TEME, Vol. XLIX, Nº 2, April – June 2025, pp. 251–268

Original research paper https://doi.org/10.22190/TEME241108019C
Received: November 8, 2024 UDC 004.8:347.4/.5
Revised: February 19, 2025
Accepted: March 3, 2025

BLACK BOX AS A JUSTIFICATION FOR STRICT LIABILITY FOR AI-RELATED DAMAGE

Mihajlo Cvetković*

University of Niš, Faculty of Law, Niš, Serbia

ORCID iD: Mihajlo Cvetković https://orcid.org/0000-0003-1440-7760

Abstract

Strict liability is increasingly recognised as an appropriate framework for governing high-risk artificial intelligence (AI) systems, particularly those with 'black-box' characteristics, where internal operations are opaque and difficult to interpret. The inherent complexity of AI, including strong black-box features and unpredictability post-deployment, challenges the applicability of traditional tort law, which relies on establishing fault or negligence. Strict liability provides a means to hold entities accountable, addressing the difficulties in attributing fault in AI contexts. This work evaluates the merits and drawbacks of strict liability, explores its implications within the general liability regime, and provides concrete examples of AI-related harms that support this approach. The principle of AI neutrality and the persistence of fault-based elements within ostensibly strict liability frameworks like the Product Liability Directive are also examined, underscoring the complexities in regulating AI. Serbian legal doctrines regarding dangerous objects and activities provide courts with flexibility to adjudicate AI-related damages. Judges must comprehend the nuances of AI, including distinctions between traditional deterministic software and AI exhibiting emergent behaviour. While strict liability is beneficial for victim compensation and risk management, it can also stifle innovation and impose burdens on small enterprises. A balanced approach is essential to manage AI-related risks while promoting innovation.

Key words: black-box AI, Product Liability, fault-based liability, strict liability, damage.

* Corresponding author: Mihajlo Cvetković, University of Niš, Faculty of Law, Trg Kralja Aleksandra 11, 18105 Niš, Serbia, mihajloc@prafak.ni.ac.rs

BLACK BOX ВЕШТАЧКЕ ИНТЕЛИГЕНЦИЈЕ КАО РАЗЛОГ ЗА ОБЈЕКТИВНУ ГРАЂАНСКОПРАВНУ ОДГОВОРНОСТ

Апстракт

Објективна одговорност све више се препознаје као одговарајући оквир за регулисање високо ризичних система вештачке интелигенције (*AI*), посебно оних са карактеристикама „црне кутије", где су унутрашњи процеси нетранспарентни. Инхерентна сложеност *AI*, условљена појединим алгоритмима, и непредвидивост након имплементације представљају изазов за традиционално утврђивање кривице. Објективна одговорност омогућава да се субјекти позову на одговорност, чиме се решавају проблеми узрочне везе и приписивања кривице у *AI* контексту. Рад процењује предности и недостатке објективне одговорности, истражује алтернативне моделе њене примене и даје конкретне примере штета изазваних *AI*. Такође, разматра се принцип *AI* неутралности и чињеница да елементи субјективне одговорности често испливају, чак и унутар оквира који номинално предвиђају објективну одговорност. Српска доктрина о опасним стварима и делатностима пружа флексибилност судовима да одлучују о штетама изазваним *AI*. Правници морају разумети основе *AI*, укључујући разлике између детерминистичког софтвера и *AI* који показује емергентно понашање. Иако је аргумент за објективну одговорност убедљив, постоје значајни контра-аргументи: оштећени не треба да буде повлашћен само због тога што га је оштетио *AI*, нарочито када је њена примена безбеднија него човек у упоредној ситуацији.

Кључне речи: проблем „црне кутије", одговорност за производе, одговорност по основу кривице, објективна одговорност, штета.

INTRODUCTION

The philosophical origins of AI lie in the ambition to mechanise human reasoning. Aristotle's formalism, which emphasises the validity of certain thought patterns based on their structural form rather than content, has profoundly influenced the field (Arkoudas & Bringsjord, 2014, p. 36). The historical emergence of AI can be traced to the mid-20th century, with the 1956 Dartmouth conference marking its formal inception. However, its conceptual roots are much older, deeply tied to early advances in formal logic and the theory of computation. Turing machines, in particular, provided an essential model for conceptualising how mental processes might be instantiated in physical systems (Arkoudas & Bringsjord, 2014, pp. 39-40). It involves replicating brain processes outside the brain.

AI remains a dynamic and rapidly evolving discipline, with dual aims of constructing intelligent systems and advancing our understanding of cognition. The overarching goal is to create artificial minds, whether inspired by human cognition or by entirely novel forms of intelligence (Frankish & Ramsey, 2014, p. 1). Due to its broad interdisciplinary scope and continuous development, AI resists any single, succinct definition. The EU's AI Act defines an 'AI system' as machine-based, capable of

functioning autonomously and adapting post-deployment. These systems are designed to infer from inputs to generate outputs such as predictions, recommendations, or decisions that affect physical or virtual environments (art. 3). The definition in the AI Act is designed for clarity, international alignment, and adaptability to technological advancements. It explicitly excludes systems that function solely based on predefined rules, thereby ensuring a precise scope that distinguishes AI from conventional software.

AI fundamentally differs from traditional software in terms of decision-making. Traditional software operates through deterministic code, allowing for predictable behaviour. In contrast, AI, particularly deep learning, utilises training data, resulting in an opaque decision-making process. Developers cannot fully anticipate AI behaviour solely by analysing the underlying code (Tai, 2022).

