UDC 004.9
https://doi.org/10.31891/CSIT-2022-1-11

EDUARD MANZIUK
Khmelnytskyi National University

INTELLIGENT INFORMATION TECHNOLOGY FOR OBTAINING TRUST DECISIONS BASED ON THE ONTOLOGY OF TRUST IN A HUMAN-CENTERED APPROACH

The paper presents the results of research on the development of intelligent information technology for obtaining trust decisions to determine the constituent elements and the interaction between them, which together provide for obtaining trust decisions. Ethical principles based on human-centered approach have been identified, which formed the basis for the development of a set of methods for the practical implementation of certain principles.

The set of developed methods to ensure certain ethical principles of trust in decisions obtained using intelligent information systems has allowed developing intelligent information technology for obtaining trust decisions. Confidence in decisions is formed by ensuring the practical implementation of ethical principles in the methods of processing input information. Information technology determines the interaction of the developed methods, which together form the trust in the decisions of the information system. The structure of formation of intelligent information technologies for obtaining trust decisions is presented, as well as the structure of interaction of components of intelligent information technologies for obtaining trust decisions. A description of ethical principles with the definition of their implementation in the structure of a set of methods is given. The set of functions which realization allows reaching application of the offered intellectual information technology is defined.

The components of the concept of trust were considered within the concept of intelligent information technology. The practical application of intelligent information technology is generally seen as a complex system of interconnected structural components to solve practical problems with building ecosystems for intelligent information technology, ensuring trust at all stages of the life cycle of intelligent information technology with a chain of trust and continuity.

Keywords: intelligent information technology, ethical principles of trust, difficult to classify information.

ЕДУАРД МАНЗЮК
Хмельницький національний університет

ІНТЕЛЕКТУАЛЬНА ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ОТРИМАННЯ ДОВІРЧИХ РІШЕНЬ НА ОСНОВІ ОНТОЛОГІЇ ДОВІРИ ЗА ЛЮДИНОЦЕНТРОВАНИМ ПІДХОДОМ

В досліджені представлено результати досліджень з розробки інтелектуальних інформаційних технологій для отримання довірчих рішень з визначення складових елементів та взаємодії між ними, які в сукупності забезпечують отримання довірчих рішень. Визначено етичні принципи, засновані на людиноцентрованому підході, які лягли в основу розробки комплексу методів практичної реалізації визначених принципів.

Сукупність запропонованих методів забезпечення визначення етичних принципів довіри до рішень, отриманих за допомогою інтелектуальної інформаційної системи, дозволив розробити інтелектуальну інформаційну технологію для отримання довірчих рішень. Довіра до рішень формується шляхом забезпечення практичної реалізації визначених етичних принципів у методах обробки вхідної інформації. Інформаційна технологія визначає взаємодію розроблених методів, які в сукупності формують довіру до рішень інформаційної системи. Представлено структуру формування інтелектуальної інформаційної технології для отримання довірчих рішень, а також структуру взаємодії компонентів інтелектуальної інформаційної технології для отримання довірчих рішень. Подано опис етичних принципів з визначенням їхньої реалізації у структурі сукупності методів. Визначено набір функцій, які досягаються шляхом застосування запропонованої інтелектуальної інформаційної технології. У рамках концепції інтелектуальної інформаційної технології було розглянуто складові поняття довіри. Практичне застосування інтелектуальної інформаційної технології загалом розглядається як складна система взаємопов'язаних структурних компонентів для вирішення практичних завдань з побудовою системи для інтелектуальної інформаційної технології забезпечення довіри на всіх етапах життєвого циклу інтелектуальної інформаційної технології з неперервністю у просторі і часі.

Ключові слова: інтелектуальна інформаційна технологія, етичні принципи довіри, складнокласифікована інформація.

Introduction

The rapid development and widespread use of information technology are due to the prospects for the practical application of computer systems. In particular, the introduction of intelligent information technology and intelligent information systems has greatly expanded human capabilities for decision-making both through the solution of complex computational problems and through autonomous decision-making. The development of intelligent information technology reaches a level at which it can be delegated tasks for decision-making and with further development creating conditions for the formation of subjectivity in decision-making. Examples of such applications are automated trading systems, disease diagnosis systems, unmanned and autopilot systems, aircraft control systems for various purposes, including for agriculture or military use, object recognition systems for tracking people or identifying the enemy, and so on.

