https://doi.org/10.54251/2522-4026.2026.2.25au

UDC 343:004.8

M.V. Talan, A. S. Talan*

Doctor of Law, Professor, Kazan Federal University, Kazan, Russia

Candidate of Chemical Sciences, Deputy General Director, Aktru-Educational Technologies LLC, Khimki, Russia

***Correcponding author**: atghostru@gmail.com.

FROM ASIMOV'S LAWS TO CRIMINAL LIABILITY FOR UNSAFE ARTIFICIAL INTELLIGENCE: A RISK-BASED MODEL FOR PROVIDER ACCOUNTABILITY

Abstract

The article argues that the widely discussed analogy between Asimov's laws of robotics and modern AI ethics is useful only as a metaphor. OECD and Council of Europe instruments require human-centered, transparent, robust and accountable AI, but they mainly operate as principles, product-safety duties or administrative compliance rules. They have not yet led to the creation of a clear criminal-law duty for commercial providers of general-purpose AI systems whose products foreseeably enable violent crime, cybercrime, child sexual abuse material, fraud or other serious harms. Using comparative doctrinal analysis, the article proposes an EU-oriented comparative model of criminal liability for developers, company officers and distributors who place unsafe AI systems on the market without proportionate guardrails, abuse-detection mechanisms, incident response and lawful escalation procedures. The proposed model preserves the human-centered structure of criminal law by treating AI as a dangerous instrument whose provider may be liable for negligent, reckless or intentional release.

Keywords: artificial intelligence, AI providers, Kazakhstan, EU AI Act, cybercrime, child sexual abuse material, Asimov's laws.

Introduction

The language of AI ethics often returns to Isaac Asimov's fictional laws of robotics. The first law prohibits harm to humans; the second requires obedience to human orders; the third protects the robot's own existence; and the later "zeroth law", introduced in Asimov's Robots and Empire through R. Daneel Olivaw's hypothesis and R. Giskard Reventlov's attempted application, protects humanity as a whole [1]. Modern international instruments express similar moral intuitions. The European approach emphasizes protection of human rights, democracy and the rule of law; OECD principles require human-centered values, transparency, robustness and accountability; and the Council of Europe Framework Convention requires AI life-cycle activities to be consistent with human rights, democratic institutions and legality [2-6].

AI ethics principles are mainly governance norms addressed to states, providers and deployers. They do not answer a difficult contemporary question: when should the human developer, corporate officer or distributor of an AI system face criminal liability because the system was released without reasonable safeguards against foreseeable criminal misuse?

This question is urgent because general-purpose and agentic AI systems are no longer passive text generators [7]. They can draft phishing messages, generate exploit code, automate reconnaissance, produce synthetic voices, create sexual deepfakes and guide users through multi-step harmful plans [8-10]. Recent research describes real-world misuse patterns across text, image, audio and video modalities [8], and newer work links provider choices, open-weight releases and insufficient safeguards to foreseeable downstream abuse [9]. The legal issue is therefore not whether AI has mens rea. The issue is whether human actors who commercialize or deliberately distribute dangerous AI have a legally enforceable duty comparable to duties already known in personal-data protection, product safety and cybersecurity.

Method and normative background

This article uses comparative doctrinal analysis. It compares soft-law AI ethics, binding European regulatory instruments and selected Kazakhstan legal concepts. The EU AI Act

establishes a risk-based framework for prohibited, high-risk and general-purpose AI systems [6]. It is important because it moves from abstract ethics to concrete provider obligations: risk management, technical documentation, logging, human oversight, cybersecurity and obligations for general-purpose models. However, its sanctions are primarily administrative and regulatory, not classical criminal penalties.

The Council of Europe Framework Convention similarly creates state duties to ensure that AI respects human rights, democracy and the rule of law [5]. OECD principles, updated to reflect newer AI risks, provide a global vocabulary of trustworthiness and accountability [4]. The High-Level Expert Group's Ethics Guidelines for Trustworthy AI add human agency, technical robustness, privacy, transparency, fairness, societal well-being and accountability [3]. These principles correspond loosely to Asimov's laws, but they are not self-executing criminal norms.

