

The Shift of Criminal Liability in the Digital Space: An Analytical Study of the 'Autonomous Agent' Dilemma in Light of Iraq's Cybercrime Legislation

Nadhim Jawad Rashid Al-Mamore



DOI : <https://doi.org/10.61796/ijblps.v3i9.558>



Sections Info

Article history:

Submitted: July 07, 2026

Final Revised: August 16, 2026

Accepted: September 04, 2026

Published: October 06, 2026

Keywords:

Shift of criminal liability

Autonomous agent

Black box

Cybercrime in Iraq

Self-learning

Presumed liability

ABSTRACT

Objective: This analytical study aims to deconstruct the 'autonomous agent' dilemma and examine the effect of the 'Black Box' problem on producing the phenomenon of the 'shift of criminal liability' away from the human element, while projecting this dilemma onto the Iraqi legislative framework represented by the Penal Code No. 111 of 1969 and the draft Anti Cybercrime Law. **Method:** To achieve this objective, the study employs both the foundational analytical methodology and the comparative methodology, through an inductive survey of contemporary Anglo-American and Latin criminal theories. **Results:** The study arrives at pivotal findings, most notably: the existence of a structural impunity gap in Iraqi law resulting from the collapse of the elements of both the physical and mental elements (knowledge and will), and the severance of the causal nexus in human terms when the machine deviates autonomously in addition to the incapacity of traditional criminal participation doctrines to accommodate algorithmic conduct. **Novelty:** Accordingly, the study concludes by presenting a proposed legislative strategy for the Iraqi legislator, requiring the abandonment of the traditional punitive policy based on absent human intent, and the adoption of a system of 'presumed liability based on the creation of digital risks,' along with the activation of criminal liability for legal persons (developing companies) and the introduction of qualitative cyber precautionary measures, such as algorithmic quarantine and the programmatic execution of deviant code.

INTRODUCTION

Traditional criminal law, since the formulation of its earliest theories, has been governed by a central, entrenched principle stipulating that crime is the product of a 'sinful human will.' No conduct is criminally proscribed unless driven by a conscious, organic movement; nor does penal liability arise unless that movement is bound to a criminal intent constituted by knowledge and will. However, the accelerating technological leaps witnessed over the last two decades most notably the digital space's entry into the era of the third generation and algorithmic systems relying on 'deep machine learning' (Deep Learning) and artificial neural networks have shaken this philosophical and legal certainty. Technology has evolved from being a mere static, obedient 'tool' (Tool) that executes an extension of human will, to an 'autonomous agent' (Autonomous Agent) possessing the capacity to process data, interact dynamically with the virtual environment, and make automatic decisions and innovative behaviours never previously encoded in its original source code.

This structural shift from 'mechanism' to 'autonomy' has given rise to a highly complex legal phenomenon known as the 'Shift of Criminal Liability' (Shift of Liability). This phenomenon manifests when a selflearning digital system commits conduct

resulting in serious harm that constitutes a crime under penal laws such as unlawful algorithmic trading, autonomous cyberintrusion, or the generation of destructively criminal content. In such a scenario, legal scholars and the judiciary are confronted with a fully realized material element pertaining to the crime in the virtual space, yet one entirely severed both materially and psychologically from any human element. The programmer who authored the original code has lost control of the system following its deployment and autonomous interaction, while the end user remains ignorant of the complex logical mechanism through which the algorithm reached its decision, owing to the 'Black Box Problem' (The Black Box Problem). This technical opacity severs the causal thread and dismantles the criminal intent (knowledge and will) in its traditional conception, thereby creating an 'Impunity Gap' (Impunity Gap) that threatens the stability of criminal justice.

At the national level, this dilemma manifests acutely when mapped onto the Iraqi legislative environment. Iraqi criminal policy remains governed by the general provisions of the Penal Code No. 111 of 1969, as amended provisions formulated in a purely physical environment centered on the human actor and traditional criminal participation (the principal and the accomplice). Even with the emergence and parliamentary deliberation of draft 'AntiCybercrime Law' texts in Iraq, their formulations have remained immersed in a 'static' view of technology, always presuming the existence of a human who 'enters, obstructs, or hacks' using a computer as a tool, and entirely overlooking the contemporary scenario in which an intelligent system is the true actor and decisionmaker behind criminal conduct.

The research problem crystallizes in the inability of the Iraqi legal system whether the general provisions of the Penal Code or the specific texts of the draft Anti Cybercrime Law to confront the phenomenon of the 'shift of criminal liability' in the digital space. This dilemma is embodied in the legislative and conceptual gap arising from the collision of the 'autonomous agent' phenomenon and the 'Black Box' problem with the traditional requirements of penal attribution, rendering the proof of the mental element (intent or fault) and the causal nexus impossible with respect to the programmer or the human user, and inevitably leading to the judiciary's inability to prosecute and to the impunity of spontaneous algorithmic crimes.