Article 6 of the AI Act establishes criteria for classifying AI systems as 'high-risk' based on their potential impact on health, safety, and fundamental rights. High-risk classifications can occur through two pathways: (1) AI systems that function as safety component of a product or as a standalone product under specific EU safety regulations (Annex I), such as machinery, toys, medical devices, and aviation, which require third-party conformity assessments before market placement; and (2) standalone AI systems used in high-risk domains outlined in Annex III, such as critical infrastructure, essential services, law enforcement, migration, and justice, as well as AI systems involving biometric identification and profiling (Wendehorst, 2022, pp. 198–199). The Act imposes stringent regulations on high-risk systems, focusing on risk management, data quality, transparency, human oversight, and cyber security, aiming to balance innovation with the protection of fundamental rights.

BLACK-BOX AI AND MACHINE LEARNING

Black-box AI refers to artificial intelligence systems whose internal mechanisms are opaque and challenging to understand, even for developers. While their technical architecture may be well-defined, the specific reasoning behind their outputs often remains elusive. This lack of transparency makes it difficult to predict system behaviour and to identify the root causes of damage (Duffourc & Gerke, 2023). Despite these issues, the high level of accuracy achieved by black-box AI remains one of its major appeals.

AI encompasses computational systems capable of performing tasks that require human-like intelligence, including decision-making, learning, and adaptation (Bathaee, 2018, p. 898). Machine-learning algorithms analyse data, identify patterns, and make predictions by adjusting the weights assigned to variables and minimising prediction errors. For

example, an algorithm predicting mile times based on factors such as height, weight, and age iteratively refines these coefficients to reduce errors. This process, called 'training,' aims to ensure the model can accurately predict outcomes when presented with new data (known as 'generalising') (Bathaee, 2018, p. 901). AI developers train systems to generalise based on training data, but they generally lack complete control over both the training data and the output. Full control would defeat the purpose of AI.

Black-box AI primarily relies on deep learning algorithms, which involve training neural networks on extensive datasets to recognise complex patterns. The intricate, multi-layered structure of these networks contributes to their non-interpretability, effectively concealing the underlying decision-making (Duffourc & Gerke, 2023, p. 11). This complicates accountability and erodes trust, particularly in high-stakes fields from Annex III.

The Black Box Problem in machine learning is particularly evident in algorithms such as Deep Neural Networks (DNNs) and Support Vector Machines (SVMs). Inspired by the human brain, DNNs process information in a distributed and intuitive way, similar to how a person instinctively knows how to ride a bike – difficult to explain in a step-by-step fashion (complexity aspect). Similarly, SVMs classify data by finding optimal boundaries in multi-dimensional spaces. Humans struggle to visualise or understand such complex geometric patterns, making their decision-making opaque (dimensionality aspect) (Bathaee, 2018, pp. 901-903).

The distinction between 'strong' and 'weak' black boxes pertains to the transparency of AI. Strong black boxes are completely opaque, making it impossible to understand the conclusions or predict future behaviour. Weak black boxes, on the other hand, allow limited reverse engineering, providing partial insights into variable influences, though the exact reasoning remains unclear. This distinction carries significant legal implications, particularly in areas related to intent and causation (Bathaee, 2018, p. 905).

In Internet of Things (IoT) systems, AI algorithms analyse large volumes of data from interconnected devices to make decisions governing system behaviour (Howells & Twigg-Flesner, 2022, p. 181). These are goal-oriented, designed to achieve objectives such as optimising energy consumption or enhancing manufacturing efficiency. Machine-learning capability allows them to continuously improve performance based on accumulated data (Howells & Twigg-Flesner, 2022, p. 181). Two primary types of AI algorithms are used: Symbolic AI and self-learning algorithms. Symbolic AI is rule-based, offering limited adaptability, whereas self-learning algorithms continuously refine their behaviour over time (Howells & Twigg-Flesner, 2022, p. 192). The complexity of AI algorithms, especially those with self-learning capabilities, poses significant liability challenges. Responsibility is often diffuse among developers, da-

ta providers, and users. The adaptive nature of AI means systems may evolve unpredictably post-deployment, complicating traditional liability that focuses on the product's condition at sale. The concept of network liability proposes treating all stakeholders within an AI-driven IoT system as one, allowing claimants to seek redress from the entire network (Howells & Twigg-Flesner, 2022, pp. 197-198).

AI RELATED DAMAGES

AI is already so widespread that its risks have become nearly unavoidable. In smart homes, AI processes data from temperature, occupancy, and weather sensors to optimise comfort and energy efficiency. In industry, AI analyses production line data and market trends to enhance efficiency, forecast maintenance, and adapt schedules (Howells & Twigg-Flesner, 2022, p. 181). AI-driven IoT systems may malfunction, resulting in security breaches or property damage due to algorithmic errors.

Generative AI, such as Generative Adversarial Networks (GANs) and Large Language Models (LLMs), present unique risks. GANs can create realistic 'deepfakes,' facilitating identity theft or fraudulent content. LLMs are prone to generating inaccurate content - 'hallucinations' that spread misinformation and erode trust. Training datasets, often scraped from the internet, raise copyright and privacy issues. Biases in training data perpetuate social inequities, and continuous updates lead to unpredictable behaviours. AI-generated content used for further training may amplify inaccuracies, creating harmful feedback loops. These risks include discrimination, misinformation, malicious use, and broader social harms, and are caused by AI 'echoes' and 'data drift' (Noto La Diega & Bezerra, 2024, pp. 6–7).

Examples of AI-related harms: (1) medical misdiagnosis – inappropriate treatments due to algorithmic flaws or biases in training data; (2) autonomous vehicle accidents due to errors in algorithms or environmental perception (Monot-Fouletier, 2022, p. 167); (3) erroneous financial decisions leading to significant monetary losses due to processing errors (Tai, 2022, pp. 127-128); and (4) discriminatory practices – AI in hiring, lending, or law enforcement can perpetuate biases, leading to damage to specific groups.

