IJBMC
International Journal of Body, Mind and Culture

Volume 13, Issue 4, pp 235-248
DOI: 10.61838/ijbmc.v13i4.1303

Article type:
Original Research

The Application of the Theory of Moral Blameworthiness in Psychology to the Determination of Punishment for Crimes Arising from Artificial Intelligence Systems

Reza. Farajpour[1], Seyed Mostafa. Mahmoudi Moghadam[1]*, Dariush. Babaei[1], Hamidreza. Rostami[1], Mohammad Javad. Kiani[2]

1 Department of Law, Yas.C., Islamic Azad University, Yasuj, Iran.
2 Department of Electrical Engineering, Yas.C., Islamic Azad University,Yasuj, Iran.

Corresponding author email address:
pmam1397@iau.ac.ir

Article history:

Received 23 Jan 2026
Revised 28 Feb 2026
Accepted 12 March 2026
Published online 01 Apr 2026

How to cite this article:
Farajpour, R., Mahmoudi Moghadam, S. M., Babaei, D., Rostami, H., & Kiani, M. J. (2026). The Application of the Theory of Moral Blameworthiness in Psychology to the Determination of Punishment for Crimes Arising from Artificial Intelligence Systems. International Journal of Body, Mind and Culture, 13(4), 235-248.

ABSTRACT

Objective: This study aimed to examine how the theory of moral blameworthiness can be applied to determining criminal responsibility and punishment for crimes arising from artificial intelligence systems.

Methods and Materials: A descriptive–analytical and interdisciplinary methodology was used. The study drew on moral psychology, philosophy of responsibility, criminal law theory, AI ethics, and contemporary AI governance frameworks. Academic literature, legal instruments, policy documents, and regulatory materials were analyzed, including sources related to moral responsibility, blame attribution, AI autonomy, the responsibility gap, and risk-based AI governance.

Findings: The analysis showed that current AI systems should not be treated as fully blameworthy moral or criminal agents because they lack consciousness, moral understanding, free will, and genuine responsiveness to moral reasons. However, AI-related harm does not eliminate responsibility. Responsibility should be assessed across the human and institutional chain surrounding the AI system, including designers, developers, data providers, model trainers, deploying institutions, corporate managers, users, and regulators. The study identified knowledge, control, foreseeability, preventability, and risk creation as the main criteria for assessing blameworthiness. Traditional criminal law concepts such as mens rea, negligence, recklessness, causation, and culpability remain necessary but require reinterpretation in light of algorithmic autonomy, opacity, distributed decision-making, and organizational responsibility.

Conclusion: Punishment for AI-related crimes should be proportionate to morally relevant culpability rather than mere causal involvement. A chain-of-responsibility model can help prevent both unjust impunity and unjust over-punishment.

Keywords: Moral Blameworthiness, Artificial Intelligence, Criminal Responsibility, Punishment, Moral Psychology.

Introduction

Artificial intelligence has rapidly moved from a specialized field of computer science into the core infrastructures of contemporary social, economic, medical, military, and legal life. AI systems now assist or replace human decision-making in areas such as medical diagnosis, autonomous transportation, predictive policing, financial risk assessment, facial recognition, judicial decision support, military targeting, and automated content moderation. This technological expansion has created not only technical and regulatory challenges but also a deeper conceptual problem: when an AI system contributes to harmful or criminal conduct, who is morally blameworthy, and on what basis should criminal responsibility and punishment be determined? (AI, 2023; Instruments, 2019; Matthias, 2004).

This question is especially difficult because modern criminal law and moral psychology have traditionally been organized around human agency. Classical criminal liability presupposes a responsible subject who acts with some combination of intention, knowledge, recklessness, negligence, or culpable omission. Similarly, theories of moral blameworthiness usually rely on ideas such as control, awareness, intentionality, capacity for moral understanding, and responsiveness to reasons. Fischer and Ravizza's account of moral responsibility, for example, emphasizes guidance control and reasons-responsiveness as central conditions for responsibility, while psychological theories of blame show that blame judgments depend on how observers evaluate causation, intentionality, obligation, knowledge, and the agent's capacity to have acted otherwise (Fischer & Ravizza, 1998; Malle et al., 2014).

AI systems disturb these assumptions because their harmful outputs may be causally connected to human design, data, training, deployment, or supervision, while the system itself lacks ordinary human mental states. A human offender can usually be assessed in terms of intention, knowledge, awareness of risk, and capacity for self-control. By contrast, present AI systems do not possess consciousness, moral understanding, or free will in the ordinary human sense. Even when an AI system appears to make autonomous decisions, it normally does so through statistical inference, optimization, pattern recognition, or programmed objectives rather than through moral deliberation. Searle's critique of strong artificial intelligence remains relevant here because it challenges the idea that the formal manipulation of symbols or data is sufficient for genuine understanding or intentionality (Searle, 1980).

The central issue, therefore, is not simply whether AI caused a harmful result, but whether blame can meaningfully be attributed to the AI itself or must instead be distributed among the human and institutional actors who designed, trained, deployed, supervised, marketed, or used the system. Moral psychology shows that people distinguish between causal responsibility and blameworthy agency: an entity may cause harm without being morally blameworthy if it lacks intention, knowledge, control, or the capacity to understand the relevant norm. This distinction is particularly important in AI-related crimes because the AI system may be causally central to the harm, while the morally and legally relevant culpability may lie elsewhere in the human or organizational chain (Cushman, 2008; Malle et al., 2014).

The problem becomes more serious in the case of autonomous and learning systems. Matthias introduced the concept of the "responsibility gap" to describe situations in which learning automata behave in ways that were not fully predictable by their manufacturers or operators. This gap is particularly significant for criminal law because punishment requires more than causal connection; it requires a justified attribution of culpability. If an autonomous vehicle kills a pedestrian, a medical AI recommends a fatal treatment, a predictive policing system reinforces discriminatory enforcement, or an autonomous weapon attacks civilians, the legal system must determine whether responsibility belongs to the developer, the data provider, the deploying institution, the user, the regulator, the corporation, or some combination of these actors (Matthias, 2004; Sparrow, 2007).