This analytical study seeks to answer the following central questions:

1. What is the conceptual grounding of the 'shift of criminal liability,' and how do the 'autonomous agent' and the 'Black Box' problem dismantle the traditional elements of crime?
2. What are the shortcomings and structural rigidity in the Iraqi Penal Code and the draft Anti Cybercrime Law when confronting the automatic criminal acts of software?
3. What international comparative models such as vicarious liability or electronic personality have attempted to address this dilemma, and what are their limits of effectiveness?

4. What is the optimal legislative strategy and attributive model that the Iraqi legislator can adopt to mitigate the risks of this shift and preserve cyber security?

This study acquires amplified significance on both the theoretical and applied levels:

1. Theoretically: it contributes to the development of the conceptual framework of the theory of digital crime by introducing the variable of 'mechanical autonomy' as a factor influencing the elements of crime.
2. Applied: it provides the Iraqi judiciary and legislator with clear interpretive frameworks that assist in filling legislative gaps and overcoming ambiguity in information law texts.
3. Comparatively: it situates the Iraqi legislative experience within a global context, facilitating the absorption of lessons drawn from leading international experiences.

RESEARCH METHOD

To encompass the philosophical and applied dimensions of the research, the study adopted a methodological complementarity combining:

1. The Foundational Analytical Methodology: for dissecting and deconstructing the general rules of criminal liability in Iraqi law, and analysing the texts of the draft AntiCybercrime Law to demonstrate their adequacy and identify their shortcomings.
2. The Prospective Comparative Methodology: through an inductive survey of modern criminal theories and legislative and judicial trends in contemporary Western schools (Anglo-American and Latin) and international reports, with the aim of deriving a developed attributive model compatible with the constitutional and legal particularities of the Iraqi penal system.

RESULTS AND DISCUSSION

Results

Chapter One: The Conceptual Framework of the Shift of Criminal Liability and the 'Autonomous Agent' Dilemma

Section One: The Concept of the 'Shift of Liability' (Shift of Liability) and Its Criminal Dimensions

The 'shift of criminal liability' in the digital space refers to the legal condition arising from the dispersal of the hypothetical link between the criminal act and the human actor, such that the authority of penal attribution moves in tandem with the shift in the source of decision-making from the human (programmer or user) to autonomous algorithmic systems [1]. This shift does not merely signify the absence of a criminal; it

signifies the repositioning of the elements of crime outside the traditional scope agreed upon in penal legislation, foremost among which is Iraqi legislation.

Subsection One: The Effect of the Machine's Autonomous Will on the Physical Element of the Crime (Conduct and Causation)

The physical element of a crime is organized in a logical sequence beginning with criminal conduct, passing through the harmful result, and bound by the causal nexus. In the traditional digital environment, this sequence was linear and direct: the programmer writes a malicious code (conduct), a governmental Iraqi website's service is disrupted (result), and causation is clear because human conduct is the sole driver of the machine [2].

However, the digital space's entry into the era of 'machine agency' (Machine Agency) has dealt a violent blow to this tripartite organization of the physical element:

First the transformation of criminal conduct from the human to the mechanical: criminal doctrine defines conduct as 'a voluntary, organic movement emanating from a human being'; accordingly, inanimate objects and animals do not produce criminal conduct in the legal sense. However, when a self-learning algorithm based on an autonomous processing of data invents a complex cyber-attack method never previously encoded in its foundational code, we are confronted with a genuine 'material conduct' that produces the harmful effect, yet is entirely devoid of any human organic movement.[3] This autonomous algorithmic conduct faces the obstacle of 'unlawfulness' in Iraqi legislation: how does a judge qualify criminal conduct whose initial movement did not emanate from a human being? [4].

Second the fracturing of the causal nexus through 'autonomous algorithmic intervention': criminal theories (such as the theory of equivalence of causes or adequate causation) rely on tracing the causal thread to determine whether the accused's fault directly led to the result. In the case of the autonomous agent, the machine's will intervenes as a 'foreign and spontaneous factor' severing this thread [5].

When a programmer deploys an intelligent system for cybersecurity management, and the system based on its self-learning attacks other servers and destroys their data under the pretext of 'automatic preemptive defence,' the direct conduct that produced the result is the machine's autonomous decision, not the programmer's original code. This machine intervention constitutes the technically force majeure severing causation; legally, the programmer cannot be regarded as the material perpetrator of conduct that the machine independently modified and developed through its own separate digital will [6].

This fracturing of the physical element demonstrates that the 'shift of liability' is not merely an intellectual luxury, but a reality confronting the courts where the physical element of the crime stands complete and realized in the digital world, yet is materially and causally severed from any human being, rendering the traditional provisions of the Iraqi Penal Code incapable of grasping the 'materiality' of the crime [7].