To address these challenges, 'explainable AI' has been developed. It employs simpler, more transparent algorithms, thereby offering some level of explanatory insight. However, these approximations often fall short of capturing the full complexity of the original black-box models. As the deployment of black-box AI expands, it is imperative that ongoing research focuses on enhancing transparency (Duffourc & Gerke, 2023, p. 12).

THE SERBIAN DOCTRINE ON DANGEROUS OBJECTS AND ACTIVITIES

Understanding *osnov odgovornosti* (*Eng.* the basis of liability) is crucial for assigning responsibility, and shaping legal remedies and the success of a compensational claim. Additionally, it determines the burden of proof and the victim's advantage over the tortfeasor, impacting the overall outcome of the case. *Osnov odgovornosti* refers to the legal justification for imposing liability on a party, distinct from *uslovi odgovornosti* (*Eng.* conditions of liability), which include damage, causality, culpability and wrongfulness. In Serbian legal theory wrongfulness is disputed. Some authors argue that there is a rebuttable legal presumption that any act causing harm to another is wrongful. Others, however, contend that the wrongfulness of a damaging act is not a requirement for establishing liability under the Act on Obligations (ZOO) but is instead subsumed within objectively understood fault. (Karanikić Mirić, 2024, p. 503). The Serbian legislator does not mention wrongfulness (Karanikić Mirić, 2024, p. 510), and domestic courts: (1) do not require proof of wrongfulness as a fourth condition for establishing liability (in addition to damage, causation, and fault) and (2) do not allow the tortfeasor to be exempted from liability by proving that the act causing harm was not unlawful (Karanikić Mirić, 2024, p. 655).

Serbian law recognises several bases or regimes: fault-based liability, strict liability, and equity-based liability which serves as a corrective. Fault-based liability is tied to blameworthy conduct, such as negligence or intentional wrongdoing. Strict liability, by contrast, applies irrespective of fault, holding entities liable due to inherent risks. Judicial interpretation significantly shapes the scope of strict liability, especially in defining dangerous activities and objects under strict liability.

The concepts of *opasna stvar* (*Eng.* dangerous object) and *opasna delatnost* (*Eng.* dangerous activity) are open-list legal standards (Karanikić Mirić, 2017, p. 353). A 'dangerous object' inherently poses an elevated risk of harm due to its nature or specific use, evaluated by a 'reasonable and careful person.' This includes inherently dangerous items like explosives and objects that become hazardous in certain contexts, such as a poorly positioned ladder. Judicial analysis assesses the nature, use, and context of these items, with examples like buildings, weapons, elevators, and manholes commonly cited. *Opasna delatnost* refers to activities that carry a heightened risk of harm, even when performed with utmost care. Examples include construction work, logs unloading or launching anti-hail rockets. Context matters, as certain activities can become dangerous depending on the circumstances, such as serving food at a crowded event. Additionally, how the activity is organised, such as transporting valuable goods without security, can make it *opasna delatnost*. As many dangerous activities involve dangerous objects, there is overlap but distinctions

exist. Some objects are inherently dangerous regardless of use, while certain activities are risky without involving dangerous items. Dangerous thing or activity trigger strict liability (art. 173-179, Act on Obligations (ZOO)).

The concept of strict liability emerged during the Industrial Revolution, when traditional fault-based liability was inadequate for addressing accidents involving machinery and hazardous substances, even with due care, focusing solely on the causal link between the dangerous activity or object and the harm. This shifts the burden of proof to the defendant, who must disprove causation (art. 173 ZOO). Strict liability serves multiple purposes: risk allocation, where costs are assigned to those engaging in inherently dangerous activities; simplified victim compensation by reducing evidentiary burdens; and risk socialisation, encouraging broader distribution of costs through insurance. It applies in areas such as product liability and animal ownership. Compared to negligence, strict liability provides a more streamlined route to compensation when proving fault is impractical. It holds individuals or entities accountable for harm resulting from hazardous activities or risky products, fixing on the harm caused rather than the mental state of the defendant. The shift enabled courts to ensure compensation for victims, stressing the principle that those who benefit from dangerous enterprises should bear the risk, promoting fairness and societal responsibility.

STRICT LIABILITY AND AI

Justifications of Strict Liability

Strict liability is advantageous for governing high-risk AI systems. It internalises the costs of harm, creating incentives for developers and operators to prioritise safety through rigorous testing, high-quality data, and effective oversight (Howells and Twigg-Flesner, 2022, pp. 193-194). Since AI systems are often complex and opaque, proving negligence is challenging, and strict liability bypasses this requirement.

Strict liability encourages the use of AI in socially beneficial ways and deters harmful applications by imposing significant liability costs. Rooted in the economic analysis of law, this framework also provides predictability, allowing companies to understand their obligations and foster responsible innovation (Heiss, 2020, p. 206). Strict liability addresses issues that arise with negligence and product liability regimes. Due to the 'black-box' nature of AI, proving negligence is often impractical. While product liability may apply to AI embedded in hardware, it is less suitable for software or multi-party systems. Strict liability thus offers a more comprehensive approach (Heiss, 2020, p. 203).

Strict liability streamlines the legal process for AI-related claims, making outcomes more predictable, and building trust in AI. This frame-

work aligns with the nature of AI and the 'do no harm' principle, promoting ethical use (Noto La Diega & Bezerra, 2024, p. 16, 17, 21). Ultimately, it can reduce compliance costs, provide economic benefits, and enhance legal certainty.

AI's unpredictability complicates risk assessments. Strict liability is typically used for inherently dangerous activities, holding parties accountable for harm regardless of intent. However, the 'Black Box Problem' introduces unpredictability that challenges effective risk management. The opaque nature of machine-learning algorithms makes it difficult to predict behaviours, undermining fault-based liability. High-frequency trading algorithms, for instance, have triggered unintended market consequences despite careful design (Bathaee, 2018).