Autonomous weapons illustrate the moral urgency of this issue with particular clarity. Sparrow argues that when lethal decisions are delegated to autonomous systems, traditional assumptions about responsibility become unstable because it may be difficult to identify a human actor who can justly be held responsible for the resulting deaths. Although the same difficulty appears in less dramatic contexts such as autonomous driving, medical decision-support systems, and algorithmic risk assessment, military AI makes the problem more visible

Ijbmc.org
E-ISSN: 2345-5802

because wrongful harm may amount not only to private injury but also to violations of humanitarian and criminal law norms (Sparrow, 2007).

This article argues that the theory of moral blameworthiness can provide a useful interdisciplinary framework for analyzing crimes arising from AI systems. The concept of blameworthiness connects psychology, moral philosophy, and criminal law because it focuses on the conditions under which an actor deserves criticism, sanction, or punishment. In the context of AI-related harms, this framework allows responsibility to be evaluated through several core questions: Who had knowledge of the relevant risk? Who had the capacity to prevent the harm? Who exercised control over the design, training, deployment, or use of the system? Was the harm reasonably foreseeable? Was there negligence, recklessness, or intentional disregard of safety obligations? And how should culpability be allocated when harm results from a distributed chain of human and machine interaction? (Fischer & Ravizza, 1998; Malle et al., 2014; Matthias, 2004).

The legal importance of these questions has increased as regulatory systems have begun to respond to AI risks. The European Union's Artificial Intelligence Act adopts a risk-based approach and imposes obligations on providers and deployers of AI systems, particularly in high-risk contexts. The revised EU Product Liability Directive also modernizes liability rules for defective products, including software and AI-related systems. At the international level, the OECD AI Principles emphasize trustworthy, human-centered, transparent, robust, and accountable AI, while the NIST AI Risk Management Framework treats AI governance as an ongoing process of identifying, measuring, managing, and governing risks. These developments show that legal systems are moving toward accountability models that no longer treat AI merely as a passive tool, although they still do not fully resolve the deeper question of criminal blameworthiness (AI, 2023; Instruments, 2019).

Accordingly, the main problem addressed in this article is the inadequacy of classical criminal responsibility frameworks when applied to harms caused or facilitated by AI systems. Traditional doctrines such as mens rea, negligence, causation, and individual culpability remain necessary, but they require reinterpretation in light of algorithmic autonomy, opacity, distributed design processes, data dependency, and organizational decision-making. The article therefore examines how moral blameworthiness, as developed in moral psychology and philosophy, can be adapted to determine punishment in AI-related crimes. It does not claim that current AI systems are moral agents in the full human sense. Rather, it argues that blameworthiness should be assessed primarily across the human and institutional chain surrounding the AI system, while also taking account of the system's autonomy, foreseeability, explainability, and risk profile (Cushman, 2008; Malle et al., 2014; Matthias, 2004; Searle, 1980).

The objectives of this study are fourfold. First, it explains the theoretical foundations of moral blameworthiness in moral psychology and moral philosophy. Second, it analyzes the relationship between AI autonomy, agency, and responsibility. Third, it evaluates the limits of existing legal frameworks for assigning criminal liability in AI-related harms. Fourth, it proposes an integrated model for assessing blameworthiness and punishment across the chain of AI design, development, training, deployment, supervision, and use (AI, 2023; Fischer & Ravizza, 1998; Malle et al., 2014).

Methodologically, the article adopts a descriptive-analytical and interdisciplinary approach. It draws on moral psychology, philosophy of responsibility, criminal law theory, AI ethics, and emerging AI regulation. This approach is necessary because crimes involving AI systems cannot be understood through a purely technical or purely legal lens. They require a combined analysis of human cognition, institutional conduct, technological opacity, moral attribution, and legal culpability. By applying the theory of moral blameworthiness to AI-related criminal harms, the article seeks to contribute to a more coherent framework for responsibility and punishment in the age of intelligent systems (AI, 2023; Fischer & Ravizza, 1998; Malle et al., 2014; Matthias, 2004).

Methods and Materials

Study Design

This study adopts a descriptive-analytical and interdisciplinary research methodology to examine the application of the theory of moral blameworthiness in psychology to the determination of punishment for

Ijbmc.org
E-ISSN: 2345-5802

crimes arising from artificial intelligence systems. The descriptive dimension of the study is used to explain the main concepts of moral blameworthiness, criminal responsibility, artificial intelligence autonomy, responsibility gaps, and algorithmic risk. The analytical dimension is used to evaluate how these concepts interact when harmful or criminal outcomes are produced by AI systems. Because the subject lies at the intersection of moral psychology, philosophy of responsibility, criminal law, AI ethics, and technology regulation, a single-discipline method would be insufficient. Therefore, the study combines conceptual analysis, doctrinal legal analysis, and normative evaluation (Fischer & Ravizza, 1998; Malle et al., 2014; Matthias, 2004).

The research is primarily library-based and document-based. Its data are collected from academic books, peer-reviewed journal articles, legal instruments, policy documents, institutional reports, and official regulatory materials. The theoretical sources include works on moral responsibility, blame, free will, intentionality, and reasons-responsiveness. The psychological sources include studies on blame attribution, moral judgment, causation, intentionality, and punishment. The legal and regulatory sources include international and regional materials concerning AI governance, especially the European Union Artificial Intelligence Act, the OECD Recommendation on Artificial Intelligence, and the NIST Artificial Intelligence Risk Management Framework. These sources are selected because they directly address the conceptual, psychological, legal, and regulatory dimensions of responsibility in AI-related harms (AI, 2023; Instruments, 2019).

The study uses conceptual analysis to clarify the core meaning of moral blameworthiness and to distinguish it from related but different concepts such as causation, liability, accountability, fault, negligence, and punishment. This distinction is essential because an AI system may causally contribute to harm without being morally blameworthy in the same sense as a human agent. In moral psychology, blame is usually understood as a structured judgment that depends on the evaluation of causation, intentionality, obligation, knowledge, and control. Therefore, the conceptual analysis in this article examines whether these elements can be applied directly to AI systems or whether they should be redirected toward the human and organizational actors surrounding the AI system (Cushman, 2008; Malle et al., 2014).