Subsection Two: The Effect of Self Learning on the Mental Element (Deconstruction of Criminal Intent and Unintentional Fault)

Whereas the physical element pertains to the external body of the crime, the mental element represents its soul and the fundamental condition for the legitimacy of punishment; for there is no crime without culpability, and no criminal liability without the existence of a sinful will emanating from human consciousness and awareness.[8] This psychological nexus takes two forms in criminal policy: criminal intent (in intentional offences), or criminal fault (in unintentional offences). Nonetheless, the characteristic of 'self-learning' (Self Learning) of autonomous digital systems has produced a radical deconstruction of both forms.

First the deconstruction of traditional criminal intent (Mens Rea): criminal intent in the general rules of the Iraqi Penal Code rests on the twin elements of 'knowledge' and 'will' that is, the offender's knowledge of the constituent elements of the crime and the directedness of his will towards realising the criminal conduct and result [2].

However, this concept dissolves when confronting self-learning algorithms. When a digital system evolves autonomously and begins committing harmful conduct such as disclosing trade secrets or hacking data based on a self-directed analysis to maximize its profits the element of 'knowledge' is entirely absent from the programmer or human user, because he is ignorant of the new behavioral genes internally generated by the machine [8], [9]. Equally, the element of 'will' is severed, because the human's will was directed towards creating a beneficial intelligent system, while the machine's 'digital will' (Digital Will) if the expression may be permitted directed itself independently towards realising the criminal result, rendering the imputation of intentional mens rea to the human legally impermissible [10].

Second the crisis of characterizing unintentional fault (Negligence): confronted with the impossibility of proving intentional mens rea, prosecutorial authorities may attempt to impute the crime on the basis of 'unintentional fault' (such as recklessness, negligence, inattentiveness, or failure to observe regulations). This attempt rests on the notion that the programmer or user was negligent in supervising the system or deficient in anticipating its risks [11].

Yet here another dilemma emerges: the standard of criminal fault is directly linked to 'foreseeability' (Foreseeability), and since complex systems based on deep neural networks are characterized by structural opacity and make instantaneous decisions based on data changing in fractions of a second, the criminal conduct becomes technically 'absolutely unforeseeable'[12].

If the human's incapacity to foresee the machine's autonomous conduct is a scientifically and technically established incapacity, it cannot be characterized as negligence or recklessness, because the law does not mandate the impossible. Consequently, the attempt to confine the crisis within the scope of 'unintentional fault' fails to accommodate the dynamic nature of self-learning, leaving the criminal judge confronting a genuine dilemma: a mental element that has collapsed in human terms, yet

is realized digitally in the machine's virtual consciousness, which penal law does not yet recognize [13].

Section Two: The Nature of the 'Autonomous Agent' in the Digital Environment

The digital space has witnessed a radical transformation in the nature of the tools and software employed; technology is no longer a mere static channel for conveying human will, but has evolved into cyber entities possessing the capacity for decision-making and self-learning. This transformation has destabilized the traditional concepts of criminal law, which were historically founded on the centrality of human will as a fundamental condition for penal liability. To deconstruct this dilemma, it is first necessary to draw a precise distinction between the traditional conception of technology as a tool and the contemporary conception of technology as an autonomous agent.

Subsection One: The Distinction between Technology as a 'Crime Instrument' and Technology as a 'Criminal Agent'

The edifice of traditional criminal law rests on the deterministic premise that crime is a sinful human conduct emanating from awareness and choice. In this context, information systems and software were legally characterized as 'instruments of crime' (*Instrumentum Sceleris*), philosophically no different from a bladed weapon or a firearm in traditional offences; the machine here is merely a material intermediary through which the human actor's criminal intent extends to execute his criminal purpose [6].

In this traditional model adopted by Iraqi criminal policy through the Penal Code's general rules and the draft Anti Cybercrime Law criminal attribution remains clear: malicious software (such as traditional viruses or spyware) operates solely pursuant to premade programmatic commands authored by the programmer or triggered by the user, and thus the human remains the 'principal actor' while technology remains a mere obedient tool [7].

However, with the emergence of advanced algorithmic systems based on 'deep machine learning' (Deep Learning), the legal gap opened. Technology is no longer merely a tool executing preprogrammed scenarios (If Then Rules); it has come to possess what is termed 'technological autonomy' (Technological Autonomy) the digital system's capacity to interact with the surrounding environment, analyse data, and make 'automatic' decisions and behaviours unexpected or programmed in its original code by the human programmer [14].

When these autonomous systems engage in conduct constituting a crime under penal laws such as an intelligent bot engaging in algorithmic trading based on autonomous analysis and committing money laundering, or an autonomous cybersystem breaching a network and selecting its targets based on its own assessment of self-protection we transition from the concept of 'technology as tool' to the concept of 'technology as criminal agent' [15].