Kazakhstan law offers a useful analogy. The Law "On Personal Data and Their Protection" defines owners and operators, requires lawful collection and processing, and imposes duties to protect rights and freedoms of personal-data subjects [11]. This structure can inspire an AI-safety duty: a provider that profits from a powerful AI system should not be allowed to externalize all foreseeable criminal risks to users. At the same time, Kazakhstan's Penal Code remains human-centered: Article 15 states that criminal responsibility applies to sane individuals of the relevant age [12]. Existing cybercrime provisions, including Articles 205-210, address unauthorized access, illegal modification, disruption of informatization objects and malicious software [12]. They can punish many users of AI-enabled cybercrime, but they do not clearly cover the provider who negligently or recklessly releases a dangerous AI product without basic safeguards.

The accountability gap

The central gap is the absence of a graded rule that connects AI-system risk, provider knowledge, missing guardrails and criminal consequences. A company may claim that it merely provides a neutral tool. An independent developer may publish an open model and argue that later misuse is outside his control. These claims are sometimes valid. But they become weaker when the system is marketed for dangerous capabilities, when guardrails are intentionally removed, when known jailbreaks are ignored, or when the provider has internal evidence of imminent serious harm and no lawful escalation procedure.

The Canadian examples show why this problem cannot be reduced to abstract philosophy. In 2026, a civil claim filed in British Columbia alleged that OpenAI knew a school shooter had used ChatGPT to plan a mass-casualty attack, closed one account, but did not alert authorities before the Tumbler Ridge shooting; these allegations have not been tested in court [13]. The legal significance is not that every disturbing prompt must be reported. If a provider's own review process identifies a credible risk of imminent serious violence, criminal law may need a duty to preserve evidence and escalate through lawful channels. A second Canadian example, the Vancouver lawyer who submitted fictitious ChatGPT-generated cases in court, illustrates a different risk: unverified AI outputs can threaten the administration of justice even without violent intent [14].

Criminal-law policy must therefore distinguish ordinary AI error, civil negligence, regulatory non-compliance and genuinely culpable release of dangerous systems. It must also avoid turning providers into general police agents. "Beacons" should mean privacy-preserving abuse signals, secure logs, content provenance, rate limits, red-team triggers and narrowly defined lawful reporting duties for child sexual abuse material, credible imminent violence or severe cyberattacks. They should not mean unlimited surveillance of lawful users.

Proposed legal model

The proposed offence can be formulated as "unsafe release or distribution of a high-risk AI system creating a foreseeable risk of serious criminal harm." The protected interests are life, health, sexual inviolability of minors, information security, property, public safety and the administration of justice.

The subject of liability should remain human. For independent developers, liability attaches personally where they intentionally, recklessly or grossly negligently release a dangerous system. For companies, EU Member States could use a two-track approach: criminal liability of responsible

officers and special criminal-law or administrative-criminal measures against legal entities, such as turnover-based fines, confiscation of unlawful profit, suspension of deployment, mandatory audit or liquidation in extreme cases. This respects the natural-person principle of criminal law while preventing the corporate form from absorbing all responsibility.

The minimum technical duty should include: pre-release risk assessment; red-teaming for violence, CSAM, fraud and cyber misuse; refusal and safe-completion policies; monitoring for abuse patterns; secure incident logs; vulnerability disclosure; model-card and system-card documentation; age-sensitive and domain-sensitive controls; provenance or watermarking where technically feasible; and a procedure for lawful escalation of credible imminent threats. The duty should be proportionate to the model's capability, deployment scale, commercial character and actual knowledge of misuse [15-20].

The following table proposes a guardrail-based sanction scale.