To address these challenges, a harmonised legal framework centred on strict liability is proposed for AI-related harms, preserving tort law's role in regulating autonomous systems (Noto La Diega and Bezerra, 2024). The authors criticise the AI Liability Directive's (AILD) reliance on fault-based models as inadequate for generative AI and autonomous agents. They argue for EU-wide harmonisation under strict liability to streamline victim compensation, incentivise safety measures, and foster public trust (Noto La Diega & Bezerra, 2024, p. 2). It should be noted that the arguments justifying strict liability for AI largely overlap with those for applying the same liability regime to dangerous objects and activities.

Arguments Against Strict Liability in AI Related Damage

Strict liability for AI presents significant challenges that may impede innovation and burden smaller entities. Imposing it without considering fault could discourage startups and smaller companies from developing AI due to fears of crippling liability for unforeseen harms. This could lead to a concentration of AI development among large corporations, stifling diversity and limiting innovation (Bathaee, 2018, p. 896). Liability caps for SMEs may help mitigate these concerns (Noto La Diega & Bezerra, 2024, p. 19). Monopolism in such a critical area is daring.

Strict liability disregards negligence or due care, potentially reducing incentives for developers to follow safety standards. The blanket imposition of liability for unforeseen AI consequences might undermine responsible development (Bathaee, 2018, p. 932). Adapting existing legal concepts like intent and causation (Cvetković, 2020) for AI may be a better solution.

Applying strict liability to foundation models presents challenges, particularly due to their broad range of applications, some of which are high-risk while others are not. Holding providers strictly liable for all harms could be unfair and impractical. Allowing defences like force majeure or unforeseeable events would provide a more balanced framework (Noto La Diega & Bezerra, 2024, p. 19). AI foundation models are

versatile, large-scale models trained on extensive data, serving as adaptable infrastructure for specific applications like language processing, image recognition, or decision-making (custom, specialised AI).

The principle of AI neutrality suggests that strict liability should not apply if an AI system consistently shows superior safety and performance compared to humans performing the same task, especially when humans are not held to a strict liability standard. For example, autonomous vehicles that cause fewer accidents than human drivers should not face strict liability, as it could hinder the adoption of life-saving technologies (Barbosa & Valadares, 2023, p. 154). Disparities in treatment between AI-driven and human-operated devices raise fairness issues. In medical contexts, for example, AI-assisted robotic systems could face strict liability for patient harm, whereas human surgeons would be evaluated based on negligence. Such discrepancies could lead to unequal compensation for similar harms, depending on whether AI or human actions were involved. In Serbian law, primarily and in most cases, liability for this damage will not be attributed to the attending physician but rather to the healthcare institution, following the rules on "Employer liability for damage caused by an employee during or in connection with their work" (ZOO, art. 170-171), with a predetermined standard of care.

Examples unsuited for strict liability include scenarios like pure economic atypical risks, where, for instance, a software agent inadvertently lowers a user's credit score. In such cases, the causal link between the software's actions and the economic harm is complex and indirect, making it challenging to establish a direct and foreseeable connection (Wendehorst, 2020, pp. 162–164). Similarly, social atypical risks—such as a spouse's excessive online gaming leading to the breakup of a marriage — demonstrate outcomes that stem from individual behaviour rather than any inherent risk by the technology itself. Holding developers responsible would stretch the boundaries of legal responsibility (Wendehorst, 2020, pp. 162-164). In Serbian law, there would be no liability here because the causal link is not adequate, regardless of whether the liability is strict or fault-based. The basis of liability cannot be considered in isolation.

A blanket application of strict liability even to all 'high-risk' AI systems is often excessive. Not all systems classified as high-risk pose the same level of danger (Arsenijević, 2023, p. 147). For instance, small robotic vacuum cleaners are far less risky compared to large industrial robots. Fairness also demands that similar devices operated by humans and AI should be subject to consistent liability standards. Aligning liability with the inherent risk profile of the device, rather than the technology used, would rectify these inconsistencies. Factors like device size, speed, and environment (for devices functioning in public areas or near vulnerable populations) should determine liability to ensure a fair and consistent framework, providing equal protection and compensation irrespective of

the device's autonomy level (Wendehorst, 2022, p. 206). These factors are very similar to those used in the Serbian legal doctrine for defining a dangerous object. According to art. 173 of the Act on Obligations (ZOO), when damage is caused by a dangerous object – and if AI is argued to fall under this category – causation is presumed precisely to ease the burden of proof for the injured party. This presumption is based on the reasonable expectation that the defendant is in a better position to prove that the causal link does not exist.

A NUANCED APPLICATION OF STRICT LIABILITY

The authors propose tailored applications of strict liability to address the unique challenges posed by AI systems.

1. 'No-Fault Compensation' for AI suggests replacing traditional tort law with no-fault schemes. In France, 'socialisation des risques' shifts compensation from individuals to the collective, such as social security, insurance, and dedicated funds. This aligns with solidarité nationale, emphasising society's duty to protect individuals from uncontrollable risks. AI complexity often makes proving fault impossible, leaving victims without recourse. No-fault schemes offer accessible compensation and encourage AI innovation by protecting developers. However, funding and moral hazard concerns remain unresolved, as developers might deprioritise safety without liability pressures (Knetsch, 2022, p. 113). Similar 'No-Fault' regimes exist in medical law regarding liability for medical malpractice.

2. 'Strict Liability with Comparative Negligence' is proposed for cases involving a high-risk AI and another party, such as humans or non-AI systems. This model ensures shared responsibility, with strict liability for the AI operator balanced by a comparative negligence defence (Heiss, 2020, p. 210). The possessor of a dangerous object is partially exempt from liability if the injured party partially contributed to the damage (art. 177-3 ZOO).