The article also applies doctrinal legal analysis to examine the limits of traditional criminal responsibility when applied to AI-related crimes. Doctrinal analysis is used to identify the basic elements of criminal liability, especially the relationship between the physical element of crime and the mental element of crime. In the context of AI, this analysis is necessary because legal doctrines such as mens rea, negligence, recklessness, causation, and culpability were developed primarily for human actors. The study therefore analyzes whether these doctrines can adequately address harms produced by autonomous, semi-autonomous, or learning AI systems, and whether they require reinterpretation in light of algorithmic opacity, distributed decision-making, and the difficulty of predicting machine behavior (Matthias, 2004; Sparrow, 2007).

A normative-analytical method is used to evaluate how responsibility and punishment should be allocated when AI systems cause or contribute to criminal harm. This part of the methodology does not merely describe existing rules; it assesses whether those rules are morally and legally adequate. The normative analysis is guided by the theory of moral blameworthiness, especially the criteria of knowledge, control, foreseeability, intentionality, capacity, and responsiveness to reasons. These criteria are then applied to different actors in the AI responsibility chain, including designers, programmers, data providers, model trainers, deploying institutions, corporate managers, end users, and regulators (Fischer & Ravizza, 1998; Malle et al., 2014; Matthias, 2004).

The study follows a chain-of-responsibility approach. Under this approach, AI-related harm is not treated as the isolated act of a machine or a single individual. Instead, it is examined as the outcome of a complex socio-technical process involving design choices, data selection, training methods, testing procedures, deployment decisions, user behavior, organizational incentives, and regulatory oversight. This approach is consistent with contemporary AI governance frameworks, which emphasize that AI risks must be managed across the life cycle of AI systems rather than only at the point of final use. The NIST AI Risk Management Framework, for example, is intended for

Ijbmc.org
E-ISSN: 2345-5802

organizations that design, develop, deploy, or use AI systems and frames AI risk management as a continuous governance process (AI, 2023).

The legal-regulatory part of the study uses the European Union Artificial Intelligence Act as one of the main contemporary reference points because it represents a comprehensive, risk-based framework for regulating AI systems. The AI Act classifies AI systems according to levels of risk and imposes specific obligations on relevant actors, especially providers and deployers of high-risk AI systems. Although this regulation is not a criminal law instrument in the strict sense, it is methodologically important because it identifies duties of risk assessment, transparency, human oversight, data governance, documentation, and accountability. These duties can help determine whether a human or institutional actor acted with negligence, recklessness, or disregard of foreseeable risks in AI-related harms.

The study also considers civil liability developments, particularly the revised EU Product Liability Directive, because civil liability frameworks help clarify how legal systems are adapting traditional responsibility doctrines to software, digital products, and AI-related harms. Although civil liability and criminal liability are not identical, both fields require analysis of causation, defect, foreseeability, fault, and the allocation of responsibility among producers, operators, and users. The comparison is useful because criminal law can draw lessons from civil liability while still preserving the stricter requirements of culpability and punishment.

The sources are selected according to four criteria. First, the source must be directly relevant to moral blameworthiness, psychological attribution of blame, AI responsibility, AI regulation, or criminal/civil liability. Second, academic sources are prioritized when they are peer-reviewed, widely cited, or foundational in the field. Third, legal and policy sources are prioritized when they are official instruments or institutional frameworks issued by recognized bodies such as the European Union, OECD, or NIST. Fourth, recent sources are preferred for AI regulation because this field changes rapidly, while classical sources are retained where they provide foundational concepts of responsibility, control, blame, and agency (AI, 2023; Fischer & Ravizza, 1998; Instruments, 2019; Malle et al., 2014).

The analysis proceeds in five stages. In the first stage, the study defines moral blameworthiness and identifies its main psychological and philosophical components. In the second stage, it examines the nature of AI agency and distinguishes causal agency, functional agency, intentional agency, and moral agency. In the third stage, it analyzes the responsibility gap created by autonomous and learning systems. In the fourth stage, it evaluates existing legal and regulatory frameworks in light of the requirements of blameworthiness. In the fifth stage, it proposes an integrated framework for allocating responsibility and punishment among the human and institutional actors involved in the AI life cycle (Cushman, 2008; Malle et al., 2014; Searle, 1980; Sparrow, 2007).

The study does not rely on empirical fieldwork, interviews, surveys, or experimental data. This limitation is justified by the primarily conceptual and normative nature of the research question. The purpose of the article is not to measure public attitudes toward AI blame or to test judicial behavior experimentally, but to construct a theoretical and legal framework for evaluating blameworthiness in AI-related crimes. Nevertheless, the study uses findings from moral psychology to support its conceptual analysis of how blame is attributed and how punishment judgments are shaped by perceived causation, intention, knowledge, and control (Cushman, 2008; Malle et al., 2014).

The scope of the study is limited to current and near-future AI systems rather than hypothetical artificial general intelligence. This limitation is important because present AI systems, even when highly autonomous, do not possess consciousness, moral understanding, or free will in the ordinary human sense. For this reason, the article does not treat AI systems as fully blameworthy moral persons. Instead, it focuses on how moral and legal responsibility can be attributed to human and institutional actors when AI systems function as complex, partially autonomous instruments within socio-technical environments (Matthias, 2004; Searle, 1980).

Finally, the methodology is guided by the principle that punishment in AI-related crimes must be connected to morally relevant forms of culpability rather than to mere causal involvement. This means that liability should be assessed according to the degree of knowledge, control, foreseeability, preventability, and risk creation attributable to each actor in the AI chain. By

Ijbmc.org
E-ISSN: 2345-5802

combining moral psychology, criminal law theory, and AI governance, the study aims to develop a framework that can distinguish between accidental harm, negligent design, reckless deployment, organizational culpability, and intentional misuse of AI systems (AI, 2023; Fischer & Ravizza, 1998; Malle et al., 2014).

Findings and Results

The findings of this study show that the application of the theory of moral blameworthiness to crimes arising from artificial intelligence systems requires a significant reconstruction of traditional assumptions about agency, culpability, and punishment. AI-related harm cannot be adequately explained through a simple model in which one human actor intentionally commits one harmful act. Instead, such harm usually emerges from a complex interaction between design choices, data practices, model training, deployment contexts, user reliance, organizational incentives, and regulatory oversight. For this reason, blameworthiness in AI-related crimes should be evaluated as a distributed and layered phenomenon rather than as a purely individual and isolated form of culpability (AI, 2023; Matthias, 2004).