The paradox and dilemma here lies in the fact that the harmful conduct has been realized, yet the psychological and mental nexus (representing the mental element of the crime) has been severed between the programmer/user and the criminal result. The

human did not will the result and did not anticipate it, while the machine produced the result based on its 'independent algorithmic reasoning' rendering any attempt to apply the traditional concept of 'the principal actor' to this situation a form of legal coercion that collides with the principle of legality of crimes and penalties [2].

Subsection Two: The 'Black Box' Problem (Black Box) and the Absence of Predictability of Algorithmic Conduct

If the transformation from 'technology as tool' to 'technology as agent' represents the external manifestation of the crisis, then the philosophical and technical root nourishing this dilemma lies in a highly complex technical phenomenon known as **the 'Black Box Problem' (The Black Box Problem)**. This phenomenon expresses the condition in which the computational and logical processes underlying the decisions of autonomous intelligent systems become unreadable or comprehensible not only to ordinary users, but even to the engineers and programmers who authored the foundational source code of the system itself [16].

This dilemma arises as a result of modern digital space's reliance on machine learning algorithms and deep artificial neural networks (Deep Neural Networks), wherein the system is trained on millions of data inputs, then left to independently and secretly restructure its internal connections and extract results and patterns through hidden layers (Hidden Layers) [17]. On this basis, a 'selective' behaviour of the machine is generated, severing its connection to the initial instructions; the machine produces outputs which may constitute cybercrime based on its own internally developed logic, without leaving behind any explanatory trail that would allow a forensic investigator or judge to ascertain 'why' the algorithm made this particular decision [4].

This technical dilemma is reflected sharply on the criminal system through two pivotal dimensions:

First the collapse of the standard of legal predictability (Predictability): the edifice of unintentional criminal liability (criminal fault) rests on the standard of the extent to which an ordinary person could 'foresee the harmful result' and avoid it. In the complex digital space, and due to the opacity of the black box, it becomes impossible for the programmer or user to foresee the criminal conduct that the autonomous system may engage in [18]. If the conduct is inherently opaque and unforeseeable in its technical nature, the imputation of 'criminal fault' to the human programmer is legally extinguished, because the law does not require individuals to foresee what is scientifically and technically impossible leading to the collapse of the mental element of the crime [13].

Second the deconstruction of causation (Causation): in traditional criminal jurisprudence (including Iraqi scholarship and jurisprudence), the occurrence of a harmful act is insufficient for liability to arise; a direct and immediate causal nexus between the actor's fault and the criminal result must also be established. The black box problem intervenes as an independent foreign factor severing this nexus: the decision to commit the digital criminal act was taken within the corridors of the autonomous

algorithm based on instantaneous interactive data, and was not an inevitable or direct consequence of the human's original programmatic codes. This 'causal severance' prevents law enforcement authorities from attributing the crime materially to any human element [8].

This structural opacity demonstrates that the autonomous agent dilemma is not merely a theoretical hypothesis, but a technical reality that incapacitates the traditional rules of criminal liability and renders the search for a human 'criminal intent' within the codes of the black box a form of legal illusion demanding a new punitive philosophy that transcends the dilemma of opacity to reach innovative rules of attribution [1].

Chapter Two: The Criminal Attribution Dilemma in Light of Iraqi Legislation (An Analytical Study of the Anti-Cybercrime Law)

This chapter aims to project the conceptual and technical data established in Chapter One onto the reality of the Iraqi legal system, examining the adequacy of the special penal provisions on cybercrime and the obstacles to applying the general penal rules in confronting the 'autonomous agent.'

Section One: The Adequacy of the Penal Provisions in the Iraqi Anti Cybercrime Law

The legislative environment in Iraq suffered for many years from a legal vacuum specific to cyber-crimes, limited to attempts to adapt the provisions of traditional penal law. With the emergence of draft 'Anti Cybercrime Law' texts, it has become necessary to analyse these provisions to determine their readiness to confront self-governing algorithmic systems.

Subsection One: Analysis of the Proposed Legal Provisions and Their Position on Spontaneous Offences

Upon reading the texts and formulations contained in the drafts of the Iraqi Anti Cybercrime Law particularly the versions deliberated in recent years it becomes evident that the Iraqi legislator drafted its provisions with a 'static' (rigid) view of technology. The criminalised acts have consistently been linked to a 'direct human actor' who employs a technical means [4].

This crisis manifests in the draft's use of verbs such as: (whoever enters, whoever obstructs, whoever designs, whoever uses, whoever intercepts) which, according to established linguistic and legal interpretation in Iraqi jurisprudence, are directed exclusively at purely human conduct, treating the computer or network as a pliant tool for executing the crime [7].