3. Strict liability should always apply to AI causing human rights violations. Given AI's unpredictability and 'black box' nature, proving causation is often infeasible, making strict liability necessary. This approach would incentivise developers to embed human rights considerations throughout the AI lifecycle (Barbosa & Valadares, 2023, p. 156) and align with legal trends imposing strict liability for distressing fundamental rights. The Serbian Anti-Discrimination Act stipulates that: "If the court has determined that an act of direct discrimination has occurred, or if this is undisputed between the parties, the defendant cannot be exempted from liability by proving the absence of fault" (art. 45-1; Tasić, 2018).

4. 'Strict Liability to the State' is suggested for incidents involving multiple high-risk AI systems. Instead of compensating individual victims directly, the liable AI operator would pay the state, which would then

compensate victims through insurance. A blanket fee system based on AI type could streamline processes. Enforcing accident reporting would rely on automated systems, using sensors and data recording within high-risk AI system itself. (Heiss, 2020)

5. The authors advocate for a strict liability for personal injury and death caused by AI. Here, the severe consequences justify shifting from the fault-based regime. As AI becomes embedded in medical, safety, and transport applications, avoiding it becomes challenging, necessitating a liability regime that transfers risk to developers and operators. Significant harm warrants stronger protection. Extending strict liability from defective products to AI ensures fairness, as both present comparable risks (Soyer & Tettenborn, 2022).

6. AI vehicle accidents pose unique challenges under traditional custodian liability, typically applied to tangible objects under human control. Extending custodian liability to AI vehicles is problematic due to the ambiguous classification of AI systems. It is unclear whether an AI system, made up of software and algorithms, qualifies as an 'object' under traditional custodian liability. While courts have sometimes classified software viruses as objects, whether this logic applies to the complex, evolving algorithms remains uncertain. Another challenge lies in identifying the custodian. Autonomous vehicles involve multiple parties, each potentially responsible for different aspects of the technology. For instance, the designer or manufacturer could be the custodian of the AI system's structure, while the user or maintenance manager might be responsible for the system's behaviour. The driver, despite a reduced role in autonomous driving, could also retain some custodial responsibility (Monot-Fouletier, 2022, p. 170). Traditional custodian liability is inadequate for AI vehicle accidents, necessitating a re-evaluation of legal concepts like 'control' and 'object.' Within the 'subject–object' dichotomy, this means that if AI systems are not objects, then they must be subjects. Arguments regarding granting legal subjectivity to AI include several perspectives. The subjectivity of AI could mean that the AI itself would be liable for damages it causes (Pavlekovic & Petrovic, 2021, p. 119). Some propose the creation of a new category of legal subject — an electronic party (ePerson). Legal subjectivity would be acquired through registration and would be appropriate to the extent of the rights and obligations that AI, as a creation of law, can bear. Legal subjectivity is a political decision of the legal system, as the fact that corporations are recognised as legal persons is not a 'natural state of affairs,' but rather a matter of legislative regulation. Additionally, the legal fiction that once classified slaves as property could, in theory, be repurposed to redefine AI as a legal person (Arsenijević, 2023, p. 141). The far-reaching nature of this idea exceeds the scope of this paper.

7. 'Determining Liability Based on the Type of Harm Caused' – AI harms can be categorised into three principal types: Physical Risks, Pure Economic Risks, and Social Risks.

Physical Risks encompass traditional safety concerns such as bodily injury, death, and damage to property. They also extend to harm involving data, interference with other digital systems, and psychological harm meeting clinical criteria. Strict liability is particularly appropriate here. Physical harm can be quantified objectively, providing a clear basis for compensation. Moreover, societal interests in protecting health and property are paramount, making strict liability justified (Wendehorst, 2020, pp. 165-166).

Pure Economic Risks refer to financial losses unconnected to physical damage, such as broken AI financial recommendations (Wendehorst, 2020, p. 161). Strict liability is less suitable here due to the complexities of causation and the subjective nature of economic loss. Expanding strict liability to these areas risks overwhelming the legal system with litigation. Instead, the non-compliance liability, where parties are liable for breaching predefined standards, is more pragmatic (Wendehorst, 2020, p. 156).

Social Risks, described as fundamental rights risks, include harms such as discrimination, manipulation, and violations of privacy and dignity (Wendehorst, 2020, p. 162). These arise in contexts like biased hiring algorithms or behaviour-modifying social media algorithms. Given the intangible nature of these harms and their resistance to monetary quantification, strict liability is inadequate. Therefore, such social risks should instead be mitigated through specialised legal regimes, including data protection laws, anti-discrimination statutes, and legislation against hate speech and harassment. (Wendehorst, 2020, p. 162). Above in the text, we saw precisely the opposite proposal in (Barbosa & Valadares, 2023).

8. 'Determining Liability Regimes Based on a Risk Type' – direct or general risks involve immediate harm caused by the AI system, such as a malfunctioning cleaning robot injuring a pedestrian, which suits strict liability due to the clear causal link. Intermediated typical risks entail a step between the AI and the harm, like medical software issuing incorrect recommendations leading to health issues; these can be addressed under strict liability if robust defences are available (Wendehorst, 2022, p. 170). Intermediated general risks involve complex causal chains where AI indirectly causes harm, such as a vulnerability in a smart heating system facilitating a burglary, which should only be covered by strict liability with substantial defences. Finally, atypical risks, such as a robot's sharp handle causing unpredictable injury, fall outside reasonable foreseeability and should not be governed by strict liability (Wendehorst, 2022, pp. 162-164). The problem with the latter two approaches (7 and 8), where the strict liability regime depends on risk classification, lies in the fact that these classifications are not generally accepted. Even the author himself highlights a classification based on the *type of harm* in one paper, while in another paper, he proposes a different classification based on the frequency and adequacy of a given risk.

