REVISTA SAN GREGORIO

http://dx.doi.org/10.36097/rsan.v1i41.1496

Artículo Original

Theoretical and Legal Bases of Artificial Intelligence Punishment System Development

Bases teóricas y legales del desarrollo de sistemas de castigo de inteligencia artificial

Authors

Ramil Rustamovich Gaifutdinov[1], Zarina Ilduzovna Khisamova[2], Elina Leonidovna Sidorenko[3], Marina Aleksandrovna Efremova[4], Tatyana Mikhaylovna Lopatina[5], Danila Vladimirovich Kirpichnikov[6]

[1]Faculty of Law, Department of Criminal Law, Kazan Federal University, Kazan, Russia, kafedra.ksu@yandex.ru

[2]Department of Planning and Coordination of Research Activities, Head, Research Department Krasnodar University of the Ministry of Internal Affairs of the Russian Federation, Krasnodar, Russia (Russian Federation), alise89@inbox.ru

[3]Department of Criminal Law, Criminal Procedure and Criminalistics, Moscow State Institute of International Relations (University), Moscow, Russia (Russian Federation), 12011979@list.ru

[4]Department of Criminal Law Disciplines, Kazan branch of the Russian State University of Justice, Kazan, Russia (Russian Federation), crimlaw16@gmail.com

[5]Department of Criminal Law, Criminal Procedure, Smolensk State University, Smolensk, Russia (Russian Federation), rectorat@smolgu.ru

[6]Department of Criminal Law and Procedure, Kazan Innovative University named after V.G. Timiryasov, Kazan, Russia (Russian Federation), danila667@outlook.com

Fecha de recibido: 2020-10-14
Fecha de aceptado para publicación: 2020-11-18
Fecha de publicación: 2020-11-20

Abstract

The article discusses the problematic aspects of artificial intelligence technology application. The author's classification of artificial intelligence types is proposed, depending on their material expression (artificial intelligence unit, artificial intelligence carrier and artificial intelligence). The study allowed us to form the theoretical and legal foundations of artificial intelligence punishment system development. The authors proposed the types of artificial intelligence punishments and described their essence. To such types of punishments, they include complete deactivation (destruction) of artificial intelligence; partial deactivation of artificial intelligence; confiscation of artificial intelligence in favour of the state; the prohibition of artificial intelligence use in certain fields of activity; the implementation of gratuitous activities by artificial intelligence in the interests of the state or society; provision of a convict status (unreliable) for artificial intelligence. They noted that in the current conditions of the digital economy development and the pace of digital technology introduction, the international community needs to develop an effective punishment system of artificial intelligence that has committed illegal acts.

Keywords: artificial intelligence, classification of artificial intelligence types, punishment system, artificial intelligence punishment types, criminal liability, criminal law.

Resumen
El artículo analiza los aspectos problemáticos de la aplicación de la tecnología de inteligencia artificial. Se propone la clasificación del autor de los tipos de inteligencia artificial, en función de su expresión material (unidad de inteligencia artificial, portador de inteligencia artificial e inteligencia artificial). El estudio nos permitió formar los fundamentos teóricos y legales del desarrollo del sistema de castigo por inteligencia artificial. Los autores propusieron los tipos de castigos de inteligencia artificial y describieron su esencia. Para este tipo de castigos, incluyen la desactivación (destrucción) completa de la inteligencia artificial; desactivación parcial de la inteligencia artificial; confiscación de inteligencia artificial a favor del estado; la prohibición del uso de inteligencia artificial en determinados campos de actividad; la implementación de actividades gratuitas por inteligencia artificial en interés del Estado o la sociedad; provisión de un estado de convicto (no confiable) para inteligencia artificial. Señalaron que en las condiciones actuales del desarrollo de la economía digital y el ritmo de introducción de la tecnología digital, la comunidad internacional necesita desarrollar un sistema de castigo efectivo de inteligencia artificial que haya cometido actos ilegales.

Palabras clave: inteligencia artificial, clasificación de tipos de inteligencia artificial, sistema de castigo, tipos de castigo de inteligencia artificial, responsabilidad penal, derecho penal.