However, giving intelligent information systems the ability to make autonomous decisions creates a problem of responsibility for decision-making. Unfortunately, technical systems can have technical and software

failures in their operation due to various reasons. An incorrect decision made by the car's autopilot is a threat to the lives of passengers, and in the case of socially critical decisions, incorrect decisions have even more significant consequences. There is a problem of safe integration of intelligent information technologies in society.

Thus, the level of development of intelligent information technologies and the benefits of wide application lead to the emergence of new challenges that arise in the formation of a new socio-technological system.

Related works

The components of the concept of trust were considered within the concept of intelligent information technology (IIT). The practical application of IIT in general is considered as a complex system of interconnected structural components to solve practical problems with building ecosystems of IIT [1], ensuring trust at all stages of the life cycle of IIT [2-6] with providing a chain of trust and continuity over time [7 –11].

Let's determine the level of confidence in the following structural components of IIT systems and systems that use elements of IIT:

1. The level of elements of information, which are sensors, components of the Internet of Things (IoT), databases, files.

2. The level of infrastructure for building IIT, including storage, transmission and processing of information.

3. The level of application software, which aims to ensure:

- functioning of IIT;
- security and non-interference;
- framework of software implementation of IIT;
- implementation of systems and algorithms of intelligent data processing (mechanisms of intelligent information systems).

One of the key system components of IIT, which is the basis of IIT and directly processes and generates application solutions, is the level of application software systems for the implementation of intelligent information system (IIS) algorithms.

The purpose of the study is development of an intelligent information system for obtaining trust decisions based on the definition of a set of ethical principles according to the human-centered approach.

Proposed technique

The projection of trust in IIT in the plane of realization of data processing mechanisms by IIS methods is considered.

The concept of “mitigation measures” ($MtMs$), which builds trust in the IIT in accordance with the developed trust ontology O, is presented at the level of subcategories in the form $MtMs = \{mtms_i\}_{i=1}^{n}$, $n = 14$, where n is a number of phrasal subcategories of the concept that determine the components of the trust concept.

The concept $MtMs$ is unambiguously and completely defined within the developed ontology of trust and presented as $MtMs_O$. Intelligent information system is a component of IIT.

Accordingly, the projection of the concepts of ensuring trust in the IIS plane is presented as follows

$$\phi_{pr} : \left(MtMs_O\right) \mapsto \left(MtMs_{IIS}\right) \; s.t. \; MtMs_{IIS} \subset MtMs_O.$$

The set of components of trust in the IIS $MtMs_{IIS}$ is defined as the allocation of a subset $MtMs_{IIS} \subset MtMs_O$. The subset is selected by the projection function of the base set $MtMs_O$. The set of components of trust $MtMs_{IIS}$ is a more compact form, which allows forming a set of components of the concept of trust. Thus, IIS mechanisms must ensure implement and meet the set of trust components $MtMs_{IIS}$. This state corresponds to the trust in the decisions formed by the means of IIS.

IIS is generally an integral part of IIT, which in the embodiment of generalized sets of aspects of the representation can be represented as $IIS \subset IIT$. Accordingly, the formation of components of trust in the projection of IIS is carried out with a description of approaches that allow ensuring the practical implementation and realization of these components by building specific mechanisms of IIS.

As a result of the practical implementation of the developed method of establishing the correspondence of the trust ontology to the IIT and the structured domain, the importance of each component of trust was determined. The importance of the components is determined by the weight fraction of the elements of trust in the total population of a certain set of trust.

Each element is presented, in addition to the description of content, as well as a mechanism of practical implementation at the level of implementation of structural, architectural and methodological solutions.

The projection of the components of trust at the level of IIS is based on the human-centered concept [12], the basic element of the relationship of trust in accordance with ISO / IEC TR 24028 is the stakeholder.