THE PROPOSED EU REGIME

The AILD Proposal harmonises procedural aspects such as evidence disclosure and the burden of proof, aligning with fault-based liability. The Product Liability Directive Proposal (PLD) modernises product liability by explicitly including software and AI systems, imposing strict liability. This dual approach aims to balance fault-based and strict liability framework.

A key problem is the contradiction between the fault-based AILD Proposal and the supposedly strict liability in PLD Proposal. Although the PLD is formally based on strict liability, in practice it often requires proving a breach of duty, such as failing to address biases in AI training data. This reliance on fault-based reasoning creates an artificial distinction, blurring the lines between two regimes. As a result, the separation of the directives into distinct frameworks undermines clarity and coherence (Hacker, 2023, p. 29). Additionally, there are no provisions for harm caused by prohibited AI. Hacker advocates for true harmonisation by merging two proposals (Hacker, 2023, p. 49). He calls for the expansion of strict liability to cover certain high-risk AI systems, particularly those causing "illegitimate harms," regardless of whether they comply with the technical requirements of the AI Act (Hacker, 2023, p. 30-31).

What follows is an illustration. An AI facial recognition system incorrectly identifies Ms. Smith as a robbery suspect, leading to her assets being frozen and causing severe financial harm. The error is traced to a known gender imbalance in the training data, which both the AI developers (SmartView Ltd.) and the bank were aware of. Under the AILD Proposal, Ms. Smith can request evidence from SmartView and the bank. If they fail to comply, a presumption of non-compliance is triggered, bolstering her case (Hacker, 2023, pp. 21-22). Proving fault remains a challenge, requiring Ms. Smith to show the bank breached its duty, potentially by violating data governance rules under Article 10(3) of the AI Act. This necessitate hiring AI experts to demonstrate that the gender imbalance in the training data was negligent.

Just as the classical tort law face challenges, so too must the major concepts of consumer protection law adapt when AI is involved. It is difficult to define 'defect' in AI, as harm can arise from design or algorithmic outcomes, not just manufacturing flaws. The 'Development Risk Defense' may not apply, as developers are aware of AI's inherent risks. (Knetsch, 2022, p. 111). Under the PLD Proposal's strict liability, Ms. Smith could request evidence from SmartView, and failure to comply would lead to a presumption of defectiveness (Hacker, 2023, p. 26). If evidence is provided, expert testimony would be needed to prove that the gender imbalance violated Article 10(3), thus establishing defectiveness. However, the PLD limits damages to property, life, or health, potentially excluding Ms. Smith's claims for pure economic loss (Hacker, 2023, p.

27). Furthermore, her case may not fit neatly into a consumer protection context, because she was a client of the bank, not a customer of SmartView. Moreover, in Serbian law, under the current consumer protection, liability for defective products cannot be applied if damage is caused by 'software only' AI (hardware-software systems are covered) (Arsenijević, 2023, p. 154).

Proving fault or defectiveness with sophisticated AI models remains difficult. The interplay between strict and fault-based liability shows that fault-based reasoning often re-emerges within supposedly strict liability frameworks like the PLD. In Ms. Smith's case, proving defectiveness requires demonstrating that the gender imbalance constituted a breach of duty, effectively merging strict liability with fault-based elements (Hacker, 2023).

Artificial intelligence systems cannot simply be reduced to a dangerous object. Even when this classification is possible, such as in the case of an autonomous vehicle or robot, the question arises as to who is liable for damages when the system's creator no longer has any ability to predict its outcomes or control the directions in which it will autonomously evolve. It appears that the creation of artificial intelligence systems should be classified as a dangerous activity (as in ZOO), with liability assigned to the entity that primarily derives economic benefits from that activity. However, a key issue remains: how to determine economic benefits, particularly when the system is freely accessible to an unlimited number of users who pay with their personal data. The entities that fall within the scope of potentially liable parties include: the AI itself (if granted legal subjectivity); the AI owner; developers; participants in the AI's creation and control; AI Product or Device Manufacturer; the AI user or operator; the Distributor, Vendor, or AI Service Provider; the Economic Beneficiary of the AI System, State and Regulatory Authorities.

CONCLUSION

Culpa represents a psychological phenomenon; a person is deemed at fault due to the intention behind tearing a book, rather than the mere physical action of doing so. AI, which operates with a certain degree of autonomy, might create the impression that elements of fault are present, akin to human-like reasoning. However, this does not translate to actual fault within the context of liability. While AI may exhibit decision-making abilities that evoke notions of fault, these should not be misread as akin to human intentionality. The attribution of human-like qualities to AI, stemming from anthropomorphising tendencies, fails to recognise the fundamental distinction: AI decisions are grounded in formal logic based on structural form rather than subjective content. Emotional or psychological elements are entirely absent. Thus, when considering liability for AI-induced harm, fault-based liability becomes untenable, leaving strict liability as more suitable.

Historically, the emergence of modern strict liability was closely tied to the Industrial Revolution, which necessitated a legal response to the increased risks introduced by rapid technological advancements. Today, as we face a new digital revolution, scholars again call for the adoption of strict liability to address the unique challenges presented by AI, particularly its opaque 'black-box' nature. Despite the utmost care in development and deployment, AI systems can cause harm, and their inherent complexity often renders it impossible to establish causation or fault. The black-box nature of these systems impedes the identification of responsible parties, as the causal chain is often obscured and, in many instances, fault is absent altogether.