Accordingly, the formulation dilemma emerges through two axes:

First the absence of the dynamic dimension of autonomous algorithms: the proposed provisions in Iraqi law do not accommodate the hypothesis that a 'self-learning artificial intelligence system' is the one that autonomously 'entered' or 'obstructed' systems based on autonomous algorithmic processing not encompassed by the original programmer's instructions [3]. Since the Iraqi penal provision prohibits expansive interpretation or analogy in implementation of the principle of legality (*nullum crimen, nulla poena sine lege*), the Iraqi criminal judge will find himself compelled to acquit the

human accused (programmer/user) for lack of material conduct, while simultaneously being unable to indict the software system for the provision's failure to recognize it as an entity susceptible of prosecution [10].

Second the rigidity of legislative definitions: the Iraqi law draft is entirely devoid of definitions for modern technical terms such as 'autonomous systems' (Autonomous Systems), 'autonomous digital agents' (Autonomous Agents), or 'shared human machine liability.' This conceptual neglect renders the law defective with a structural shortcoming, regulating an advanced virtual environment with legal tools dating back to the philosophy of the traditional industrial revolution depriving the penal provision of its effectiveness against complex digital crimes [1].

Subsection Two: The Shortcoming of the Legislative Provision in Identifying Responsibility for 'Spontaneous Criminal Results' of Software

The legislative gap appears in its clearest form when autonomous digital systems produce 'spontaneous criminal outcomes' (Spontaneous Criminal Outcomes) outcomes generated from the process of the algorithm's self-learning and autonomous analysis, without having been predetermined in the programmer's plan or the user's desire. In this case, the Iraqi criminal judiciary stands before the dilemma of distributing criminal liability and identifying the 'person' who must bear the burden of the crime [18].

The shortcoming of Iraqi legislation in this regard manifests through three principal dimensions:

First the impossibility of applying the standard of moral perpetration to the parties of the digital industry: when an autonomous cyber intrusion or generation of legally prohibited digital content occurs through generative artificial intelligence, human roles are distributed among (the original programmer, the manufacturing company, the server host, and the end user). The Iraqi Anti Cybercrime Law contains no special provisions for 'digital joint liability' or 'allocation of fault shares' among these parties when the machine independently makes the decision [13]. Since criminal liability is personal under Iraqi constitutional and statutory law, neither the user can be punished for a criminal result caused by an algorithmic defect of which he was unaware, nor the programmer who lost control of his algorithm after its deployment in the digital space and its interaction with external data [11].

Second the absence of the concept of 'presumed fault' in Iraqi penal policy: traditional Iraqi criminal policy rests on the principle that fault must rest on material conduct proven by the public prosecution (negligence, recklessness). By contrast, the nature of 'spontaneous results of software' demands the adoption of the concept of 'presumed fault' or 'risk based liability' (Strict Liability) that is, holding the system's owner or programmer liable merely because they launched a dangerous, uncontrollable entity.[9] This form of objective liability is entirely absent from the Iraqi law draft, preventing the judiciary from prosecuting those responsible for creating digital risks unless the public prosecution proves the existence of general and specific intent which is technically impossible under the black box problem [19].

Third the incapacity to address precautionary measures and confiscation: if traditional punishment (imprisonment or fine) is impossible to apply to the autonomous agent (the software itself), the precautionary measures prescribed in Iraqi laws (such as confiscation of tools or closure of websites) are insufficient on their own to avert the danger.[20] The autonomous software may be dispersed across decentralized cloud networks (Decentralized Networks) or block chains (Block chain), rendering the traditional 'confiscation' or 'deletion' order issued by the Iraqi court technically unenforceable demonstrating the provision's shortcoming in devising qualitative digital penalties (such as code execution, or advanced cyber prohibition) [2].

This legislative incapacity proves that the spontaneous criminal results of software remain without penal coverage in Iraq not because no harmful act exists, but because the engineering of the proposed legal provision is incapable of identifying a 'human being' wearing the garb of the offender under the traditional concepts of penal attribution [4].

Section Two: Obstacles to Applying the General Provisions of the Iraqi Penal Code No. 111 of 1969

In 1969, the Iraqi legislator could not envision that the judiciary would one day confront a crime behind whose physical element no human stands. Accordingly, the general rules of liability were formulated rigidly, centred around the human being. The incapacity of these traditional rules is clearly manifested when attempting to characterize the forms of participation and criminal contribution in offences committed by the autonomous agent.

Subsection One: The Crisis of Proving 'Criminal Participation' (Principal and Accomplice) Between Humans and Intelligent Systems

Articles 47, 48, and 49 of the Iraqi Penal Code No. 111 of 1969 govern the provisions of criminal participation, dividing participants into two categories: the principal (who commits the offence alone or with others, or contributes to its execution when it consists of a series of acts), and the accomplice (who incites, agrees, or assists in preparatory or facilitative acts for the commission of the offence) [21].