Introduction

Uncontrolled technology for Artificial Intelligence could do more harm than good without human rights and ethics. During RightsCon 2018, UNESCO held an interactive session focusing on existing debates on developing and applying big data and AI technologies. The participants emphasized the urgent need for these emerging technologies to develop human rights and ethical standards. Uncontrolled application of artificial intelligence technology (hereinafter - AI) creates certain threats to the safe development of society and the state, determined by the lack of adapted technologies, in accordance with the AI specifics, criminal law mechanisms for public relation protection, the presence of legal uncertainty concerning the issue of criminal liability subject in case of direct damage by AI technology (Z. I. Khisamova & Begishev, 2019a). Artificial intelligence and other new digital technologies, such as the Internet of Things or distributed ledger technologies, have the ability to bring about a better transformation of our communities and economies. However, in order to reduce the risk of damage, these technologies can cause, such as bodily injury or other damage, their deployment must come with appropriate safeguards. A significant number of works have been devoted to the study of these issues, as well as the aspects of the public relation legal regulation arising from AI use (Alzou'bi et al., 2014; I. R. Begishev, 2020; I. R. Begishev et al., n.d.; Ildar R. Begishev et al., 2019, 2020; Ildar R. Begishev & Khisamova, 2018; Bikeev et al., 2019; Bokovnya et al., 2019; Cameron, 1990; Z. Khisamova et al., 2019; Z. I. Khisamova & Begishev, 2019b; Zarina I. Khisamova et al., 2019; Latypova et al., 2019; Rademacher, 2020; Shestak et al., 2019; Shestak & Volevodz, 2019; Simmler & Markwalder, 2019; Sukhodolov et al., 2020; Sukhodolov & Bychkova, 2018).

Objective

Based on their content expression (unit of artificial intelligence, artificial intelligence carrier and artificial intelligence), the author's classification of artificial intelligence types is proposed. The writers suggested the types of punishments for artificial intelligence and defined their meaning.

Material and Methods

The materials for work were the articles posted in scientific journals and on websites.
The methodological basis of the study is the combination of scientific knowledge methods, including abstract logical, comparison and correlation analysis.

Results and Discussion

In order to build a system of artificial intelligence (hereinafter - AI) punishment, it is necessary to touch upon such an important aspect as its material expression. Not only the way and mechanism of punishment application to it but also their types, as well as their measure, directly depends on the way the AI is represented in external reality. It should be noted that a similar problem does not arise in relation to a person, since their harming always occurs by affecting the phenomena and the objects of the world.

In the affected context, the human body is its material expression, which is quite obvious and does not cause discussions. Regarding AI, there is no such certainty and consistency. It can also manifest itself in autonomous, self-driving cars (as

in the situation with a car accident under the control of the Uber AI company, which caused death); as an integral part of a certain subject that is not capable of functioning outside its intended purpose prescribed by AI (for example, an unmanned aerial vehicle); or it may exist only in cybernetic reality without material expression. The definition of AI punishments is made dependent on its material expression.

We offer the following classification of AI types, depending on their material expression:
AI unit - an object of the material world, which is an integral part of AI, created to ensure the AI potential implementation and capable of influencing the surrounding reality through mechanical movements. This item does not represent value when AI is removed from it since it is directly intended for the implementation of its functions.

An AI carrier is an object of the material world that contains AI technology but is incapable of fulfilling its functions and mechanical actions. Its main purpose is to keep AI technology on itself.
An AI program is a cybernetic creation that does not have a material expression and exists exclusively in the digital space. According to its qualities, it is capable of carrying out actions involving criminal law consequences. Moreover, the AI program can be used in modern digital networks.
Based on the foregoing, we can conclude that it is necessary to develop a theoretical and legal approach to AI punishment system development, depending on its material expression.

As for the system of AI punishment, the authors proceed from the fact that it represents the application of coercive measures against the subject in its epistemological essence with the aim of inducing him to lawful behaviour, forming his conviction of the impossibility or inappropriateness of criminal law prohibition violations, eliminating the antisocial attitude, and creating respect for the law. Besides, the punishment implements general and private criminal law prevention and protects society from the entities that pose a threat of harm to public relations protected by criminal law.

The punishment contains two elements: legal - the deprivation or restriction of the right and freedom, which consists in the person's ability elimination to realize a certain type and measure of possible behaviour (scope of actions), which follows from the alienated right or freedom, which is used to restore social justice, correct a convicted person, and to prevent the commission of new crimes; and actual - the creation of conditions, factual circumstances in which the exercise of this right is impossible.

The following is fundamental - based on the court position, punishment is deprivation or restriction of a right or freedom, from which a certain type and measure of a person possible behaviour follows guaranteed by law, but alienated from him as a measure of state coercion, which is expressed in creating conditions that exclude the possibility of exercising the right, that is, the ability to act in a certain way.