In Serbian law, the open-ended concepts of dangerous things and activities provide courts with the flexibility to adjudicate AI-related damages under strict liability. This is especially applicable for AI classified as high-risk under the EU AI Act. Advanced algorithms such as self-learning systems, Deep Neural Networks, and Support Vector Machines, especially those exhibiting a 'strong black-box,' provide grounds for addressing cases under strict liability. Such an approach is more favourable to the injured party, as it establishes liability irrespective of fault and shifts the burden of proving causation away from the victim. These same advantages are reflected in both PLD and AILD.

The Serbian Act on Obligations (ZOO) offers adequate civil protection to victims of AI-related harm, provided that the judiciary is competent to articulate why a particular AI constitutes a dangerous entity or activity. This determination is a legal question, not a matter for expert witnesses, underscoring the importance of theoretical work in this area for both educational and practical purposes. Judges must grasp the conceptual differences between classical deterministic software and AI that manifests emergent, non-deterministic behaviour. Understanding the role of AI inference and its inherent unpredictability post-deployment is essential. Furthermore, judges should be familiar with the business models underpinning AI systems, particularly contractual relationships between AI developers, manufacturers, and users, as well as the dynamics between foundational and specialised, fine-tuned AI models.

Nevertheless, strict liability is not a universal remedy, and its application should not be unduly broadened, as this could yield adverse consequences. Overextending risks stifling innovation, deterring small enterprises from entering the AI market, and imposing disproportionate burdens on developers who may have limited control over an AI's evolving behaviour. Although the case for strict liability is compelling, there are substantial counterarguments, particularly the principle of AI neutrality. Transitional and nuanced solutions, such as linking liability to specific types of risk or adopting a sliding scale approach, are suitable to strike a balanced legal framework. In certain cases, strict liability is inappropriate,

prompting scholars to suggest specialised liability regimes akin to anti-discrimination or privacy protection laws. However, such regimes are not without their own limitations, as they often involve strict liability as well.

Even in instances where strict liability is embedded in legislative instruments such as the PLD, elements of fault-based liability tend to surface in practical applications, especially when proving a breach of duty becomes necessary. Strict liability is, therefore, not a permanent or comprehensive solution, particularly as advancements in 'explainable AI' may provide more transparency and accountability in the future.

ACKNOWLEDGEMENTS: *This research has been financially supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia (Contract No. 451-03-137/2025-03/200120 from 04.02.2025). SDG: 9,10,11,16,12.)*

REFERENCES

Arkoudas, K., & Bringsjord, S. (2014). Philosophical foundations. In W. M. Ramsey & K. Frankish (Eds.), *The Cambridge handbook of artificial intelligence* (pp. 34–63). Cambridge: University Press. doi: 10.1017/CBO9781139046855.004

Arsenijević, B. (2023). Odgovornost za štetu od veštačke inteligencije [Liability for Damage Caused by Artificial Intelligence]. In: Petrović, Z., Čolović V., Obradović D. (Ed.): *XXVI International scientific conference - Causation of Damage, Damage Compensation and Insurance* (135-155). Beograd, Valjevo: The Institute of Comparative Law, The Association for Tort Law and Judicial Academy. doi: 10.56461/ZR_23.ONS.08

Barbosa, F., & Valadares, L. (2023). Artificial intelligence: A claim for strict liability for human rights violation. *Revista de Direito Internacional*, *20*(2), 149-158. doi: 10.5102/rdi.v20i2.9119

Bathaee, Y. (2018). The Artificial Intelligence Black Box and the Failure of Intent and Causation. *Harvard Journal of Law & Technology (Harvard JOLT)*, *31*(2), 889–938.

Cvetković, M. (2020). Causal Uncertainty: Alternative Causation in Tort Law. *Teme-Časopis Za Društvene Nauke*, *44*(1), 33–47. doi: 10.22190/TEME191115007C

Duffourc, M., & Gerke, S. (2023). Decoding U.S. Tort Liability in Healthcare's Black-Box Al Era: Lessons from the European Union. *Stanford Technology Law Review*, *27*(1), 1-70.

Frankish K. & Ramsey M. (Eds.). (2014). *The Cambridge Handbook of Artificial Intelligence*. Cambridge: University Press. doi: 10.1017/CBO9781139046855

Hacker, P. (2023). The European AI liability directives–Critique of a half-hearted approach and lessons for the future. *Computer Law & Security Review*, *51*, 1-42. doi: 10.1016/j.clsr.2023.105871

Heiss, S. (2020). Towards Optimal Liability for Artificial Intelligence: Lessons from the European Union's Proposals of 2020. *Hastings Sci. & Tech. LJ*, *12*, 186-224.

Howells, G., & Twigg-Flesner, C. (2022). Interconnectivity and Liability: AI and the Internet of Things. In C. Poncibò, L. A. DiMatteo, & M. Cannarsa (Eds.), *The Cambridge Handbook of Artificial Intelligence: Global Perspectives on Law and Ethics* (pp. 179–199). Cambridge: University Press. doi: 10.1017/9781009072168.019

Karanikić Mirić, M. (2017). General Clause on Strict Liability in Comparative Perspective. In B. Milisavljević, T. Petrović Jevremović, & M. Živković (Eds.), Law and Transition. Collection of Papers, Belgrade (pp. 345–356).