Accordingly, at the applied level of IIS methods, the constituent elements of trust are defined as follows:

- Interpretability, explainability, transparency, traceability. The basic generalizing concept is transparency. Transparency is defined in relation to the stakeholder as the ability of the system to be understood and explained both by the results of the decisions made and by the steps of decision-making. Accordingly, the IIS method of decision-making is divided into successive steps of implementation with the requirement of interpretability and transparency of their implementation in relation to the stakeholder. Interpretability, explainability are key components of transparency [13]. Let's define interpretability as the internal state of the system that corresponds to the ability to be understood by the stakeholder. Let's define interpretability as the external state of the system that corresponds to the ability to be understood by the stakeholder. Decision-making in IIS is carried out by a model that is built and validated accordingly. The interpretability of the model is determined by the interpretability of the IIT algorithms used to build it [13, 14]. Thus, the interpretability of the model is determined by the level of interpretability of the algorithm used to build it. For IIS systems, we define three levels of definition of interpretability, explainability and transparency: premodel, model, postmodel. The pre-model level is determined by the stages that precede the construction of the working model. The model level corresponds to the stage of the constructed model. The postmodel level corresponds to the stage of the generated model decision. At the level of developed methods, the components are implemented through the use of group decision-making in the form of an ensemble of models. The consolidated decisions of the ensemble are interpreted according to the model that has the maximum interpretation in relation to the models of the ensemble.

- Comprehensiveness of consideration. Decision-making has a high level of responsibility, so every decision must be made objectively and taking into account various aspects. For each decision, there must be a certain level of certainty as to its infallibility and the impossibility of causing harm. Decisions must also be fair and unbias. The unbias is characterized by an unbiased assessment of the baseline value in the data. To ensure the practical implementation of the components of trust, groups of models are used, in which each model makes its own independent decision. Models are selected for the group according to the quality indicators of model decisions and the availability of alternative decisions from a certain data group. The availability of alternative decisions is an indicator of the model and allows including models in the ensemble, providing the maximum variety of decisions to implement the comprehensiveness of the review. The value of each model is determined by the availability of alternative decisions. This makes it impossible to form groups of models with exactly the same decisions across the data set. Based on this, the data is segmented according to the consolidated decisions of the group of models and the availability of alternative decisions. Accordingly, the set of data on consolidated decisions is a set of trusty decisions based on the comprehensiveness of decision-making. Also, multiple data are formed concerning which there are models that have alternative decisions to the decisions of group models. Such a set of data cannot be considered credible in relation to the decisions of the model group. Thus, the problem of binary classification is formed in relation to the affiliation of data to the set of trust decisions and the set of distrust decisions. On the set of distrustful decisions for each unit of data there is a set of alternative decisions. To achieve qualitative indicators of decisions, i.e. the characteristics of classification, within a unit of data decisions are grouped by the parameter of the probability of belonging to a class within binary groups by decisions. For each group of model decisions, a general group decision is formed, which is transferred to the next level of the hierarchy, where it is combined with the decision of another group. Thus the general decision of an ensemble of models in a segment of the data of alternative decisions is formed. The study considers the form of binary base groups by decisions, which is presented in the form of alternatives. Accordingly, the level of the structural hierarchy is formed of three layers with two corresponding transitions between layers. The upper level is the ultimate decision of the ensemble of models.

Segmentation of data by decisions of groups of models, allows forming a set of typical data on the basis of consolidated decisions. This set is formed by data which in relation to other data of this set are typical on criterion of the generality of decisions of models. That is, the data on the set of features of the data on the decisions of the models are weakly different within this formed set. A set of atypical data is also formed. Each element of this set has at least one alternative decision to the decisions of other models. The number and distribution of alternative decisions is not regulated. This set of data is atypical because each data element in relation to other elements of this set has significant differences in a certain set of features according to the criterion of consolidation of decisions. Significant differences are determined by the fact that the decisions of a group of models do not match, and therefore at least one model has identified in certain units of data a certain aspect that was not taken into account by other models. That is, in general, decisions about a unit of data are heterogeneous and ambiguous in their entirety, which calls into question each of the decisions of the models regarding this data element. The data element has features whose magnitude within the ensemble of models allows forming a state of alternative decisions, and therefore is atypical. Combining such data forms a set of atypical data.