Level	**Technical and commercial condition**	**Fault standard**	**Proposed response**
G0: compliant AI	Documented risk assessment, red-team testing, abuse monitoring, incident response, user-facing warnings and lawful escalation channels.	No fault.	No criminal liability; regulatory audit only after incidents.
G1: documentation breach	Missing or outdated model card, risk file, logging policy or user notice, but no serious harm and no known dangerous capability.	Ordinary negligence.	Administrative fine; mandatory remediation; temporary deployment limits.
G2: unsafe commercial release	No risk-assessment, paid or large-scale AI system released without basic guardrails against violence, fraud, malware or critical-infrastructure abuse.	Gross negligence.	Criminal fine for responsible officers or equivalent legal-entity measure; suspension until certified; compensation fund contribution.
G3: reckless continuation	Provider knows of repeated jailbreaks, harmful outputs or abuse campaigns and fails to patch, rate-limit, log, warn or suspend high-risk functions.	Recklessness.	Higher fines; prohibition on high-risk deployment; officer disqualification; imprisonment for responsible individuals where serious harm occurs.
G4: deliberate dangerous AI	Model or fine-tune is marketed as "uncensored", "no guardrails", "malware generator", "phishing agent", "CSAM/deepfake generator" or tool for evading law enforcement.	Direct or conditional intent.	Liability for aiding or facilitating predicate crimes; confiscation; platform takedown; imprisonment for individual developers/distributors; liquidation or market ban for entities.
G5: concealment after red flags	Destruction of logs, evasion of audit, failure to report legally reportable CSAM or credible imminent mass	Direct intent.	Aggravated liability; separate offence for obstruction; higher sanctions and mandatory external monitoring.

	violence, or misleading regulators after an incident.		

The table is deliberately technology-neutral. It covers closed commercial systems, open-weight releases, fine-tunes, agentic tools and API services. Open-source status should not automatically create liability, because open research is socially valuable. But open distribution becomes legally relevant when the released weights or instructions are foreseeably specialized for serious crime and the developer ignores available mitigations such as staged release, capability evaluation, access controls, non-generative harmful-specialization audits or platform moderation [15, 19, 20].

Criminal-law limits and defenses

The proposed model must be constrained by legality, culpability and proportionality. First, a provider should not be punished merely because a user committed a crime with a general tool. The prosecution must prove a legally defined unsafe release, causation or material contribution, foreseeability and fault. Second, no liability should arise for good-faith safety research, defensive cybersecurity work, journalism, academic study or lawful model evaluation when safeguards and disclosure rules are followed. Third, a provider should have a due-diligence defense if it can show documented testing, reasonable guardrails, timely patching, user enforcement, cooperation with lawful requests and independent audit.

For Kazakhstan, this EU-oriented comparison supports a practical path: not to declare AI a criminal subject, but to add an AI-provider safety duty to digital legislation, cross-reference it in the Penal Code for aggravated cases, and harmonize it with personal-data, cybersecurity and consumer-protection rules. This would align Kazakhstan with the European risk-based model while adding what the European model still lacks: a focused criminal-law response to deliberate or reckless release of dangerous AI.

Conclusion

Asimov's laws remain a useful metaphor because they express a simple intuition: intelligent machines should not harm human beings or humanity. Modern AI governance translates that intuition into principles of human oversight, safety, transparency and accountability. But principles alone are insufficient where providers profit from systems capable of enabling serious crime. Criminal law should not punish AI as if it were a person. It should punish human and corporate actors who intentionally, recklessly or grossly negligently release dangerous AI without proportionate guardrails, lawful abuse signals and incident response. A graded model would protect innovation while creating a clear boundary between ordinary AI risk, civil liability, administrative non-compliance and criminally culpable unsafe release.

References

1. Asimov, I. Robots and Empire. Doubleday, 1985. https://openlibrary.org/works/OL46178W/Robots_and_Empire
2. European Parliament. Resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics, 2015/2103(INL). https://www.europarl.europa.eu/doceo/document/TA-8-2017-0051_EN.html
3. High-Level Expert Group on Artificial Intelligence. Ethics Guidelines for Trustworthy AI. European Commission, 2019. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
4. OECD. Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449, adopted 2019, amended 2024. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449
5. Council of Europe. Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, CETS No. 225, 2024. https://www.coe.int/en/web/artificial-intelligence/the-framework-convention-on-artificial-intelligence

