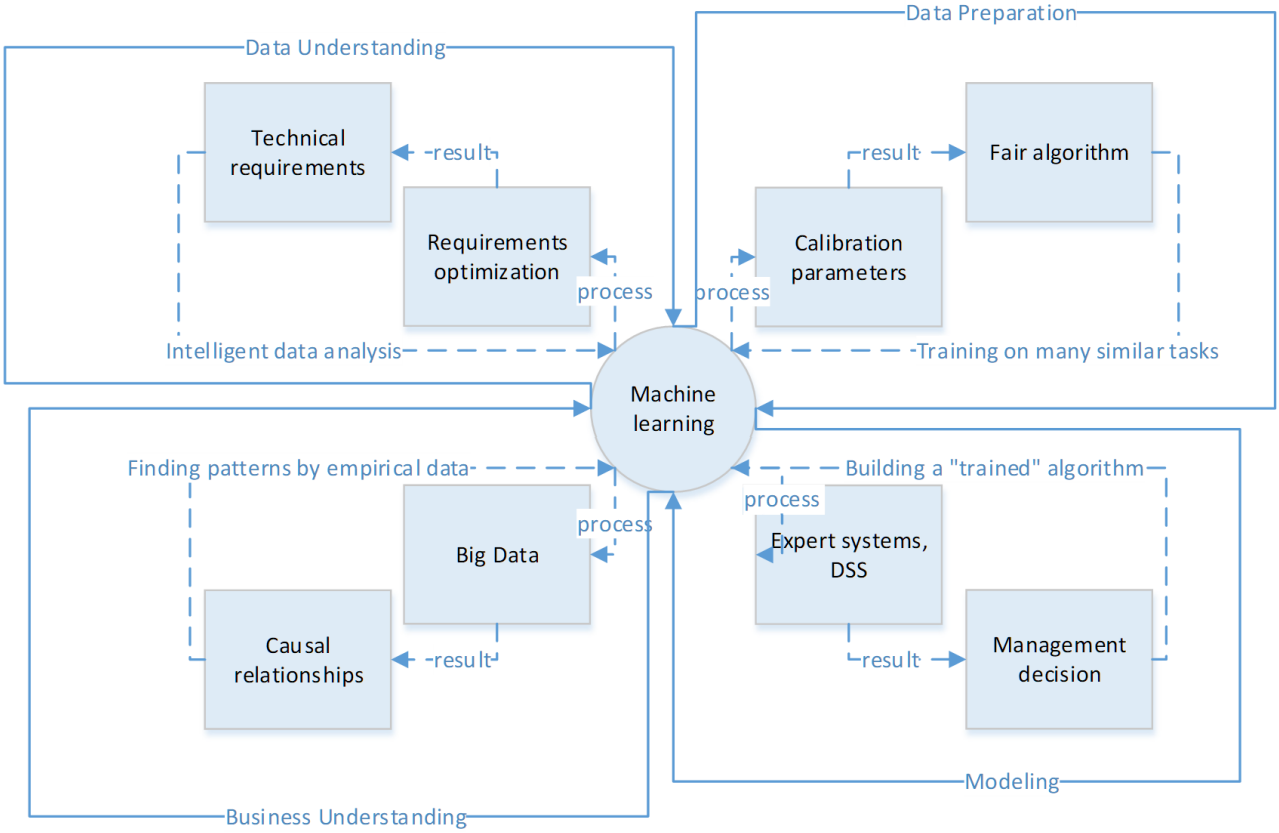

Fig. 2. Verification of parameters for evaluating ethical restrictions

For each of the algorithmic process subprocesses in Figure 2, a separate verification cycle is allocated. The cycle is constructed in PDCA notation (Plan-Do-Check-Act), the visualization of which is the Deming Cycle [22]. The cycle includes four segments: Task ML-Process-Result-Implementation. The sequence of cycles repeats the logic of the algorithmization subprocesses and the CRISP-DM methodology. Parameters are verified by comparing them with empirical data. For example, the "causal relationship" parameter is verified by the results of Big Data processing (Big Data). Depending on the task of machine learning, Big Data is structured. After the parameter is verified, it can be used to assess ethical restrictions. Thus, we answered the second question of the research on the method of measuring ethical restrictions.

4 Conclusion

In this study, an approach to establishing ethical constraints for the algorithmization process was formalized. In addition, the study answers the question of how to measure ethical limitations in the development of a "learning algorithm". The toolkit used in the study once again shows that the problem of machine learning ethics is in the interdisciplinary field of scientific knowledge. This study shows the possibilities of applying various methods of scientific knowledge, which form the basis of the Theory of Computational Learning, Theory of Limiting Systems, Methodologies of Data Mining, and the Deming Cycle. However, this is far from a complete and by no means universal list of methods for solving problems in the field of machine learning ethics.