The abovementioned clearly implies the possibility of AI punishment application, since the only difficulty may be the creation of conditions (that is, the actual element), and not the fundamental impossibility of criminal punishment application to AI as a legal model, which, on the contrary, seems reasonable and is quite possible from a legal point of view.

Accordingly, the use of AI punishments in any of its material manifestations seems motivated and expedient, and they are exclusively practical, and their overcoming is dependent on the efforts made and does not have fundamentally unremovable obstacles.

However, we believe that the full comparison of punishment mechanism for a man and AI would be wrong. If in the first case, the main goal is to correct the convict, then in the second case the main goal is to ensure the safety of society, the individual and the state from the illegal activities of AI, and prevent the commission of crimes. We do not exclude that some of the proposed types of punishments can create a conviction in AI about the inadmissibility of illegal behaviour. Along with this, there is no sufficient reason to count on this result unambiguously. Moreover, in relation to AI, the activity on the formulation of punishment types is not limited to those humanistic principles and norms that are unconditional and internationally recognized for humans, which is a certain part predetermines the potential prevalence in those types of punishments application that consist in AI destruction, or in a significant restriction of its functionality.

In Russian criminal law, the purpose of criminal punishment is to restore social justice and correct the convicted person, as well as to prevent the commission of new crimes. There are reasonable doubts about the effectiveness of traditional types of criminal punishment application to AI, such as a fine or imprisonment for the purpose of an "AI criminal" change (Z. I. Khisamova & Begishev, 2019a).

The legal literature presents several positions regarding the definition of punishment list. Kopfstein (2017) proposes, in particular, the following list:

- deactivation;
- reprogramming;
- granting the status of a criminal (which, from his point of view, will have a preventive effect in relation to other participants of the legal relationship).

Evaluating this list comprehensively, we conclude that it is not without drawbacks, which boil down to the following: when formulating the types of punishments, the author does not explain their nature, order and conditions of use, which forms some incompleteness of presentation and ambiguity in understanding the terms used. In particular, the question arises about the content of deactivation as a process: does it involve the physical destruction of the material component of AI? Besides, it seems that the list provided by the author is not capable of ensuring the individualization of punishment and its compliance with the nature and degree of public danger of the crime committed, since the first two types of punishment presuppose the termination of the AI in the form in which it existed before the forced impact. Then the third type in its punitive potential is rather similar to another measure of a criminally legal nature.

Software, for example, unmanned vehicles or devices for automatic exchange trading, does not have self-awareness or a proper degree of self-awareness, which excludes subjective wrongfulness, as well as the achievement of punishment goals. In such cases, it is more efficient to reprogram or replace the device without using a criminal law mechanism. Actually, criminal punishment application for such devices can only achieve the goal of crime commission prevention, and to restore social justice to a lesser extent and the correction of the convicted person it is extremely unlikely (Mosechkin, 2019).
F.V. Uzhov takes a somewhat categorical position in relation to the prospects of AI punishment (Uzhov, 2017), which indicates that the "re-education" of AI can only be implemented by its complete reprogramming, which, according to the scientist, can be "compared with lobotomy in relation to a person," that is, the author gives an additional explanation - absolute and irreversible change the properties of AI. The second way, in his opinion, is the disposal of the machine (in this case, the content of the coercive effect does not change per se. Complete destruction is supposed in both proposed variants). The author's attempt to formulate the actual conditions for the fairness of the punishment application to AI, to which he refers is of interest: the value of the machine, the number of "mistakes" made by it, the ability to eliminate these shortcomings by technical means (Uzhov, 2017). Nevertheless, the possibility of the given criteria application is doubtful, since the author has not explained by what types of punishment an alternative is formed in principle: they are only offered destruction per se.

G. Hallevy rightly asks the question of criminal liability applicability to the AI that committed the crime and, thus, of criminal punishment goal achievement (Hallevy, 2010). G. Hallevy's reasoning seems more constructive, which is logically justified, although it is of a theoretical nature. Scientists are particularly focused on the fact that the existing concepts of criminal punishment require doctrinal review, and the existing obstacles of punishment application to AI are completely removable and practical (Hallevy, 2015).

Conclusion

Conventional wisdom holds that punishing AI is incompatible with universal concepts of criminal law, such as the right to be guilty and the need for a guilty mind. We prove that AI punishment cannot be categorically ruled out with fast theoretical arguments based on analogies to corporate and strict criminal responsibility, as well as familiar concepts of imputation. AI punishment can result in general deterrence and expressive advantages, and negative constraints such as punishment in excess of guilt do not need to run afoul. The following types of AI punishments are offered; their essence is described:

1. complete deactivation (destruction) of AI;
2. partial deactivation of AI;
3. confiscation of AI in favour of the state;
4. The ban on AI application in certain areas of activity;
5. the implementation of AI gratuitous activities in the interests of the state or society;
6. giving AI the status of a convicted (unreliable).