Karanikić Mirić, M. (2024). Obligaciono pravo (2. izd.) [Law of Obligations]. Beograd: Službeni glasnik. https://plus.cobiss.net/cobiss/sr/sr_latn/bib/147456265

Knetsch, J. (2022). Are Existing Tort Theories Ready for AI?: A Continental European Perspective. In C. Poncibò, L. A. DiMatteo, & M. Cannarsa (Eds.), *The Cambridge Handbook of Artificial Intelligence: Global Perspectives on Law and Ethics* (pp. 99–115). Cambridge: University Press. doi: 10.1017/9781009072168.013

Monot-Fouletier, M. (2022). Liability for Autonomous Vehicle Accidents. In C. Poncibò, L. A. DiMatteo, & M. Cannarsa (Eds.), *The Cambridge Handbook of Artificial Intelligence: Global Perspectives on Law and Ethics* (pp. 163–178). Cambridge: University Press. doi: 10.1017/9781009072168.018

Noto La Diega, G., & Bezerra, L. C. (2024). Can there be responsible AI without AI liability? Incentivizing generative AI safety through ex-post tort liability under the EU AI liability directive. *International Journal of Law and Information Technology*, *32*(1), 1-21. doi: 10.1093/ijlit/eaae021

Pavlekovic, B., & Petrovic, J. (2021). Civil Law Aspects of Artificial Intelligence in Medicine. *Pravni letopis*, *1*, 103–124.

Proposal for a Directive of the European Parliament and of the Council on Adapting Non-Contractual Civil Liability Rules to Artificial Intelligence (AI Liability Directive) (2022).

Proposal for a Directive of the European Parliament and of the Council on Liability for Defective Products (2022).

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying down Harmonised Rules on Artificial Intelligence and Amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (2024).

Soyer, B., & Tettenborn, A. (2022). Artificial intelligence and civil liability—Do we need a new regime? *International Journal of Law and Information Technology*, *30*(4), 385–397. doi: 10.1093/ijlit/eaad001

Tai, E. T. T. (2022). Liability for AI Decision-Making. In C. Poncibò, L. A. DiMatteo, & M. Cannarsa (Eds.), *The Cambridge Handbook of Artificial Intelligence: Global Perspectives on Law and Ethics* (pp. 116–131). Cambridge: University Press. doi: 10.1017/9781009072168.014

Tasić, A. (2018). Терет доказивања у антидискриминационим парницама на примеру одлуке Врховног касационог суда [Burden of Proof in Anti-Discrimination Lawsuits: An Example from a Supreme Court of Cassation Decision]. *Зборник Радова Правног Факултета у Нишу*, *57*(78), 325–336. doi:10.5937/zrpfni1878323T

Wendehorst, C. (2020). Strict Liability for AI and other Emerging Technologies. *Journal of European Tort Law*, *11*(2), 150–180. doi: 10.1515/jetl-2020-0140

Wendehorst, C. (2022). Liability for Artificial Intelligence: The Need to Address Both Safety Risks and Fundamental Rights Risks. In O. Mueller, P. Kellmeyer, S. Voeneky, & W. Burgard (Eds.), *The Cambridge Handbook of Responsible Artificial Intelligence: Interdisciplinary Perspectives* (pp. 187–209). Cambridge: University Press. doi: 10.1017/9781009207898.016

Zakon o obligacionim odnosima [Act on Obligations], Sl. list SFRJ. br. 29/78, 39/85, 45/89 - odluka USJ, 57/89. Sl. list SRJ. br. 31/93. Sl. list SCG. br. 1/2003 - Ustavna povelja. Sl. glasnik RS. br. 18 (2020)

Zakon o zabrani diskriminacije [Act on the Prohibition of Discrimination], Sl. glasnik RS. br. 22 (2009). 52 (2021)

"BLACK BOX" ВЕШТАЧКЕ ИНТЕЛИГЕНЦИЈЕ КАО РАЗЛОГ ЗА ОБЈЕКТИВНУ ГРАЂАНСКОПРАВНУ ОДГОВОРНОСТ

Михајло Цветковић
Универзитет у Нишу, Правни факултет, Ниш, Србија

Резиме

Објективна одговорност је начелни оквир за регулисање одговорности за вештачку интелигенцију (*AI*) која делује аутономно, нарочито када су процеси одлучивања непрозирни и сложени. *AI* системи не поседују емоционалне или психолошке елементе, већ се њихове одлуке заснивају на формалној логици без људске намере, због чега одговорност заснована на кривици није прикладна. Појава објективне одговорности била је правни одговор на ризике индустријске револуције, а данас, у дигиталној револуцији, поново је релевантна због непредвидивости и сложености *AI* система. Чак и уз највећу бригу у развоју и имплементацији, *AI* може изазвати штету, а због комплексности често није могуће идентификовати одговорну страну. Судови у Србији могу штету изазвану вештачком интелигенцијом третирати као последицу опасне ствари или делатности, омогућавајући тако примену објективне одговорности, што пружа бољу заштиту оштећенима. Напредни алгоритми, као што су самоучећи системи и дубоке неуронске мреже са карактеристикама „црне кутије", захтевају објективну одговорност, као и пребацивање терета доказивања на штетника. Закон о облигационим односима (ЗОО) омогућава заштиту оштећенима, уз услов да судије разумеју разлике између традиционалног софтвера и *AI* система са емергентним понашањем. Правници морају упознати уговорне односе програмера, произвођача и корисника *AI*, као и основне техничке карактеристике *AI* модела како би одлучивали о правним питањима, уместо да све зависи од вештака. Разумевање *AI* инференције и инхерентне непредвидивости након имплементације је од суштинског значаја. Међутим, прекомерна примена објективне одговорности може угушити иновације и негативно утицати на мала предузећа, обесхрабрујући инвестиције у *AI*. Иако је аргумент за објективну одговорност убедљив, постоје значајни контра-аргументи, посебно принцип неутралности: оштећени не треба да буде повлашћен само због тога што га је оштетио *AI*, нарочито када је *AI* безбеднији него човек у упоредној ситуацији. Потребно је пронаћи баланс између одговорности и иновација. Прелазна нијансирана решења, као што је повезивање режима одговорности са специфичним врстама ризика или усвајање клизне скале, од суштинског су значаја за постизање уравнотеженог правног оквира.