UDC 347.918:004.8]:174
DOI https://doi.org/10.32782/2521-6473.2025-3.2

O. A. Shamov, Intelligent systems researcher head
of Human Rights Educational Guild

FROM ETHICAL PRINCIPLES TO CERTIFIED PRACTICE: A TIERED REGULATORY MODEL FOR ARTIFICIAL INTELLIGENCE IN ONLINE DISPUTE RESOLUTION

This article explores the lack of effective mechanisms for ensuring the ethical use of artificial intelligence (AI) in online dispute resolution (ODR) systems. Introduction: The rapid development and implementation of AI-powered ODR platforms are significantly outpacing the formation of adequate legal and ethical frameworks. This creates substantial risks to the fundamental rights of citizens, particularly the right to a fair trial, equality of arms, and due process. Existing international documents, such as the Council of Europe's Charter and UNESCO's Recommendations, establish important high-level principles (transparency, non-discrimination, human control). However, these are largely declarative and do not offer concrete tools for their practical implementation and verification. Consequently, a dangerous gap emerges between proclaimed ethical goals and the actual functioning of ODR platforms, which can lead to systemic errors, the amplification of societal biases, and the erosion of trust in digital justice. The core problem is the absence of a verifiable and enforceable bridge between principle and practice, leaving users vulnerable to opaque, potentially biased algorithmic decision-making.

The Purpose of this article is to develop and substantiate a comprehensive regulatory model capable of bridging the gap between ethical principles and the practical application of AI in ODR. The objective is to propose a concrete, actionable framework that moves beyond abstract guidelines to create enforceable standards for ODR platform certification. The research methodology includes the dialectical method of cognition, a systematic analysis of existing ethical and legal frameworks, a comparative legal method to study regulatory approaches in other high-stakes domains (specifically, the EU AI Act and the U.S. FDA model for medical devices), and a modeling method to design the new regulatory structure. This multi-faceted approach allows for a thorough examination of the problem's theoretical underpinnings and the formulation of a practical, evidence-based solution. The research demonstrates that existing principle-based approaches are insufficient. As a key element of scientific novelty, a model of mandatory, differentiated (tiered) technical-procedural certification for ODR platforms is proposed. This model is risk-based and envisions three tiers of certification, with requirements escalating in stringency according to the legal complexity and social significance of the disputes. Tier 1 (low risk, e.g., e-commerce disputes) mandates basic transparency and data security. Tier 2 (medium risk, e.g., consumer credit disputes) adds requirements for regular bias audits and the use of Explainable AI (XAI) tools. Tier 3 (high risk, e.g., court-annexed ODR) mandates a crucial shift from mere «explainability» to full model «interpretability», ensuring the legal coherence of decisions, and guarantees an unconditional right to review by a human judge. The study further distinguishes between post-hoc «explainability» and the more rigorous standard of «interpretability», arguing the latter is essential for satisfying the principles of a reasoned legal decision in high-stakes contexts.

The proposed certification model allows for the transformation of abstract ethical principles into specific, verifiable, and legally meaningful standards. It offers a flexible and scalable approach to regulation that fosters innovation while safeguarding fundamental rights. It is proposed to codify the duties of the ODR platform as the «Fourth Party» in the dispute, establishing a regime of legal accountability. This framework creates a foundation for building trust in digital justice and ensuring its sustainable development, balancing efficiency with the non-negotiable demands of due process.

Key words: *online dispute resolution, ODR, artificial intelligence, AI, AI ethics, AI regulation, certification, tiered model, Explainable AI, XAI, due process, digital justice.*

О. А. Шамов. Від етичних принципів до сертифікованої практики: багаторівнева модель регулювання штучного інтелекту в онлайн-вирішенні спорів

У статті досліджено проблему відсутності дієвих механізмів для забезпечення етичного використання штучного інтелекту (ШІ) у системах онлайн-вирішення спорів (ODR). Проблема полягає в тому, що стрімкий розвиток та впровадження ODR-платформ на базі ШІ значно випереджають формування адекватних правових та етичних рамок. Це створює суттєві ризики для фундаментальних прав громадян, зокрема права на справедливий суд, рівність сторін та належну правову процедуру. Існуючі міжнародні документи, такі як Хартія Ради Європи та Рекомендації ЮНЕСКО, встановлюють важливі високорівневі принципи (прозорість, недискримінація, людський контроль), однак вони мають переважно декларативний характер і не пропонують конкретних інструментів для їх практичної імплементації та верифікації. Таким чином, виникає небезпечний розрив між проголошеними етичними цілями та реальною практикою функціонування ODR-платформ, що може призвести до системних помилок, посилення соціальної упередженості та підриву довіри до цифрового правосуддя.