The first major finding is that current AI systems should not be treated as morally blameworthy agents in the full human sense. Although AI systems can produce outputs, make classifications, recommend actions, and even operate with a high degree of autonomy, they do not possess consciousness, moral understanding, practical reason, or free will in the ordinary sense required for moral blame. This means that AI systems may be causally responsible for harm but not morally blameworthy in the same way that human beings are. The distinction between causal contribution and moral blame is therefore central to any legal framework dealing with AI-related crimes (Malle et al., 2014; Searle, 1980).

The second finding is that moral blameworthiness should primarily be attributed to the human and institutional actors surrounding the AI system. These actors include system designers, software developers, data providers, model trainers, testing teams, deploying institutions, corporate managers, users, and regulators. Each of these actors may contribute to the final harmful outcome in a different way. For example, designers may create unsafe architectures, data providers may supply biased datasets, deploying institutions may use the system in unsuitable contexts, and users may rely on automated outputs without adequate supervision. Therefore, responsibility should not be reduced to the final user or the corporation alone; it should be assessed across the entire AI life cycle (AI, 2023; Instruments, 2019; Matthias, 2004).

The third finding is that the psychology of blame offers a useful framework for identifying the morally relevant elements of responsibility in AI-related crimes. Psychological theories of blame show that blame judgments are shaped by evaluations of causation, intentionality, knowledge, obligation, control, and preventability. These elements can be adapted to AI cases by asking whether a human or institutional actor knew or should have known about a risk, had the capacity to prevent the harm, violated a duty of care, ignored available warnings, or failed to exercise appropriate oversight. This approach prevents punishment from being based merely on the occurrence of harm and instead connects punishment to morally relevant culpability (Cushman, 2008; Malle et al., 2014).

The fourth finding is that the traditional concept of mens rea requires reinterpretation in AI-related crimes. In ordinary criminal law, mens rea is usually connected to a human mental state such as intention, knowledge, recklessness, or negligence. In AI-related cases, however, the harmful result may be produced through automated decision-making rather than direct human action. This does not eliminate the need for mens rea, but it shifts the inquiry toward the mental states of the human actors who designed, deployed, authorized, or supervised the system. In this context, concepts such as design recklessness, deployment negligence, algorithmic awareness, and organizational disregard of risk become especially important (Fischer & Ravizza, 1998; Matthias, 2004).

The fifth finding is that the responsibility gap is one of the most serious obstacles to assigning criminal responsibility for AI-related harms. The responsibility gap arises when an autonomous or learning system produces harmful conduct that was not fully intended or predicted by any individual actor, while the system itself cannot be meaningfully punished as a moral agent. This creates a situation in which harm occurs, but traditional legal categories struggle to identify a clearly blameworthy subject. The gap is especially significant in cases involving autonomous vehicles, medical AI

Ijbmc.org
E-ISSN: 2345-5802

systems, algorithmic risk assessment tools, and autonomous weapons (Matthias, 2004; Sparrow, 2007).

The sixth finding is that the responsibility gap should not be used as an excuse for non-liability. Even if no single actor fully intended the harmful outcome, the chain of human and institutional decisions that made the harm possible can still be evaluated. The relevant question is not only who directly caused the harm, but who created, increased, ignored, or failed to manage the relevant risk. This means that responsibility may be distributed among multiple actors according to their knowledge, control, foreseeability, and capacity to prevent harm. Such a model is more consistent with both moral psychology and modern AI governance than a model that searches only for one direct perpetrator (AI, 2023; Malle et al., 2014; Matthias, 2004).

The seventh finding is that foreseeability is a central criterion for determining blameworthiness in AI-related crimes. If a harmful outcome was reasonably foreseeable in light of available technical knowledge, prior warnings, known limitations, testing results, user complaints, or risk assessments, then the actors who ignored those indicators may be blameworthy. By contrast, if the harm was genuinely unforeseeable at the time of design or deployment, the degree of blameworthiness may be reduced. Therefore, criminal punishment in AI-related cases should be sensitive to the difference between unforeseeable accidents, negligent oversight, reckless deployment, and intentional misuse (Fischer & Ravizza, 1998; Malle et al., 2014).

The eighth finding is that control is another essential condition of blameworthiness. An actor should not be punished merely because they were connected to the AI system; they should be punished only to the extent that they had meaningful control over the relevant risk. Control may exist at different stages of the AI life cycle. Designers control architecture and safety constraints; data teams control dataset selection and labeling; model trainers control training procedures and evaluation methods; deployers control the operational context; users control how outputs are interpreted and acted upon; and regulators control approval, monitoring, and enforcement structures. The degree of control held by each actor should therefore affect both the attribution of responsibility and the severity of punishment (AI, 2023; Fischer & Ravizza, 1998).

The ninth finding is that knowledge and epistemic access must be evaluated carefully in AI-related crimes. Because many AI systems operate through complex and opaque processes, not every actor has the same ability to understand or predict system behavior. However, opacity should not automatically excuse responsibility. If opacity results from poor documentation, inadequate testing, intentional secrecy, lack of explainability, or commercial decisions that prioritize speed over safety, then opacity itself may become evidence of culpability. In such cases, the relevant actors may be blameworthy not because they fully understood the harmful output, but because they created or tolerated conditions under which harmful outputs could not be adequately understood or controlled (AI, 2023; Matthias, 2004).

The tenth finding is that organizational culpability plays a central role in AI-related crimes. Many harmful AI outcomes are not the result of one individual's decision but of organizational practices, including inadequate safety culture, pressure to deploy quickly, insufficient testing, poor data governance, weak internal auditing, misleading marketing, or failure to respond to known risks. Therefore, legal systems should give greater attention to corporate and institutional responsibility. A company that knowingly deploys a high-risk AI system without adequate oversight may be more blameworthy than a low-level employee who merely follows organizational procedures (AI, 2023; Instruments, 2019).