To a greater extent, this study focuses on the problem of convergence of human ethics based on the intuition of justice and algorithmic logic. Defining the ethics of machine learning as a special case of a mathematical model of a dilemma, we avoid deviations in the field of futurology, seeing the perspective in the algorithmization of ethical constraints. This working model can be applied today, formalizing the intuition of justice for the "learning" algorithm in the yes or no answers. However, we are also optimistic about the future, when a strong AI can make intuitively fair decisions as a person.

The research was supported by the grant of the President of the Russian Federation according to the state support of leading scientific schools (grant № NSh-5449.2018.6).

References

1. J. Rawls, Philosophy and Public Affairs, *The Priority of Rights and Ideas of the Good*, **17, 4,** 251-276, (1988)
2. A. Mesoudi, P. Danielson, Theory in Biosciences, *Ethics, evolution and culture*, 127, **3**, 229-240, (2008)
3. M. Loi, M. Christen, Ercim News, *How to Include Ethics in Machine Learning Research*, **116**, 5, (2019)
4. I. Rahwan, M. Cebrian, Nature, *Machine behaviour*, **568, 7753,** 477-486, (2019)
5. E. Awad, S. Dsouza, R. Kim, Nature, *The Moral Machine experiment*, **563, 7729,** 59-68, (2018)
6. U. Kose, Brain-Broad Research in Artificial Intelligence and Neuroscience, *Are We Safe Enough in the Future of Artificial Intelligence? A Discussion on Machine Ethics and Artificial Intelligence Safety*, **9, 2,** 184-197, (2018)
7. P. Danielson, Ethics and Information Technology, *Designing a machine to learn about the ethics of robotics: the N-reasons platform*, **12, 3,** 251-261, (2010)
8. P. Danielson, Nature, *Moral Machines: Teaching Robots Right from Wrong*, **457, 7229,** 540, (2009).
9. D. Madras, T. Pitassi, R. Zemel, Nips 2018, *Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer*, **31**, 11, (2018).
10. P. Hors, M. C. Rousset, Expert Systems with Applications, *The sustainability of structural knowledge is a formal basis based on the logic of description*, **8, 3,** 371-380, (1995).
11. W. Carnielli, A. Rodrigues, WCP, *On the Philosophy and Mathematics of the Logics of Formal Inconsistency*, **152** 57-88.
12. G. K. Golubev, Problemy Peredachi Informatsii, *On a method of empirical risk minimization*, **40, 3,** 21-32, (2004)
13. K. V. Vorontsov, Computing Center RAS *Combinatorial theory of learning reliability by precedents: Dis. doc Phys.-Mat. Sciences: 05-13-17*, (2010)
14. R. Wirth, J. Hipp, *CRISP: DM Towards a Standard Process Model for Data Mining,* Retrieved from: http://citeseerx.ist.psu.edu/viewdoc/download?, doi=10.1.1.198.5133&rep=rep1&type=pdf
15. V. N. Vapnik, A. Y. Chervonenkis, Theory of Probability and its Applications, *Necessary and sufficient conditions for the uniform convergence of means to their expectations*, **26 (3),** 532-553, (1981).
16. Y. Zhou, Z. Lu, K. Cheng, W. Yun, Mechanical Systems and Signal Processing, *A Bayesian Monte Carlo-based method for efficient computation of global sensitivity indices*, **117,** 498-516, (2019).
17. S. L. Anderson, AI and Society, *Asimov's "Three Laws of Robotics" and machine metaethics*, **22, 4,** 477-493, (2008).
18. *Perspectives on Issues in AI Governance - Google AI*, Retrieved from: https://ai.google/static/documents/perspectives-on-issues-in-ai-governance.pdf
19. *British Standard BS 8611:2016, ethical design and application of robots*, Retrieved from: http://www.machinebuilding.net/ta/t1028.htm
20. S. A. Applin, M. D. Fischer, Istas, *New Technologies and Mixed-Use Convergence How Humans and Algorithms are Adapting to Each Other*, (2015).
21. H. William Dettmer, Milwaukee, *Goldratt's Theory of Constraints: A Systems Approach to Continuous Improvement*, (1997).
22. P. Arveson, The Deming Cycle, Retrieved from: https://www.balancedscorecard.org/BSC-Basics/Articles-Videos/The-Deming-Cycle
23. Pthe application: 01.06.2019