Метою статті є розробка та обґрунтування комплексної регуляторної моделі, здатної подолати розрив між етичними принципами та практичним застосуванням ШІ в ODR. Методи дослідження включають діалектичний

метод пізнання, системний аналіз існуючих етичних та правових рамок, порівняльно-правовий метод для вивчення регуляторних підходів в інших сферах (зокрема, Закону ЄС про ШІ та моделі FDA), а також метод моделювання для розробки нової регуляторної структури.

В роботі доведено, що існуючі принципові підходи є недостатніми. Запропоновано наукову новизну – модель обов'язкової, диференційованої (багаторівневої) техніко-процесуальної сертифікації ODR-платформ. Ця модель базується на оцінці ризиків і передбачає три рівні сертифікації, вимоги до яких зростають відповідно до правової складності та соціальної значущості спорів. Запропонована модель сертифікації дозволяє перетворити абстрактні етичні принципи на конкретні, перевірювані та юридично значущі стандарти. Вона пропонує гнучкий та масштабований підхід до регулювання, який стимулює інновації, водночас захищаючи фундаментальні права.

Запропоновано кодифікувати обов'язки ODR-платформи, як «четвертої сторони» у спорі, запровадивши для неї режим юридичної відповідальності. Це створює основу для побудови довіри до цифрового правосуддя та його сталого розвитку.

Ключові слова: *онлайн-вирішення спорів, ODR, штучний інтелект, ШІ, етика ШІ, регулювання ШІ, сертифікація, багаторівнева модель, Пояснюваний ШІ, XAI, належна правова процедура, цифрове правосуддя.*

Formulation of the problem. Online Dispute Resolution (ODR) is undergoing a phase of transformational development, evolving from simple communication tools into complex ecosystems that actively use artificial intelligence (AI) technologies. These platforms promise unprecedented efficiency, speed, and access to justice, especially in resolving mass disputes of low to medium value [1]. However, this rapidly accelerating technological evolution poses a fundamental challenge to the modern legal system.

The implementation of algorithmic systems capable of analyzing evidence, predicting outcomes, and even proposing ready-made decisions is occurring in a regulatory vacuum.

The connection of this problem to important scientific and practical tasks is direct. Firstly, at stake are fundamental constitutional principles such as the right to a fair trial, equality of parties before the law, and due process. The opacity of algorithms (the «black box problem») and the risk of algorithmic bias, where AI reproduces and amplifies existing social stereotypes, pose a direct threat to these principles [2]. Secondly, the absence of clear standards undermines public trust in digital justice, which can negate all the potential benefits of ODR. Thirdly, the legal community faces a practical task: how to ensure effective supervision and control over technologies that are inherently complex and opaque to most lawyers and judges. Thus, developing adequate ethical and legal frameworks for the use of AI in ODR is not just an academic exercise but an urgent necessity for ensuring the stability and fairness of the legal system in the digital age.

Analysis of recent research and publications. The foundation for the ethical regulation of AI in justice has been laid at the international level. Key documents here are the «European Ethical Charter on the use of Artificial Intelligence in judicial systems and their environment», developed by the Council of Europe [3], and the «Recommendation on the Ethics of Artificial Intelligence» by UNESCO [4]. These documents established five fundamental principles: respect for human rights, non-discrimination, quality and security, transparency and fairness, and ensuring human control. They serve as an important moral and conceptual guide.

The academic community is actively researching this issue. Leading scholars such as Ethan Katsh and Orna Rabinovich-Einy, who are pioneers of ODR, have formulated the concept of the ODR platform as the «fourth party» in a dispute (in addition to the two parties and the neutral mediator) [5]. This «fourth party»-the technology-actively influences the process and, therefore, must have not only functions but also duties. Richard Susskind, developing the idea of «online courts», emphasizes the inevitability of the technological transformation of justice while also raising questions about preserving its core values [6]. Works dedicated directly to ethical risks focus on problems of algorithmic bias, opacity, and the dehumanization of the dispute resolution process [2, 7]. In response to the «black box problem», the field of Explainable AI (XAI) is actively developing, seeking technical ways to increase the transparency of algorithmic decisions [8].

However, despite a significant number of publications, a key part of the general problem remains unresolved: the lack of a practical, effective, and legally binding mechanism that would connect high-level ethical principles with the technical reality of ODR platforms. Existing frameworks are predominantly declarative.

Academic discussions excel at identifying risks but rarely offer comprehensive, ready-to-implement regulatory models. The question remains open: how, in practice, can one verify an ODR platform's compliance with the principle of «fairness» or «transparency»? How can a system be created that is both flexible, able to adapt to technological development, and rigid enough to protect fundamental rights? This article is dedicated to solving this very problem-building a bridge between principle and practice.