- Analysis of information. Within the study, the analysis of information is determined by establishing the level of influence of data characteristics on the delimitation of difficult to classify data, which are formed by a set of atypical data. In the area of atypical data, the boundaries of the boundary division of classes by classification methods are formed. Using a group approach to the formation of the decision of the ensemble of models, the set of

atypical data allows creating a certain zone of uncertainty, which is due to the availability of alternatives in making decisions about this data. Each data element, among others, which forms a set of atypical data, has a certain feature or set of them, which, depending on the degree of influence, are taken into account by a particular model or several models from the group of ensemble models. Detection of a data element by the criterion of an alternative decision can be both true and false according to a certain decision of a particular model. However, in general, the models compensate for possible shortcomings in the construction of other models and allow forming the most objective decision in the case of consolidation. If there are alternative decisions, each data element has the value of a feature or set of them, which according to the consolidated decision of the ensemble does not correspond to the set of typical data. Atypical data are not trusted data. Such data cannot be used for training and modeling, especially as models are sensitive to outlier. However, such data are important and valuable in terms of research and analysis. This is due to the fact that these data in some cases may indicate the emergence of new processes and phenomena. That is, they indicate the early stages of these phenomena, in which models cannot reliably classify them. However, the whole set of ensemble models can detect this data element by the criterion of alternative. The set of atypical data in general represents a certain transition zone in the conditions of one-class classification. This zone is a distrust zone, but is important from the point of view of analysis to determine the characteristics of data that have certain dimensional characteristics, due to which they do not belong to the class in the same class classification or do not lie in space outside the class. The structure of the formation of information technology is presented in Fig. 1.

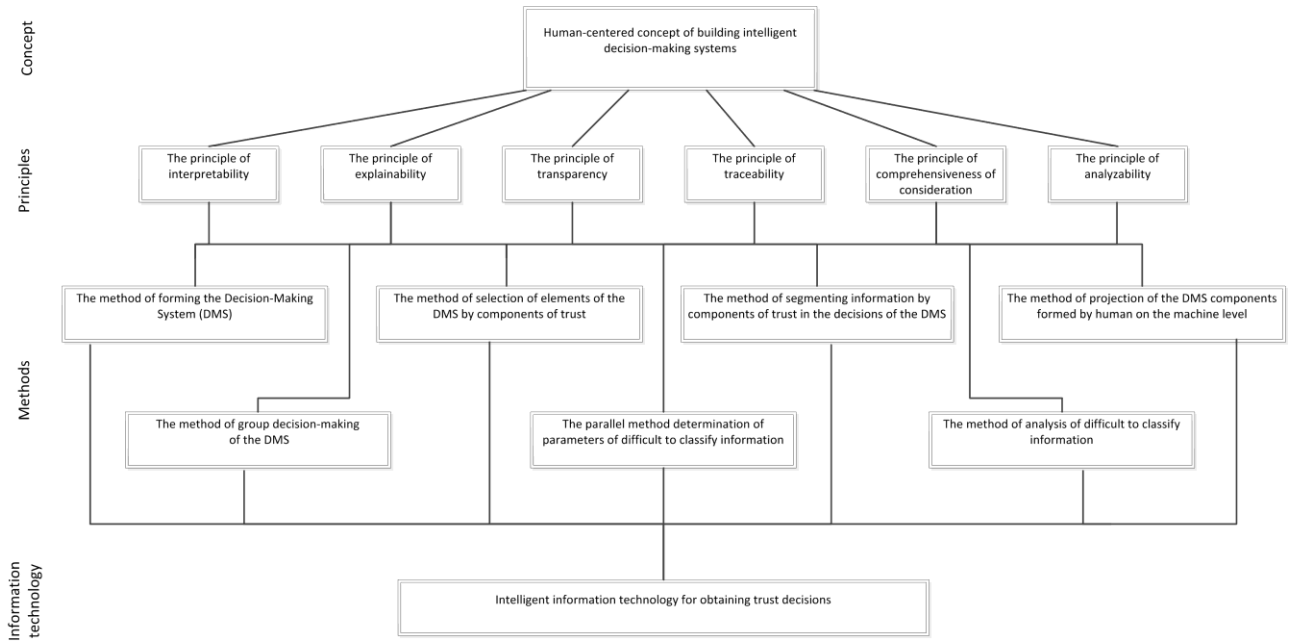

Fig. 1. The structure of the formation of intelligent information technology to obtain trust decisions