The eleventh finding is that existing legal frameworks are moving toward risk-based accountability, but they remain incomplete for criminal blameworthiness. The European Union Artificial Intelligence Act classifies AI systems according to risk and imposes obligations on providers and deployers, especially for high-risk systems. This approach is useful because it links legal duties to the level of danger created by the system. However, a regulatory risk-based model does not by itself determine criminal punishment. Criminal law still requires a separate inquiry into fault, culpability, causation, and proportionality.

The twelfth finding is that civil liability developments can support, but not replace, criminal responsibility analysis. The revised EU Product Liability Directive expands the legal understanding of defective products in a digital and software-based environment. This development is important because many AI harms may be addressed through product liability, compensation,

and risk allocation. However, criminal responsibility has a different function: it expresses moral condemnation and imposes punishment where culpability is sufficiently serious. Therefore, civil liability may compensate victims, but it cannot fully answer the question of moral blameworthiness in serious AI-related crimes (Fischer & Ravizza, 1998).

The thirteenth finding is that AI-related crimes require a proportional model of punishment. Punishment should be proportionate not only to the harm caused but also to the actor's degree of knowledge, control, foreseeability, and preventability. For example, a corporation that deploys a known unsafe AI system in a high-risk environment for economic gain should face more severe consequences than a user who reasonably relied on a system presented as safe. Likewise, intentional misuse of AI for fraud, discrimination, surveillance abuse, or violence should be treated differently from negligent failure to supervise an otherwise lawful system (Cushman, 2008; Malle et al., 2014).

The fourteenth finding is that strict liability may be useful for compensation, but it is insufficient as the main basis for criminal punishment. Strict liability can encourage safety and ensure that victims are compensated even when fault is difficult to prove. However, criminal punishment normally requires a stronger connection to culpability. Therefore, strict liability may be appropriate in civil or administrative contexts, but criminal liability should remain tied to morally relevant factors such as recklessness, negligence, knowledge, or intentional wrongdoing (Fischer & Ravizza, 1998).

The fifteenth finding is that the concept of automation bias is important for evaluating blameworthiness among users and supervisors of AI systems. Automation bias occurs when humans place excessive trust in automated systems and fail to question or verify their outputs. In AI-related crimes, this phenomenon may reduce or increase blame depending on the circumstances. It may reduce blame if the user reasonably relied on a system that was presented as reliable and safe. However, it may increase blame if the user ignored clear warnings, failed to exercise required professional judgment, or relied on automation in a context where independent verification was legally or ethically required (Cushman, 2008; Malle et al., 2014).

The sixteenth finding is that high-risk AI systems require heightened duties of care. Systems used in criminal justice, healthcare, transportation, employment, biometric identification, border control, education, and military contexts can seriously affect life, liberty, equality, and fundamental rights. Therefore, actors involved in these systems should be held to a higher standard of knowledge, testing, documentation, transparency, and oversight. The more dangerous the context of deployment, the stronger the duty to anticipate and prevent harm (Instruments, 2019).

The seventeenth finding is that transparency and documentation are not merely technical requirements; they are conditions for legal and moral accountability. Without adequate documentation, audit trails, explainability, and risk records, it becomes difficult to determine who knew what, when they knew it, and what they could have done to prevent harm. Therefore, failure to document design choices, data sources, testing limitations, and deployment risks should be treated as a serious factor in evaluating blameworthiness. In this sense, documentation functions as an evidentiary bridge between AI governance and criminal responsibility (AI, 2023).

The eighteenth finding is that AI-related criminal responsibility should be assessed through a chain-of-responsibility model. This model examines each stage of the AI life cycle: design, data collection, model training, testing, validation, deployment, monitoring, user interaction, and post-incident response. At each stage, the model asks whether the relevant actor had knowledge of risk, control over the system, capacity to prevent harm, access to relevant information, and a duty to act. This model avoids both over-punishment and under-punishment by ensuring that responsibility is allocated according to each actor's actual role in creating or failing to prevent the harm (AI, 2023; Matthias, 2004).

The nineteenth finding is that current AI systems should be treated as objects of regulation and instruments of human action rather than as independent criminal subjects. Although some theories propose electronic personhood or limited legal personality for advanced AI systems, current systems lack the moral capacities required for blame and punishment. Punishing an AI system would not achieve the traditional purposes of criminal punishment, such as moral condemnation, deterrence through rational choice,

Ijbmc.org
E-ISSN: 2345-5802

rehabilitation, or retributive desert. Therefore, the more coherent approach is to regulate AI systems and assign responsibility to the humans and institutions that control or benefit from them (Fischer & Ravizza, 1998; Searle, 1980).

The twentieth finding is that the proposed blameworthiness-based approach can distinguish between different categories of AI-related wrongdoing. Accidental AI harm may require compensation and technical correction but not criminal punishment. Negligent AI harm may justify liability where an actor failed to meet a reasonable duty of care. Reckless AI harm may justify stronger punishment where an actor consciously disregarded a substantial risk. Intentional AI misuse may justify the most severe punishment where AI is deliberately used as a tool for crime. This classification allows legal systems to respond to AI harms in a more proportionate and morally justified manner (Cushman, 2008; Malle et al., 2014).

Overall, the findings demonstrate that moral blameworthiness remains a necessary concept in the age of artificial intelligence, but it must be adapted to the realities of socio-technical systems. AI does not eliminate human responsibility; rather, it changes where and how responsibility must be located. The most defensible model is one that combines moral psychology, criminal law, and AI governance to evaluate knowledge, control, foreseeability, preventability, organizational conduct, and the distribution of risk across the AI life cycle. Such a model can help legal systems avoid both extremes: blaming machines as if they were human moral agents, or allowing human actors to escape responsibility by hiding behind the autonomy and complexity of machines (AI, 2023; Fischer & Ravizza, 1998; Malle et al., 2014; Matthias, 2004).

Discussion and Conclusion

The findings of this study demonstrate that the emergence of artificial intelligence systems requires a reconsideration of the traditional relationship between moral blameworthiness, criminal responsibility, and punishment. Classical theories of criminal liability were developed around human agents who can form intentions, understand rules, control their conduct, and respond to moral or legal reasons. AI systems, however, complicate this structure because they may generate harmful or criminal outcomes without possessing consciousness, moral understanding, or free will in the human sense. Therefore, the central problem is not merely whether AI systems can cause harm, but whether the conditions that justify blame and punishment can be meaningfully applied to the human and institutional actors involved in the AI life cycle (Fischer & Ravizza, 1998; Malle et al., 2014; Searle, 1980).