Thus, in the current conditions of the digital economy development, in order to ensure the effectiveness of public relation protection from unlawful attacks by AI, the world community needs to develop theoretical and legal foundations for AI punishment system development.
Based on the foregoing, the punishment system of the AI seems possible to determine in the following edition:

Complete deactivation (destruction) of AI is a forced impact on the neural connections of AI. Thus, they completely lose their properties. The specified type of punishment is the most repressive.

It is believed that it is necessary to use complete deactivation (destruction) as the only form of punishment in the case of an AI committing a grave and especially grave crime.

The next type of punishment in terms of repressive potential is the partial deactivation of AI - transfer to the category of weaker AI, by its potential intellectual reduction. The specified type of punishment, in our opinion, is advisable to implement by forced impact onto AI neural connections. Thus, the ability to process a certain amount of information from the outside world can be eliminated. At the same time, part of the information necessary for the implementation of socially significant functions will be available, but clearly insufficient for conscious-willful autonomous behaviour.

We believe that when they resolve the issue of this type of punishment application, courts should also establish the economic feasibility of continuing the functioning of AI. It seems possible to match the specified type of punishment with a specific probationary period during which the activities of AI will be constantly monitored by the supervisory authority.

Confiscation of AI in favour of the state is a forced withdrawal of AI from the owner who disposes of it and turning it in favour of the state in order to perform useful public-state functions.

The specified type of punishment is even more dependent on the economic feasibility of AI functioning maintenance than the previous one. However, in view of the AI possible willingness to commit crimes, or the incomplete elimination of the antisocial attitude, it is advisable to use it in the event of a crime commission of small or medium gravity for the first time, or due to an accidental combination of circumstances.

We also consider it is necessary to take into account that at the present stage of development, each AI represents significant scientific, industrial, defence, and equally different social or state value. These circumstances caused us to deduce this type of punishment for AI, as to confiscation in favour of the state. We are convinced that it would be unreasonable to destroy AI completely, or reduce its properties significantly when AI committed a crime that does not have a sufficient degree of public danger; it is quite possible to achieve the goals of punishment in the ways that do not involve causing harm to AI.

Next, let's consider the prohibition of AI use in certain areas of activity, which, in our opinion, may look like the deprivation of the right to occupy certain positions or engage in certain activities as a punishment applied to a person. The specified punishment is advisable to apply only in cases where the AI committed an offence in certain types of activities (for example, it carried out the incorrect diagnosis of patients, violated the road rules, etc.). At the same time, the need arises to fix these offences and to develop a database to make a decision on the prohibition of functioning in certain areas.

The implementation of AI gratuitous activities in the interests of the state or society is the forced implementation of community service by AI. The specified type of punishment is directed to a greater extent as the means of social justice material restoration and can be used as an additional type of punishment. In other words, the main value of this type of punishment is that in the case of an AI crime of small or medium gravity causing damage to property, the AI is given the opportunity to compensate for the loss by its own labour. At the same time, a balance is maintained between the elimination of the sentence execution cost and the compensation of the damage caused by the AI activities. To clarify the essence of this type of punishment, it is quite possible to compare it with compulsory work. The difference, however, lies in the fact that a person carries out his own re-education when performing the latter, and this is the punishment goal. In the case of AI, the focus is on compensation for caused material damage.

Giving AI the status of a convicted (unreliable). The specified negative legal consequence cannot be related to the types of punishment with sufficient justification. It can be attributed to other measures of a criminal-legal nature, because of small repressive potential. This measure consists in obligatory informing the counterparty of the AI about the fact of an AI crime commission, which will inevitably entail reputation losses for the AI and, in our opinion, will have a tangible deterrent, especially in the field of commercial activity, thereby fully implementing general and private prevention. The indicated measure seems reasonable and expedient for use only in the case of a minor offence.

Acknowledgements

The work is performed according to the Russian Government Program of Competitive Growth of Kazan Federal University.

References

Alzou'bi, S., Alshibl, H., & Al-Ma'aitah, M. (2014). Artificial intelligence in law enforcement, a review. *International*

Journal of Advanced Information Technology, *4*(4), 1.

Begishev, I. R. (2020). Organization of the hacker community: Criminological and criminal law aspects. *All-Russian Criminological Journal*, *14*(1), 96–105.