Considering the formation of data seals that belong to the class in the hyperspace of features atypical data are at the boundaries of spatial seals. Data analysis should be able to identify signs of changing values which can improve the spatial grouping of atypical data. That is, those features of the data that affect the differentiation of atypical data are detected. In the class data spatial data consolidation model, atypical data is located at the class boundaries, and the set of atypical data includes class-related and non-class data. Thus, a local spatial zone of data delimitation is formed, which is limited to a set of atypical data. Accordingly, among atypical data, it is possible to locally detect features that affect the differentiation of data in the local segment of atypical data. It is also necessary to determine the degree of impact and the importance of the characteristics of the data, the change of which improves the delimitation, which necessitates the development of a set of relevant metrics. Accordingly, the analysis is realized by applying the method of detecting atypical features on a set of atypical data. It is important to localize the research of features, which is limited to the set of atypical data, because the change of features does not apply to the set of typical data for consolidated decisions. The defined principles form the basis for the development of intelligent information technology for obtaining trust decisions.

Changing the characteristics of the data does not affect the spatial shift of the set of typical data, which allows not changing the informativeness of this data, thus preserving the original values of data and ensuring confidence in them and the decisions of models built using this data. Analysis of atypical data based on the detection of signs of impact on spatial grouping allows identifying the conditions for the formation of spatial boundary locations of data to improve the delimitation of data and identify signs of significant impact on it.

The set of principles are formed using a human-centered concept, which is the defining concept of socio-technological development of society. Next, using the developed trust ontology, the components of trust were

identified. Using a comprehensive method of determining the alignment between the trust ontology and the structured domain, a set of ethical principles was established, which will form the basis for the design and development of intelligent information technology in terms of trust decisions. Using the developed syntactic, structural and semantic methods of alignment of ethical principles to the components of the structured domain, the importance of each principle in relation to the general set of principles was established. Ensuring the receipt of trust decisions is carried out by building information technology on certain ethical principles.

In order to ensure the implementation and practical implementation of the established set of ethical principles, a number of methods of practical orientation and functional purpose have been developed.

Intelligent information technology is a holistic form of combining the developed methods into a single system, which is subject to ensuring trust in the decisions obtained with intelligent information system. Obtaining trust decisions determines the general trust in intelligent information technology within certain ethical principles of human-centered approach.

The developed information technology to ensure trust in intelligent information systems in the form of interaction of the proposed methods is presented in Fig. 2.