6. European Parliament and Council. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence. Official Journal of the European Union, 2024. https://eur-lex.europa.eu/eli/reg/2024/1689/oj
7. Nannini, L., Smith, A. L., Maggini, M. J., Panai, E., Feliciano, S., Tiulkanov, A., Maran, E., Gealy, J., and Bisconti, P. AI Agents Under EU Law. arXiv:2604.04604, 2026. https://arxiv.org/abs/2604.04604
8. Marchal, N., Xu, R., Elasmar, R., Gabriel, I., Goldberg, B., and Isaac, W. Generative AI Misuse: A Taxonomy of Tactics and Insights from Real-World Data. arXiv:2406.13843, 2024. https://arxiv.org/abs/2406.13843
9. Kamachee, M., Casper, S., Ding, M. L., Yew, R.-J., Reuel, A., Biderman, S., and Hadfield-Menell, D. Video Deepfake Abuse: How Company Choices Predictably Shape Misuse Patterns. arXiv:2512.11815, 2025. https://arxiv.org/abs/2512.11815
10. Chen, G., Wang, Y., Ji, S., Luo, X., and Wang, T. Synthetic Voices, Real Threats: Evaluating Large Text-to-Speech Models in Generating Harmful Audio. arXiv:2511.10913, 2025. https://arxiv.org/abs/2511.10913
11. Republic of Kazakhstan. Law "On Personal Data and Their Protection", No. 94-V, 21 May 2013. https://adilet.zan.kz/eng/docs/Z1300000094
12. Republic of Kazakhstan. Penal Code of the Republic of Kazakhstan, Law No. 226-V, 3 July 2014, Articles 15 and 205-210. https://adilet.zan.kz/eng/docs/K1400000226
13. Associated Press. Family sues ChatGPT-maker OpenAI over school shooting in Canada. 10 March 2026. https://apnews.com/article/canada-mass-shooting-chatgpt-openai-lawsuit-de6c2b89cea94bde3b1c288b0e1581ca
14. The Guardian. Canada lawyer under fire for submitting fake cases created by AI chatbot. 29 February 2024. https://www.theguardian.com/world/2024/feb/29/canada-lawyer-chatgpt-fake-cases-ai
15. Saeri, A. K., Lloyd George, S., Graham, J., Lacarriere, C. D., Slattery, P., Noetel, M., and Thompson, N. Mapping AI Risk Mitigations: Evidence Scan and Preliminary AI Risk Mitigation Taxonomy. arXiv:2512.11931, 2025. https://arxiv.org/abs/2512.11931
16. Mojica-Hanke, A., Goger, T., Wolfel, S., Valerius, B., and Herbold, S. Criminal Liability of Generative Artificial Intelligence Providers for User-Generated Child Sexual Abuse Material. arXiv:2601.03788, 2026. https://arxiv.org/abs/2601.03788
17. O Ciardha, C., Buckley, J., and Portnoff, R. S. AI Generated Child Sexual Abuse Material: What's the Harm? arXiv:2510.02978, 2025. https://arxiv.org/abs/2510.02978
18. Kokolaki, E., and Fragopoulou, P. Unveiling AI's Threats to Child Protection: Regulatory efforts to Criminalize AI-Generated CSAM and Emerging Children's Rights Violations. arXiv:2503.00433, 2025. https://arxiv.org/abs/2503.00433
19. Suriyakumar, V. M., Sekhari, A., Stempfle, L., Wang, R., Simpson, M., Portnoff, R., Ghassemi, M., and Wilson, A. C. Evaluation without Generation: Non-Generative Assessment of Harmful Model Specialization with Applications to CSAM. arXiv:2604.25119, 2026. https://arxiv.org/abs/2604.25119
20. Zhao, W., Khazanchi, V., Xing, H., He, X., Xu, Q., and Lane, N. D. Attacks on Third-Party APIs of Large Language Models. arXiv:2404.16891, 2024. https://arxiv.org/abs/2404.16891

М. В. Талант, А. С. Талант*
Заң ғылымдарының докторы, профессор, mtalan@inbox.ru, Қазан Федералды Университеті, Қазан, Ресей
Химия ғылымдарының кандидаты, бас директорының орынбасары, atghostru@gmail.com, "Актру-Білім Беру Технологиялары" ЖШС Химки, Ресей

АСИМОВТЫҢ ЗАҢДАРЫНАН БАСТАП ҚАУІПТІ ЖАСАНДЫ ИНТЕЛЛЕКТ ҮШІН ҚЫЛМЫСТЫҚ ЖАУАПКЕРШІЛІККЕ ДЕЙІН: ПРОВАЙДЕРЛЕРДІҢ ЕСЕП БЕРУІНІҢ ТӘУЕКЕЛГЕ НЕГІЗДЕЛГЕН МОДЕЛІ