Begishev, I. R., Khisamova, Z. I., & Mazitova, G. I. (n.d.). Information infrastructure of safe computer attacks, Helix, 2019, Vol. 9, No. 5. *DOI*, *10*, 2019–5639.

Begishev, Ildar R., & Khisamova, Z. I. (2018). Criminological Risks of Using Artificial Intelligence. *RUSSIAN JOURNAL OF CRIMINOLOGY*, *12*(6), 767–775.

Begishev, Ildar R., Khisamova, Z. I., & Mazitova, G. I. (2019). Criminal legal ensuring of security of critical information infrastructure of the Russian Federation. *Revista Gênero & Direito*, *8*(6), 283–292.

Begishev, Ildar R., Latypova, E. Y., & Kirpichnikov, D. V. (2020). Artificial Intelligence as a Legal Category: Doctrinal Approach to Formulating a Definition. *Actual Probs. Econ. & L.*, 79.

Bikeev, I., Kabanov, P., Begishev, I., & Khisamova, Z. (2019). Criminological risks and legal aspects of artificial intelligence implementation. *Proceedings of the International Conference on Artificial Intelligence, Information Processing and Cloud Computing*, 1–7.

Bokovnya, A. Y., Khisamova, Z. I., & Begishev, I. R. (2019). Study of Russian and the UK Legislations in Combating Digital Crimes. *Helix*, *9*(5), 5458–5461.

Cameron, J. (1990). Artificial intelligence, expert systems, microcomputers and law enforcement. *The Police Chief*, *57*(3), 36–41.

Hallevy, G. (2010). The criminal liability of artificial intelligence entities-from science fiction to legal social control. *Akron Intell. Prop. J.*, *4*, 171.

Hallevy, G. (2015). *Liability for crimes involving artificial intelligence systems*. Springer.

Khisamova, Z., Begishev, I., & Gaifutdinov, R. (2019). On methods to legal regulation of artificial intelligence in the world. *SCOPUS-2019-9-1-SID85075304864*.

Khisamova, Z. I., & Begishev, I. R. (2019a). Criminal liability and artificial intelligence: Theoretical and applied aspects. *All-Russian Journal of Criminology*, *13*(4), 574.

Khisamova, Z. I., & Begishev, I. R. (2019b). Legal regulation of artificial intelligence. *Baikal Research Journal*, *10*(2).

Khisamova, Zarina I., Begishev, I. R., & Sidorenko, E. L. (2019). Artificial Intelligence and Problems of Ensuring Cyber Security. *International Journal of Cyber Criminology*, *13*(2), 564–577.

Kopfstein, J. (2017). *Should Robots Be Punished for Committing Crimes? Vocativ Website*.

Latypova, E. Y., Nechaeva, E. V., Gilmanov, E. M., & Aleksandrova, N. V. (2019). Infringements on Digital Information: Modern State of the Problem. *SHS Web of Conferences*, *62*, 10004.

Mosechkin, I. N. (2019). Artificial Intelligence and Criminal Liability: Problems of Becoming a New Type of Crime Subject. *Vestnik Saint Petersburg UL*, 461.

Rademacher, T. (2020). Artificial intelligence and law enforcement. In *Regulating artificial intelligence* (pp. 225–254). Springer.

Shestak, V. A., & Volevodz, A. G. (2019). Modern requirements of the legal support of artificial intelligence: A view from Russia. *Russian Journal of Criminology*, *13*(2), 197–206.

Shestak, V. A., Volevodz, A. G., & Alizade, V. A. (2019). On the possibility of doctrinal perception of artificial intelligence as the subject of crime in the system of common law: Using the example of the US criminal legislation. *Russian Journal of Criminology*, *13*(4), 547–554.

Simmler, M., & Markwalder, N. (2019). Guilty Robots?–Rethinking the Nature of Culpability and Legal Personhood in an Age of Artificial Intelligence. *Criminal Law Forum*, *30*(1), 1–31.

Sukhodolov, A. P., Bychkov, A. V., & Bychkova, A. M. (2020). *Criminal Policy for Crimes Committed Using Artificial Intelligence Technologies: State, Problems, Prospects*.

Sukhodolov, A. P., & Bychkova, A. M. (2018). Artificial intelligence in crime counteraction, prediction, prevention and evolution. *RUSSIAN JOURNAL OF CRIMINOLOGY*, *12*(6), 753–766.

Uzhov, F. V. (2017). Artificial intelligence as a subject of law. *Probely v Rossiiskom Zakonodatel'stve*, *3*, 357–360.