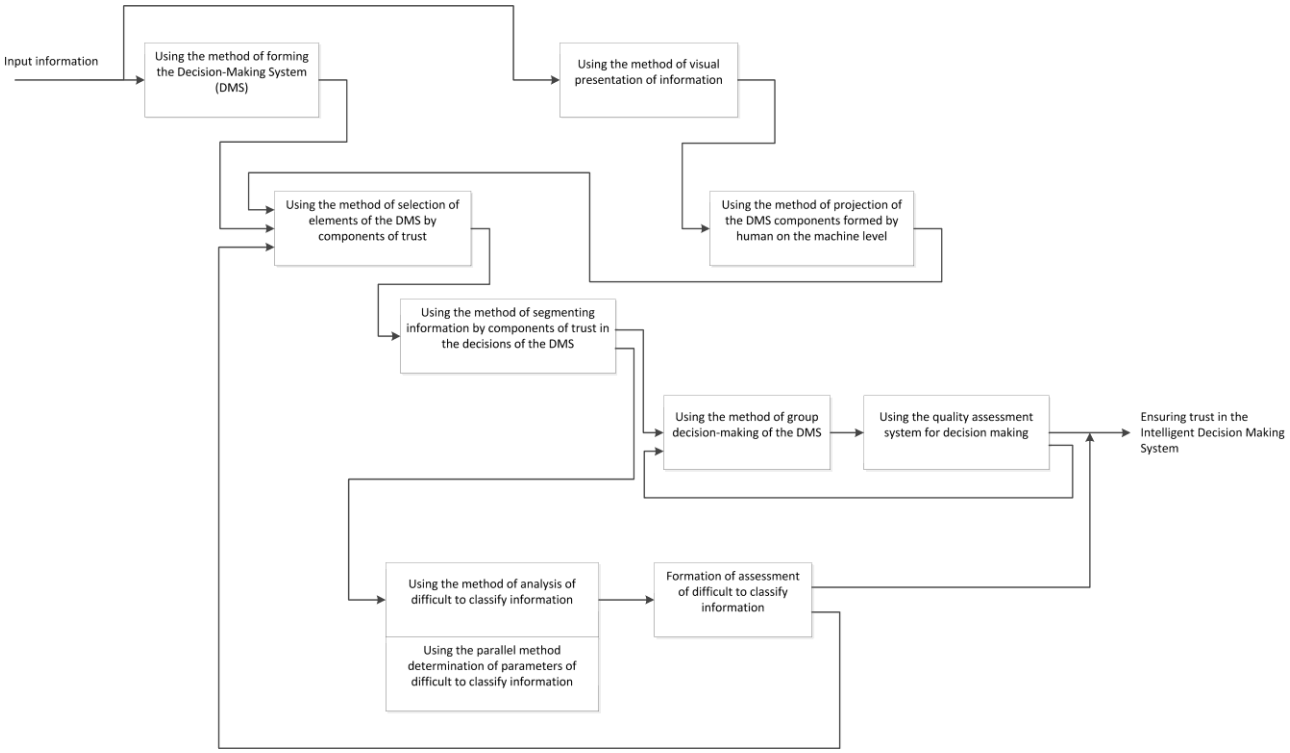

Fig. 2. The structure of intelligent information technology components interaction to obtain trust decisions

Intelligent information technology for obtaining trust decisions defines a set of functions:

- formation of a segment of information on the basis of input information decisions on which are confidential and established components of the intelligent decision-making system based on certain ethical principles;
- formation of a segment of information on the basis of input information decisions on which are distrustful, which is established by the components of the intelligent decision-making system, such information is defined as difficult to classify;
- obtaining an opinion on the qualitative metrics of classification of input information on the total set of information, including reliable information and difficult to classify information;
- obtaining an opinion on the features that have a decisive influence on the delimitation of difficult to classify information depending on the classes of information;
- obtaining an opinion on the spatial arrangement of information classes on hyperspace features and visual presented for visual analysis and control.

Conclusions

Thus, the set of developed methods to ensure certain ethical principles of trust in decisions obtained using intelligent information systems has allowed developing intelligent information technology for obtaining trust decisions. Confidence in decisions is formed by ensuring the practical implementation of the set of ethical principles in the methods of processing input information. Information technology determines the interaction of the developed methods, which together form the trust in the decisions of the information system. The structure of formation of

intelligent information technology for obtaining trust decisions is presented, as well as the structure of interaction of components of intelligent information technology for obtaining trust decisions. The set of functions which are realized by application of the offered intellectual information technology is defined.