The aim of this research is to develop and scientifically substantiate a comprehensive regulatory model for the use of artificial intelligence in online dispute resolution that would ensure compliance with ethical standards and the protection of human rights through the implementation of a mandatory, differentiated certification system.

To achieve this aim, the following objectives were set:

To analyze existing international ethical frameworks and scientific doctrines on the use of AI in justice to determine their strengths and limitations.

To identify and systematize the key ethical risks associated with the use of AI in ODR, particularly algorithmic bias, opacity, and the erosion of procedural guarantees.

To study, using the comparative legal method, successful risk-based regulatory models, particularly the EU Artificial Intelligence Act and the FDA's medical device certification system, to identify possibilities for their adaptation to the ODR sphere.

To develop the concept of a tiered (differentiated) certification model for ODR platforms, which involves different levels of requirements depending on the complexity and social significance of the disputes

To justify the need for a shift from the standard of «explainability» to the standard of «interpretability» for high-risk ODR platforms as a key condition for ensuring the right to a reasoned decision.

To formulate specific proposals for the implementation of the proposed model and to identify prospects for further scientific research in this direction.

Presenting main material. The core thesis of this research is that existing ethical frameworks, despite their fundamental importance, are insufficient for the effective regulation of AI in ODR. They establish the «what»-the goal to be striven for (fairness, transparency) but do not offer the «how» a concrete, verifiable, and legally binding mechanism for achieving this goal. This gap creates an illusion of security, while in practice, ODR platforms can function as «black boxes» that make legally significant decisions without proper oversight or the possibility of effective appeal.

To bridge this gap, a transition from declarations to effective tools is necessary. An analogy can be drawn with other areas where technology carries high risks. The most relevant example is the approach embedded in the EU Artificial Intelligence Act. This act does not attempt to regulate all AI uniformly but introduces a differentiated model based on four levels of risk: unacceptable, high, limited, and minimal [9]. For each category, specific requirements are established: from a complete ban (for social scoring systems) to strict requirements for data quality, transparency, human oversight, and cybersecurity (for high-risk systems). A similar approach has been used for decades by the U.S. Food and Drug Administration (FDA), which classifies medical devices into three classes depending on the potential harm to the patient, establishing a distinct control regime for each class [10].

This risk-based approach is an ideal model for regulating AI in ODR. After all, the risks posed by an algorithmic decision in a dispute over a $100 refund for a product in an online store are incomparably lower than the risks in a consumer credit debt collection case or a family dispute handled within a court-annexed ODR program.

Based on this, a model of mandatory, differentiated technical-procedural certification for ODR platforms is proposed, which includes three tiers:

Tier 1: Low Risk.

– Scope: Mass, low-value disputes with clearly defined rules, for example, disputes on e-commerce platforms (product returns, non-conformity with description).

– Certification Requirements:

1. Transparency of Use: Clear and unambiguous information provided to the user that AI is involved in resolving the dispute.

2. Data Security: Compliance with basic data protection standards (e.g., GDPR).

3. Availability of Escalation: A simple, understandable, and accessible mechanism for escalating the dispute to a human operator in case of disagreement with the automated decision.

Tier 2: Medium Risk.

– Scope: Disputes that have significant financial or social consequences for an individual, for example, disputes in the areas of consumer credit, insurance, and rental agreements.

– Certification Requirements (in addition to Tier 1):

1. Bias Audit: A requirement for developers to conduct regular (e.g., annual) audits of algorithms and the datasets on which they were trained to identify and minimize the risks of discrimination based on protected characteristics (gender, race, age, etc.). Audit reports must be submitted to the certification body

2. Explainability of the Decision (XAI): The platform must be technically capable of providing the user with an explanation of the decision made. This is not just a statement of the result, but a provision of the key factors that influenced it. Particularly effective are so-called counterfactual explanations («Your loan application would have been approved if your verified income had been 10% higher») [8].

3. Human-on-the-loop: For this category of disputes, not just an escalation mechanism but proactive supervision by a qualified specialist must be ensured, who can intervene in the process and correct or cancel the algorithmic decision.

Tier 3: High Risk.

– Scope: Disputes that are directly integrated into the state justice system (court-annexed ODR programs), as well as disputes concerning sensitive areas (e.g., certain categories of family disputes, labor disputes regarding dismissal).

– Certification Requirements (in addition to Tiers 1 and 2):

1. Model Interpretability: At this level, mere «explainability» is no longer sufficient. The requirement must be «interpretability». While explainability is an attempt to peek into the «black box» post-hoc, interpretability requires

that the AI model itself be built in such a way that its logic is understandable and consistent with legal norms and principles [11]. The decision of such a system must not only be explained but also legally coherent and reasoned, as required by the right to a fair trial.