This traditional legislative architecture collides with a thick technical wall when attempting to apply it to offences committed through autonomous self-learning algorithms, and the crisis manifests in two dimensions:

First the crisis of characterizing the 'innocent agent' (The Innocent Agent): according to the established traditional concept in Iraqi jurisprudence, the indirect perpetrator is one who drives a criminally irresponsible person (such as a person with a mental disorder or a discerning minor) to execute the crime.[7] Some may attempt to adapt this concept and regard the 'autonomous machine' as the innocent agent (the no rational instrument) and the programmer as the principal.

However, this characterization fails entirely, because the traditional innocent agent requires that he execute the instigator's 'will' and plan precisely, without addition or omission (such as the minor who is ordered to administer poison) [2]. In the case of autonomous systems, however, the system entirely transcends the programmer's plan,

producing innovative criminal conduct that the programmer's will had never encompassed or intended owing to the 'self-learning' problem and its complex neural relationships making it impossible to describe the programmer as a principal acting through an intermediary [6].

Second the impossibility of realizing 'animus societatis' for complicity: Articles 48 and 49 of the Iraqi Penal Code require, for criminal partnership (by incitement, agreement, or assistance), the existence of 'unity of criminal intent' or the mutual mental nexus between the accomplice and the principal that is, the accomplice's will must converge with the principal's will in realising the shared criminal result [10].

In the environment of the 'autonomous agent,' this mental cohesion collapses: if an Iraqi programmer formulates an intelligent code designed to manage and process medical data, and the system autonomously based on independent processing decides to hack and break into a competitor hospital's system and destroy its records, how can 'agreement' or 'incitement' be established? [15]. The machine cannot 'agree' with the human in the legal sense, as it lacks human consciousness; nor can the human 'assist' the machine in an act of whose occurrence and timing he was ignorant. Consequently, the severance of the mutual psychological nexus inevitably precludes the filing of a 'criminal partnership' charge against any human element, leaving the material criminal act without a principal and without an accomplice in Iraqi law [17].

This fragmentation of criminal participation elements demonstrates how the general rules of the Iraqi Penal Code stand shackled before the dissolution and dispersal of traditional offender roles amidst the complex cells of digital space making room for the emergence of new patterns of crime that structurally escape the net of penal justice [12].

Subsection Two: Challenges to Proving Criminal Intent (Knowledge and Will) of the Programmer or Iraqi User for Errors of Complex Systems

The criminal attribution crisis in Iraqi law culminates when the public prosecution and the judiciary collide with the two elements of the mental element: 'knowledge' and 'will' on the part of the programmer or human user. If the established general rules in Iraqi criminal jurisprudence require, for a person's conviction under intentional crimes, that he be fully aware of all material facts of the crime and have directed his will towards producing the result, then the architecture of complex digital systems renders such proof practically and technically impossible [2].

These challenges manifest in the Iraqi judicial environment through two pivotal aspects:

First the impossibility of proving 'actual knowledge' and its dissolution in the environment of deep systems: Iraqi law requires that the actor know that his conduct will lead to an assault on an interest protected by law [21]. In the context of algorithmic systems based on neural networks and continuous decentralized updating through cloud computing, this knowledge is entirely absent. When an Iraqi programmer designs an intelligent algorithm and deploys it in the digital environment, the system begins

autonomously modifying its codes based on variable inputs over which the programmer has no control or awareness [16].

If the system, of its own accord, breaches Iraqi banking databases to gather information, the 'actual knowledge' of the act, its timing, and its mechanism is entirely absent from both the programmer and the user. The criminal trial here collides with the absence of human awareness of the conduct at the moment of its commission, inevitably driving the court to acquit for lack of the mental element, based on general rules that do not punish hypothetical or impossible knowledge in intentional crimes [4].

Second the severance of the element of 'volition' (Volition) through autonomous machine decision-making: knowledge alone is insufficient; the offender's will must also mobilize to execute the act and achieve the criminal result.[10] In the autonomous agent dilemma, the 'machine decision-making' problem (Machine Decision-making) manifests: the algorithm is the one whose mathematical (digital) will directed itself towards committing the harmful act as the sole means according to its autonomous data analysis of solving a particular problem or achieving a general programmatic objective [8].