The discussion of moral blameworthiness shows that causation alone is not sufficient for criminal punishment. An AI system may be the immediate causal mechanism through which harm occurs, but moral and criminal responsibility require additional factors such as knowledge, control, foreseeability, intention, recklessness, negligence, or breach of duty. This distinction is crucial because without it, legal systems may either overextend punishment to actors who were only remotely connected to the harm or under-punish actors who created serious risks but avoided direct involvement in the final harmful act. The theory of blameworthiness therefore helps distinguish between mere causal participation and culpable contribution (Cushman, 2008; Malle et al., 2014).

A key implication of this study is that current AI systems should not be treated as fully blameworthy moral agents. Although advanced AI systems can classify, predict, recommend, generate, and act in ways that appear autonomous, they do not possess the psychological and moral capacities traditionally required for blame. They do not understand moral reasons as reasons, nor do they experience guilt, remorse, responsibility, or normative self-reflection. For this reason, attributing criminal blame directly to the AI system would be conceptually weak and practically ineffective. Punishment directed at an AI system would not satisfy the usual aims of criminal punishment, such as moral condemnation, retribution, deterrence, rehabilitation, or responsibility-based accountability (Fischer & Ravizza, 1998; Searle, 1980).

However, the fact that AI systems themselves are not fully blameworthy does not mean that no one is responsible when they cause harm. On the contrary, one of the central conclusions of this study is that AI-related harms often reveal a distributed structure of human and institutional responsibility. Designers, developers, data providers, model trainers, corporate decision-makers, deploying institutions, professional users, and regulators

Ijbmc.org
E-ISSN: 2345-5802

may each contribute to the harmful outcome in different ways. The question, therefore, should not be "Is the AI guilty?" but rather "Which human or institutional actors created, increased, ignored, or failed to control the relevant risk?" (AI, 2023; Matthias, 2004).

The responsibility gap remains one of the most important theoretical and legal challenges in this field. When autonomous or learning AI systems behave in ways that were not specifically intended or fully predicted by any individual actor, traditional doctrines of responsibility may struggle to identify a clearly culpable subject. Nevertheless, this gap should not be understood as a complete absence of responsibility. Rather, it should be treated as a signal that legal systems need more refined tools for assessing distributed culpability, organizational fault, and risk-based duties. In this sense, the responsibility gap is not only a problem of AI autonomy; it is also a problem of legal design and institutional accountability (Matthias, 2004; Sparrow, 2007).

From the perspective of moral psychology, blame attribution is structured around several recurring elements: causation, intentionality, knowledge, obligation, capacity, and control. These elements remain relevant in AI-related crimes, but they must be applied to the network of actors surrounding the AI system rather than to the machine alone. For example, a company that deploys a high-risk AI system despite known safety weaknesses may be blameworthy because it had knowledge of the risk and the capacity to prevent harm. A user who blindly follows an automated recommendation in a context requiring professional judgment may also be blameworthy if independent verification was reasonably expected. Thus, moral psychology provides criteria for evaluating degrees of culpability across the AI chain (Cushman, 2008; Malle et al., 2014).

The findings also show that the traditional concept of mens rea needs reinterpretation in the context of AI-related crimes. Since AI systems do not possess human mental states, the mental element of criminal liability should be examined through the knowledge, intention, recklessness, or negligence of the relevant human actors. In this context, concepts such as design recklessness, deployment negligence, algorithmic awareness, and organizational disregard of risk become important. These concepts help translate classical criminal law categories into the realities of AI development and deployment without abandoning the culpability requirement that gives criminal punishment its moral legitimacy (Fischer & Ravizza, 1998; Matthias, 2004).

The study further suggests that criminal responsibility in AI-related cases should be proportional and differentiated. Not all AI-related harms should produce criminal punishment. Some harms may be accidental and require compensation, correction, or regulatory intervention rather than criminal sanction. Other harms may result from negligence, such as inadequate testing or poor supervision. More serious cases may involve recklessness, such as deploying a system despite known safety failures. The most severe cases involve intentional misuse, where AI is deliberately used to commit fraud, discrimination, surveillance abuse, violence, or other crimes. A blameworthiness-based model allows legal systems to distinguish among these categories and impose proportionate responses (Cushman, 2008; Malle et al., 2014).

Another important point is that organizational culpability must be taken seriously. Many AI-related harms are not caused by one isolated wrongdoer but by corporate structures, institutional incentives, inadequate safety cultures, weak documentation, poor data governance, and pressure to deploy quickly. Criminal law often struggles with such distributed wrongdoing because it tends to search for identifiable individual fault. However, AI systems are developed and deployed within organizations; therefore, the organization itself may be the most appropriate object of blame when harmful outcomes result from systemic negligence or reckless business practices. A blameworthiness-based approach should therefore include both individual and corporate responsibility (AI, 2023; Instruments, 2019).

The regulatory developments examined in this study indicate that contemporary legal systems are beginning to move toward risk-based accountability. The European Union Artificial Intelligence Act classifies AI systems according to risk and imposes obligations on providers and deployers, especially in high-risk contexts. This model is important because it links duties of care to the level of potential harm created by the system. However, risk-based regulation is not equivalent to criminal responsibility. It can define duties, standards, and compliance obligations, but criminal punishment still

requires an additional analysis of fault, causation, proportionality, and blameworthiness.

Civil liability developments also provide useful lessons, especially in relation to defective digital products, software, and AI-related harm. The revised EU Product Liability Directive reflects an effort to modernize liability rules for technologically complex products. Such reforms are important for victim compensation and risk allocation, but they cannot replace the moral function of criminal law. Civil liability can compensate harm even where culpability is limited, while criminal liability expresses condemnation and requires a stronger justification. Therefore, civil and criminal responses to AI-related harm should be complementary rather than identical (Fischer & Ravizza, 1998).