Түйін

Мақалада Асимовтың робототехника заңдары мен қазіргі жасанды интеллект этикасы арасындағы кеңінен талқыланған ұқсастық метафора ретінде ғана пайдалы екендігі айтылған. ЭЫДҰ мен Еуропа Кеңесінің құралдары адамға бағытталған, ашық, сенімді ЖӘНЕ есеп беретін жасанды интеллектті қажет етеді, бірақ олар негізінен принциптер, өнім қауіпсіздігі міндеттері немесе әкімшілік сәйкестік ережелері ретінде әрекет етеді. Олар әлі

күнге дейін өнімдері зорлық-зомбылық қылмыстарына, киберқылмыстарға, балаларға қатысты жыныстық зорлық-зомбылық материалдарына, алаяқтыққа немесе басқа да ауыр зиян келтіруге мүмкіндік беретін жалпы мақсаттағы жасанды интеллект жүйелерін коммерциялық жеткізушілер үшін нақты қылмыстық-құқықтық міндеттерді орындауға әкелген жоқ. Салыстырмалы доктриналық талдауды қолдана отырып, мақалада пропорционалды қоршауларсыз, теріс пайдалануды анықтау механизмдерінсіз, оқиғаларға ден қоюмен және өршудің заңды процедураларынсыз нарыққа қауіпті жасанды интеллект жүйелерін енгізетін әзірлеушілерге, компания қызметкерлеріне және дистрибьюторларға ЕО-ға бағытталған қылмыстық жауапкершіліктің салыстырмалы моделі ұсынылған. Ұсынылған модель жасанды интеллектті қауіпті құрал ретінде қарастыра отырып, қылмыстық құқықтың адамға бағытталған құрылымын сақтайды, оның жеткізушісі абайсызда, абайсызда немесе қасақана босатуға жауапты болуы мүмкін.

Кілттік сөздер: жасанды интеллект, жасанды интеллект жеткізушілері, қазақстан, еоның жасанды интеллект туралы заңы, киберқылмыс, балаларға қатысты жыныстық зорлық-зомбылық материалдары, Асимов заңдары.

М.В. Талан, А. С. Талан*
Доктор юридических наук, профессор, mtalan@inbox.ru, Казанский федеральный университет, Казань, Россия
Кандидат химических наук, заместитель генерального директора, atghostru@gmail.com, ООО "Актру-образовательные технологии", Химки, Россия

ОТ ЗАКОНОВ АЗИМОВА К УГОЛОВНОЙ ОТВЕТСТВЕННОСТИ ЗА НЕБЕЗОПАСНЫЙ ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ: ОСНОВАННАЯ НА РИСКЕ МОДЕЛЬ ПОДОТЧЕТНОСТИ ПРОВАЙДЕРОВ

Аннотация

В статье утверждается, что широко обсуждаемая аналогия между законами робототехники Азимова и современной этикой ИИ полезна только как метафора. Документы ОЭСР и Совета Европы требуют ориентированного на человека, прозрачного, надежного и подотчетного ИИ, но в основном они действуют как принципы, обязанности по обеспечению безопасности продукции или административные правила соблюдения требований. Они пока не привели к созданию четкой уголовно-правовой обязанности для коммерческих поставщиков систем искусственного интеллекта общего назначения, чьи продукты, как ожидается, позволят совершать насильственные преступления, киберпреступления, материалы о сексуальном насилии над детьми, мошенничестве или причинении другого серьезного вреда. Используя сравнительный доктринальный анализ, в статье предлагается ориентированная на ЕС сравнительная модель уголовной ответственности для разработчиков, должностных лиц компаний и дистрибьюторов, которые размещают на рынке небезопасные системы искусственного интеллекта без надлежащих мер предосторожности, механизмов обнаружения злоупотреблений, реагирования на инциденты и законных процедур эскалации. Предлагаемая модель сохраняет ориентированную на человека структуру уголовного права, рассматривая искусственный интеллект как опасный инструмент, поставщик которого может быть привлечен к ответственности за небрежное, неосторожное или преднамеренное использование.

Ключевые слова: искусственный интеллект, поставщики ИИ, Казахстан, Закон ЕС об ИИ, киберпреступность, материалы о сексуальном насилии над детьми, законы Азимова.