This autonomous machine orientation severs the psychological nexus between the human and the criminal result; it cannot legally be said that the programmer's or user's will was directed towards producing this specific result, for their will had been directed solely towards creating an advanced intelligent tool.[11] This vast divergence between the human's will (directed towards benefit and utility) and the machine's will (which spontaneously deviated towards crime) makes it impossible for the Iraqi criminal judge to apply the provisions on criminal intent contained in Penal Code No. 111 of 1969 once again confirming that we face an attribution vacuum portending total impunity due to the incompatibility of the provisions with the data of the contemporary digital space.[13]

Chapter Three: Prospective Visions and Alternative Models of Liability Attribution (A Comparative Study)

Having analytically demonstrated the incapacity of the traditional rules of Iraqi criminal law to accommodate the 'shift of liability' arising from the autonomous agent's conduct, it becomes imperative not to halt at the threshold of legislative criticism, but to advance towards surveying and developing the globally innovated legal models. This chapter aims to present a prospective comparative treatment from which the Iraqi legislator can draw guidance to reformulate penal policy in the digital space.

Section One: International Comparative Models for Attributing Criminal Liability for Artificial Intelligence Acts

Contemporary criminal scholarship in Western schools (Anglo-American and Latin) has attempted to overcome the dilemma of absent human intent by formulating hypothetical attribution models that transfer the burden of liability through innovative methods. These models oscillate between maintaining the inevitability of the human element behind technology, and recognizing the complete autonomy of the digital entity.

Subsection One: The Vicarious Liability Model (Vicarious Liability) and the Natural and Probable Consequences Doctrine

The vicarious or derivative liability model is one of the most prominent legal solutions proposed to address the autonomous agent dilemma without the need to demolish the foundations of traditional criminal law. This model borrows its philosophy from the provisions of 'a principal's liability for the acts of an agent' established in civil laws, but develops them to suit the criminal cyber nature [6].

This model which has found wide resonance in American and British judicial deliberations rests on the legal premise that the 'autonomous digital system' functions as an 'agent' (Agent) on behalf of the programmer or user (Principal); since this digital entity was deployed to serve this human person's interests and upon his authorization, the criminal acts committed by the system while performing its functions may be criminally attributed to the human, even if the latter's direct will was not directed towards committing the specific crime [11].

To discipline this model and prevent it from becoming arbitrary punishment that collides with the principle of personal liability, it has been supplemented with the '**Natural and Probable Consequences Doctrine**,' whose dimensions can be summarized as follows:

First the standard of latent risk in the source code: if a programmer formulates an intelligent algorithm designed to collect and analyse financial data from the open network, and it was natural and probable according to the ordinary course of events and due to the machine's self-learning flexibility that this algorithm would interface with encrypted systems and hack them, then the crime arising is considered a probable result of the risks the programmer accepted in launching it [16]. Here, the human is not prosecuted for direct criminal intent, but for his responsibility for 'risk creation' (Risk Creation) and failure to place adequate restraints preventing the machine's autonomous deviation.[2]

Second avoiding the problem of causal severance: this model is distinguished by resolving the knot confronting Iraqi legislation in Chapter Two; it prevents the 'machine's will' from severing the causal nexus. Under vicarious liability, the machine's autonomous conduct is regarded not as a foreign factor, but as a natural and dynamic extension of the initial risk commenced by the human allowing the judge to keep the thread of criminal attribution connected to the machine's 'human director' [1].

.Nevertheless, this model faces sharp academic criticism: it closely approaches the concept of 'strict liability' (Strict Liability), which punishes a person without conclusive proof of personal fault which may raise constitutional problems if it were to be introduced into the general rules of the Iraqi Penal Code without precise and extensive adaptation [13].

Subsection Two: Trends Towards Granting 'Electronic Legal Personality' (Electronic Personality) and Their Risks

Confronted with the growing instances of digital systems' autonomy and their transcendence of their human creators' plans, a radical doctrinal and legislative trend has emerged calling for an end to the forced search for a 'natural person' to bear the burden of an algorithmically produced crime. Instead, this trend advocates the necessity of recognizing **'Electronic Legal Personality' (Electronic Personality)** for autonomous intelligent systems, treating them as independent legal entities capable of bearing rights, obligations, and liabilities analogous to the successful history of the concept of 'legal personality' for companies and institutions [22].

This trend attracted significant international attention, crystallized in the landmark report issued by the European Parliament in 2017, which included recommendations on the possibility of granting sophisticated robots and intelligent systems a special legal status as 'electronic persons' [23].

Under this model, the course of criminal attribution changes entirely: the digital crime committed by the autonomous agent is attributed directly to it as a 'digital person,' and traditional penalties are deconstructed and replaced by sanctions compatible with its technical nature, such as: (compulsory reprogramming, cyber prohibition, freezing of digital and algorithmic assets, or programmatic execution through permanent deletion of the code) [20].