The discussion also reveals that transparency, documentation, and explainability are not merely technical ideals; they are essential conditions for responsibility. If a system's design, data sources, training process, testing limitations, risk assessments, and deployment decisions are not documented, it becomes difficult to determine who had knowledge, control, or the capacity to prevent harm. In this sense, transparency is not only a governance requirement but also an evidentiary condition for criminal accountability. Failure to maintain adequate documentation may itself support a finding of negligence or recklessness where harm was foreseeable (AI, 2023).

The findings further indicate that high-risk AI systems require heightened duties of care. AI systems used in healthcare, transportation, criminal justice, biometric identification, employment, education, military operations, and border control can directly affect life, liberty, equality, and human dignity. Therefore, the standard of responsibility for actors involved in these systems should be more demanding than for low-risk systems. The more serious and foreseeable the potential harm, the stronger the duties of testing, monitoring, human oversight, transparency, and post-deployment review should be (Instruments, 2019).

5.2. Theoretical Implications

Theoretically, this study contributes to the literature by showing that moral blameworthiness remains relevant in the age of artificial intelligence, but it must be reconstructed for socio-technical systems. Traditional theories of responsibility focus primarily on individual agents, while AI-related harms often arise from distributed processes involving many actors and technological components. Therefore, a modern theory of blameworthiness must include both individual moral agency and institutional risk creation. This does not require treating AI systems as persons; rather, it requires expanding the analysis of human responsibility across the AI life cycle (Fischer & Ravizza, 1998; Matthias, 2004).

The study also clarifies the distinction between causal agency, functional agency, and moral agency. AI systems may possess causal agency because they can produce effects in the world. Some may possess a limited form of functional agency because they can select outputs or actions based on goals, models, or optimization functions. However, current AI systems do not possess full moral agency because they lack moral understanding, consciousness, and genuine reasons-responsiveness. This distinction prevents conceptual confusion and helps avoid both technological exaggeration and legal irresponsibility (Malle et al., 2014; Searle, 1980).

Another theoretical implication is that blameworthiness should be understood as a matter of degree. In AI-related harms, different actors may bear different degrees of responsibility depending on their role, knowledge, control, and capacity to prevent harm. This layered approach is more suitable than a binary model that treats actors as either fully responsible or not responsible at all. It also reflects the psychological reality that blame judgments are sensitive to variations in intention, foreseeability, causation, and preventability (Cushman, 2008; Malle et al., 2014).

5.3. Legal and Policy Implications

Legally, the study supports the development of AI-specific doctrines of criminal responsibility. Traditional concepts such as mens rea, negligence, recklessness, causation, and corporate liability remain essential, but they should be adapted to the realities of algorithmic decision-making. Courts and lawmakers should consider whether an actor had algorithmic awareness, whether the harm was foreseeable, whether proper testing and oversight were performed, whether deployment occurred in a high-risk context, and whether the organization ignored known warnings. These factors can help determine whether criminal punishment is justified (AI, 2023; Matthias, 2004).

Ijbmc.org
E-ISSN: 2345-5802

The study also suggests that legal systems should develop a chain-of-responsibility model for AI-related crimes. Under this model, responsibility is examined at each stage of the AI system's life cycle: design, data collection, training, testing, validation, deployment, use, monitoring, and post-incident response. At each stage, investigators should ask who had control, who had knowledge, who had a duty to act, who benefited from the risk, and who could have prevented the harm. This model can reduce the risk that corporations or individuals escape responsibility by attributing harm to the complexity or autonomy of the system (AI, 2023; Matthias, 2004).

From a policy perspective, AI governance should combine preventive regulation with responsibility-based enforcement. Preventive measures include mandatory risk assessment, independent auditing, documentation, transparency obligations, human oversight, and post-deployment monitoring. Enforcement measures include civil compensation, administrative sanctions, corporate liability, and criminal punishment in cases of serious culpability. This layered approach is preferable because not all AI harms require the same legal response, and criminal law should be reserved for cases involving significant fault or serious disregard of risk (Instruments, 2019).

The study also supports the creation or strengthening of specialized institutions for AI-related incidents. Because AI harms often involve technical complexity, legal ambiguity, and distributed responsibility, ordinary courts and regulatory agencies may lack sufficient expertise. Specialized AI investigation bodies, technical audit authorities, and expert judicial panels could help determine causation, foreseeability, compliance failures, and organizational culpability. Such institutions would improve both victim protection and the fairness of responsibility attribution (AI, 2023; Instruments, 2019).

5.4. Limitations of the Study

This study has several limitations. First, it is primarily conceptual and normative rather than empirical. It does not conduct interviews, surveys, experiments, or case-based statistical analysis. Therefore, its conclusions concern the theoretical and legal structure of blameworthiness rather than measured public attitudes, judicial behavior, or organizational practices. Future empirical work is needed to test how judges, juries, regulators, developers, and users actually attribute blame in AI-related harm scenarios (Cushman, 2008; Malle et al., 2014).

Second, the study focuses on current and near-future AI systems rather than artificial general intelligence. This limitation is necessary because present AI systems lack consciousness, moral understanding, and free will in the ordinary human sense. If future AI systems were to develop capacities closer to moral understanding or autonomous practical reasoning, the conclusions of this article would require reconsideration. However, under current technological conditions, the most defensible approach is to treat AI as a complex instrument embedded in human and institutional systems rather than as a fully responsible moral subject (Matthias, 2004; Searle, 1980).

Third, the study relies heavily on selected international and European regulatory frameworks, especially the EU AI Act, OECD AI Principles, and NIST AI Risk Management Framework. These sources are valuable because they are influential and contemporary, but they do not represent all legal systems. Different jurisdictions may adopt different approaches to criminal responsibility, corporate liability, product liability, and AI governance. Therefore, comparative legal research is needed to examine how the proposed blameworthiness-based model could be adapted to national legal systems, including civil law, common law, Islamic law, and mixed jurisdictions (AI, 2023; Instruments, 2019).

Fourth, the study is limited by the rapid pace of technological change. AI systems are evolving quickly in terms of autonomy, scale, opacity, and social integration. Legal and philosophical frameworks may therefore require continuous revision. A model that is adequate for current machine learning systems may not be sufficient for future systems with more advanced planning, autonomy, embodiment, or interaction with physical environments. This limitation reinforces the need for flexible, principle-based legal frameworks rather than rigid rules that may soon become outdated (AI, 2023; Instruments, 2019).