References

1. Stix C. A Survey of the European Union's Artificial Intelligence Ecosystem / C. Stix. — Rochester, NY : Social Science Research Network, 2019.
2. Baker-Brunnbauer J. Trustworthy AI Implementation (TAII) Framework for AI Systems / J. Baker-Brunnbauer. — Rochester, NY : Social Science Research Network, 2021.
3. Cammarota R. Trustworthy AI Inference Systems: An Industry Research View / R. Cammarota, M. Schunter, A. Rajan, F. Boemer, Á. Kiss, A. Treiber, C. Weinert, T. Schneider, E. Stapf, A.-R. Sadeghi, D. Demmler, H. Chen, S. U. Hussain, S. Riazi, F. Koushanfar, S. Gupta, T. S. Rosing, K. Chaudhuri, H. Nejatollahi, N. Dutt, M. Imani, K. Laine, A. Dubey, A. Aysu, F. S. Hosseini, C. Yang, E. Wallace, P. Norton // arXiv:2008.04449 [cs]. — 2020.
4. Janssen M. Data governance: Organizing data for trustworthy Artificial Intelligence / M. Janssen, P. Brous, E. Estevez, L. S. Barbosa, T. Janowski // Government Information Quarterly. — 2020. — Vol. 37, № 3. — Pp. 101493. doi: 10.1016/j.giq.2020.101493.
5. Kaur D. Requirements for Trustworthy Artificial Intelligence – A Review / D. Kaur, S. Uslu, A. Durresi // Advances in Networked-Based Information Systems. Advances in Intelligent Systems and Computing. Cham. — 2021. — Pp. 105–115. doi: 10.1007/978-3-030-57811-4_11.
6. Martínez-Fernández S. Developing and Operating Artificial Intelligence Models in Trustworthy Autonomous Systems / S. Martínez-Fernández, X. Franch, A. Jedlitschka, M. Oriol, A. Trendowicz // Research Challenges in Information Science. Lecture Notes in Business Information Processing. Cham. Research Challenges in Information Science. — 2021. — Pp. 221–229. doi: 10.1007/978-3-030-75018-3_14.
7. Brundage M. Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims / M. Brundage, S. Avin, J. Wang, H. Belfield, G. Krueger, G. Hadfield, M. Anderljung. // arXiv:2004.07213 [cs]. — 2020.
8. Crockett K. A. Building Trustworthy AI Solutions: A Case for Practical Solutions for Small Businesses / K. A. Crockett, L. Gerber, A. Latham, E. Colyer // IEEE Transactions on Artificial Intelligence. — 2021.
9. Li B. Trustworthy AI: From Principles to Practices / B. Li, P. Qi, B. Liu, S. Di, J. Liu, J. Pei, J. Yi, B. Zhou // arXiv:2110.01167 [cs]. — 2021.
10. Serban A. Practices for Engineering Trustworthy Machine Learning Applications / A. Serban, K. van der Blom, H. Hoos, J. Visser // 2021 IEEE/ACM 1st Workshop on AI Engineering - Software Engineering for AI (WAIN). — 2021. — Pp. 97–100. doi: 10.1109/WAIN52551.2021.00021.
11. Toreini E. The relationship between trust in AI and trustworthy machine learning technologies / E. Toreini, M. Aitken, K. Coopamootoo, K. Elliott, C. G. Zelaya, A. van Moorsel // Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. FAT* '20. New York, NY, USA. — 2020. — Pp. 272–283. doi: 10.1145/3351095.3372834.
12. Linardatos P. Explainable AI: A Review of Machine Learning Interpretability Methods / P. Linardatos, V. Papastefanopoulos, S. Kotsiantis // Entropy. — 2021. — Vol. 23, № 1. — Pp. 18. doi: 10.3390/e23010018.
13. Cheng L. Socially Responsible AI Algorithms: Issues, Purposes, and Challenges / L. Cheng, K. R. Varshney, H. Liu // Journal of Artificial Intelligence Research. — 2021. — Vol. 71. — Pp. 1137–1181. doi: 10.1613/jair.1.12814.
14. Yang Z. Hierarchical Attention Networks for Document Classification / Z. Yang, D. Yang, C. Dyer, X. He, A. J. Smola, E. H. Hovy // HLT-NAACL. 2016.

Eduard Manziuk **Едуард Манзюк**	PhD, Associate Professor of Computer Science, Khmelnytskyi National University, Khmelnytskyi, Ukraine, e-mail: eduard.em.km@gmail.com orcid.org/0000-0002-7310-2126, Scopus Author ID: 57203157573, ResearcherID: AAG-9231-2019 https://scholar.google.com/citations?user=bwW-dBEAAAAJ&hl	кандидат технічних наук, доцент кафедри комп'ютерних наук, Хмельницький національний університет, Хмельницький, Україна