2. Guaranteed Right to Review: The user must have an unconditional, clearly defined right to a full review of their case by a qualified judge or arbitrator, whose decision will be final.

3. Highest Data Standards: A requirement for full data provenance for the data on which the model was trained and operates, to ensure its quality, relevance, and impartiality.

The implementation of such a model allows for the codification of the duties of the ODR platform as the «fourth party» in the dispute [5]. The platform ceases to be a passive tool and becomes an active subject of the process, bearing legal responsibility for compliance with certification requirements. This creates a legal basis for holding developers and operators liable in case of violations of users' rights due to algorithmic errors or bias.

Conclusions. The rapid integration of artificial intelligence into online dispute resolution systems is an inevitable stage in the evolution of justice. However, this process must not be uncontrolled. The efficiency and speed offered by AI cannot be achieved at the cost of eroding fundamental rights and the principles of due process.

This research has demonstrated a critical gap between high-level ethical principles, which are largely declarative, and the practical absence of mechanisms for their implementation and oversight. To bridge this gap, a scientific proposal – a tiered model of mandatory technical-procedural certification for ODR platforms based on risk assessment – has been developed and substantiated

The main conclusions are:

1. A «one-size-fits-all» approach to regulating AI in ODR is ineffective. A differentiated, risk-based model, similar to that embedded in the EU AI Act, is the most appropriate and flexible.

2. The proposed three-tiered certification system (low, medium, high risk) allows for the establishment of proportional requirements for ODR platforms, balancing the need for innovation with the necessity of protecting citizens' rights.

3. For high-risk disputes, it is necessary to demand a shift from the standard of «explainability» to the more stringent standard of «interpretability» of the AI model, which is a key condition for upholding the right to a reasoned and legally sound decision.

The implementation of such a certification system allows for the legal codification of the duties of the ODR platform as the «fourth party» in the dispute, creating a legal basis for its liability.

The proposed model is a concrete and practical solution that transforms abstract ethics into verifiable standards, contributing to the construction of a truly fair and reliable digital justice system.

Prospects for further research in this area include:

– Developing detailed technical protocols and methodologies for auditing ODR platforms for algorithmic bias.

– Researching the organizational and legal forms for the creation and functioning of an independent national or international body responsible for the certification of ODR platforms.

– Conducting empirical studies on the psychological impact of interaction with automated justice systems on the parties to a dispute and their perception of procedural fairness.

– Analyzing issues of cross-border recognition and enforcement of decisions made by certified ODR platforms.

Bibliography:

1. Larson, D. The Future of Online Dispute Resolution (ODR): Definitions, Standards, Disability Accessibility, and Legislation. *Faculty Scholarship*. 2022. 559. URL: https://surl.li/rprlsi

2. Cotton, N., Kelly, E. and Williams, D. Consider the Risks of Algorithmic Bias. *The Jabian Journal*. 2023. URL: https://surl.li/gubqhu

3. European Commission for the Efficiency of Justice (CEPEJ). European ethical Charter on the use of Artificial Intelligence in judicial systems and their environment. *Council of Europe Website*. 2018. 24 p. URL: https://rm.coe.int/ethical-charter-en-for-publication-4-december-2018/16808f699c

4. UNESCO. Recommendation on the Ethics of Artificial Intelligence. *UNESCO Website.* 2021. 46 p. URL: https://unesdoc.unesco.org/ark:/48223/pf0000380455

5. Katsh E., Rabinovich-Einy O. Digital Justice: Technology and the Internet of Disputes. *Oxford University Press*. 2017. 288 p. URL: https://academic.oup.com/book/27240

6. Susskind R. Online Courts and the Future of Justice. *Oxford University Press*. 2019. 400 p. URL: https://surl.li/nxsimw

7. Shen, J., DiPaola, D., Ali, S., Sap, M., Won Park, H. and Breazeal C. Empathy Toward Artificial Intelligence Versus Human Experiences and the Role of Transparency in Mental Health and Social Support Chatbot Design: Comparative Study. *National Institutes of Health Website*. 2024. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC11464935/

8. Wachter, S., Mittelstadt, B. & Russell, C. Counterfactual explanations without opening The black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*. 2018. URL: https://surl.li/qtbwhf

9. Veale M., Borgesius F. Demystifying the Draft EU Artificial Intelligence Act. *Computer Law Review International*. Volume 22. Issue 4. 2021. URL: https://surl.li/sjvxht

10. U.S. Food and Drug Administration. Regulatory Controls. *U.S. Food and Drug Administration Website*. 2018. URL: https://surl.li/inktmt

11. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*. 2019. Vol. 1. P. 206–215. URL: https://www.nature.com/articles/s42256-019-0048-x

Дата надходження статті: 04.07.2025
Дата прийняття статті: 09.07.2025
Опубліковано: 01.10.2025