5.5. Suggestions for Future Research

Future research should examine how legal decision-makers actually attribute blame in AI-related cases. Experimental studies could present judges, jurors, lawyers, and ordinary citizens with scenarios involving autonomous vehicles, medical AI, predictive policing, algorithmic discrimination, and autonomous weapons.

Ijbmc.org
E-ISSN: 2345-5802

Such research would help determine whether people blame developers, users, corporations, regulators, or the AI system itself, and how factors such as foreseeability, control, transparency, and harm severity affect these judgments (Cushman, 2008; Malle et al., 2014).

Further research should also develop more precise legal categories for AI-related culpability. Concepts such as algorithmic negligence, design recklessness, deployment recklessness, organizational algorithmic fault, and failure of human oversight require deeper doctrinal analysis. These categories could help lawmakers and courts translate traditional criminal law concepts into AI contexts without abandoning the principle that punishment must be based on culpability (Fischer & Ravizza, 1998; Matthias, 2004).

Comparative legal research is also necessary. Different legal systems may vary in how they treat corporate criminal liability, negligence, strict liability, product liability, causation, and the mental element of crime. Future studies should compare approaches in the European Union, United States, United Kingdom, Iran, China, and other major legal systems. Such research could identify whether a global or regional framework for AI-related criminal responsibility is possible, or whether responsibility models must remain jurisdiction-specific (Instruments, 2019).

Future research should also investigate the role of technical explainability in legal accountability. If courts require proof of knowledge, control, causation, or foreseeability, they will need reliable methods for understanding how AI systems produced harmful outputs. Therefore, interdisciplinary work between lawyers, computer scientists, psychologists, and ethicists is needed to develop explainability tools that are legally meaningful rather than merely technically descriptive (AI, 2023).

Finally, future studies should examine the relationship between AI governance and criminal punishment. Regulatory duties such as documentation, testing, auditing, transparency, and human oversight may become important evidence in criminal proceedings. Research should therefore explore how violations of AI governance standards can support findings of negligence, recklessness, or organizational culpability in serious harm cases (AI, 2023).

5.6. Conclusion

This study concludes that the theory of moral blameworthiness provides a valuable framework for determining responsibility and punishment in crimes arising from artificial intelligence systems. AI systems challenge traditional criminal law because they can contribute to harmful outcomes without possessing the human capacities normally required for blame. However, this does not eliminate responsibility. Instead, it requires responsibility to be relocated and redistributed across the human and institutional actors who design, train, deploy, supervise, regulate, and benefit from AI systems (Malle et al., 2014; Matthias, 2004).

The central conclusion is that current AI systems should not be treated as fully blameworthy moral or criminal agents. They may be causally involved in harm, but they lack consciousness, moral understanding, free will, and genuine responsiveness to moral reasons. Therefore, direct punishment of AI systems is neither conceptually justified nor practically useful. The more coherent approach is to evaluate the blameworthiness of human and organizational actors according to their knowledge, control, foreseeability, preventability, and role in creating or managing risk (Fischer & Ravizza, 1998; Searle, 1980).

The article also concludes that AI-related responsibility is often distributed rather than individual. Harm may result from unsafe design, biased data, inadequate testing, poor documentation, reckless deployment, excessive user reliance, weak corporate governance, or regulatory failure. A blameworthiness-based framework can identify how much responsibility each actor bears and what legal response is proportionate. This approach avoids both unjust over-punishment and unjust impunity (AI, 2023; Matthias, 2004).

The proposed model emphasizes five core criteria: knowledge, control, foreseeability, capacity to prevent harm, and degree of risk creation. These criteria allow legal systems to distinguish between accidental harm, negligent harm, reckless harm, and intentional misuse of AI. They also support a more nuanced approach to punishment, where the severity of legal response corresponds to the degree of culpability rather than merely to the severity of the outcome (Cushman, 2008; Malle et al., 2014).

Ijbmc.org
E-ISSN: 2345-5802

In conclusion, artificial intelligence does not make moral blameworthiness obsolete. Rather, it forces legal and psychological theory to refine the concept of blameworthiness for a world in which harmful actions may emerge from complex interactions between humans, organizations, and autonomous systems. The most defensible path is not to blame machines as if they were human moral agents, nor to allow human actors to avoid responsibility by invoking machine autonomy. Instead, legal systems should develop a chain-of-responsibility model grounded in moral psychology, criminal law, and AI governance. Such a model can preserve the moral legitimacy of punishment while responding effectively to the distinctive risks of intelligent systems (AI, 2023; Fischer & Ravizza, 1998; Malle et al., 2014; Matthias, 2004).

Acknowledgments

The authors express their gratitude and appreciation to all participants.

Declaration of Interest

The authors of this article declared no conflict of interest.

Ethical Considerations

The study protocol adhered to the principles outlined in the Helsinki Declaration, which provides guidelines for ethical research involving human participants. Ethical considerations in this study were that participation was entirely optional.

Transparency of Data

In accordance with the principles of transparency and open research, we declare that all data and materials used in this study are available upon request.

Funding

This research was carried out independently with personal funding and without the financial support of any governmental or private institution or organization.

Authors' Contributions

All authors equally contribute to this study.

References

AI, N. (2023). Artificial intelligence risk management framework (AI RMF 1.0). *URL: https://nvlpubs. nist. gov/nistpubs/ai/nist. ai*, 100-101. https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*(2), 353-380. https://doi.org/10.1016/j.cognition.2008.03.006

Fischer, J. M., & Ravizza, M. (1998). *Responsibility and control: A theory of moral responsibility*. Cambridge university press. https://doi.org/10.1017/CBO9780511814594

Instruments, O. L. (2019). Recommendation of the council on artificial intelligence. *Organization for Economic Cooperation and Development*. https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological inquiry*, *25*(2), 147-186. https://doi.org/10.1080/1047840X.2014.877340

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and information technology*, *6*(3), 175-183. https://doi.org/10.1007/s10676-004-3422-1

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and brain sciences*, *3*(3), 417-424. https://doi.org/10.1017/S0140525X00005756

Sparrow, R. (2007). Killer robots. *Journal of applied philosophy*, *24*(1), 62-77. https://doi.org/10.1111/j.1468-5930.2007.00346.x