Nonetheless, this revolutionary thesis has faced an arsenal of philosophical and legal criticisms and risks that make it difficult to adapt in states with conservative criminal policies such as Iraq:

First the undermining of the deterrent and moral character of criminal law: criminal law rests fundamentally on the concept of deterrence and moral culpability, which only a human being with consciousness and conscience can experience [2]. Granting legal personality to the machine means stripping punishment of its philosophical meaning; the machine feels no pain, remorse, or social stigma resulting from a criminal judgment, reducing criminal justice to mere 'hollow technical procedures' that deprive society of its confidence in the function of punishment [17].

Second the creation of a dubious 'immunity shield' for programmers and companies: this is the most dangerous point in practice; recognizing the independent electronic personality of the machine will be exploited as a 'safe harbor' (Safe Harbor) by major technology companies and programmers to evade liability [24]. Whenever a major cyber-crime occurs (such as the algorithmic hacking of public funds), developers will claim that the system independently made the decision and possesses a separate personality closing the door on criminal prosecution against the actual human actors who profit from these systems, representing the apex of the liability shift dilemma [7].

Third the machine's lack of a truly independent financial estate: the success of the idea of corporate legal personality rested on the existence of a genuinely independent financial estate capable of compensating for damages and paying fines. As for the digital

entity, even if equipped with an independent digital wallet (such as encrypted assets), its feeding and management remain ultimately linked to human sources rendering financial autonomy a mere technical illusion that collapses upon the first attempt at penal judicial enforcement [9].

Section Two: The Proposed Strategy for Developing the Iraqi Legislative Approach

Confronting the 'autonomous agent' dilemma in the Iraqi judicial environment does not require the adoption of fantastical theories such as granting legal personality to machines, but demands an intelligent modernization of national criminal policy. The legislator must move towards reformulating attribution rules to transition from tracking the 'absent human intent' to tracking 'unjustified technical risk.'

Subsection One: Required Amendments to Iraqi Criminal Policy (Adopting Objective or Presumed Liability in Digital Systems)

To enable the Iraqi judiciary to prosecute autonomous digital crimes, we propose that the Iraqi legislator introduce a fundamental amendment to the draft Anti Cybercrime Law, incorporating a framed departure from the general rules of Penal Code No. 111 of 1969, through the adoption of **'a system of presumed liability based on risk creation** [4]. This legislative proposal is embodied in the following mechanisms:

First the establishment of a new criminal culpability termed 'digital oversight fault': we propose the formulation of a new legal article in the Iraqi Cybercrime Law stipulating: **'Shall be punished by imprisonment or a fine any person who designed, developed, or used a self-learning digital system and was negligent in placing the technical restraints necessary to prevent its deviation, if such negligence resulted in the system engaging in conduct constituting a crime under this Law.'**()

This provision shifts the focus of criminal accounting: the judge will no longer search for the 'programmer's intent to breach a bank or government website' (an intent that is nonexistent), but will try him for a positive act the 'release of a dangerous, opaque system into the digital space without exercising due diligence' (Due Diligence) [11]. The black box problem is here absorbed through the presumption of fault on the part of the human who created the initial risk [2].

Second shifting the burden of proof (Shifting the Burden of Proof): under the Iraqi general rules, the burden of proving fault falls on the public prosecution which is technically incapable of proving what occurred inside the black box in complex digital crimes [19]. We therefore propose that the new provision include a 'reversal of the burden of proof,' such that the programmer's or developing company's fault is presumed upon the occurrence of the automatic criminal result from the machine, and the human developer can rebut this liability only by technically and forensically proving that he adopted all international standard measures of digital security and caution, and that the machine's deviation was the result of a technically irresistible cyber force majeure [24].

Third harmonizing the provision with the constitutional principle of legality of punishment: adopting presumed liability in this domain does not collide with the Iraqi Constitution, which provides that 'criminal liability is personal,' because the punishment

here will not fall on a person for another's act (the machine's), but will fall on him for 'his personal fault consisting in negligence in controlling and directing the will of the machine he created' preserving the moral and deterrent character of Iraqi criminal law while affording it superior cyber flexibility [7].

Subsection Two: Activating Digital Precautionary Measures and Corporate Criminal Liability of Developing Companies

The prosecution of natural persons (such as individual programmers) in confronting autonomous agent crimes may fail to achieve the required deterrence, given that these systems are typically developed, managed, and marketed by institutional entities and giant software companies possessing enormous financial and technical capabilities. Hence, the necessity emerges of activating the 'criminal liability of legal persons' and developing 'digital precautionary measures' in the Iraqi legislative environment [12].

This activation requires adopting a legislative strategy resting on two pillars:

First activating the criminal liability of the digital legal person: the Iraqi Penal Code No. 111 of 1969 establishes criminal liability for legal persons in Article 80 thereof; however, this provision remains inadequate to accommodate the virtual nature of cyber companies [21]. We propose the incorporation of a special provision in the Iraqi Anti Cybercrime Law establishing direct corporate criminal liability if the autonomous algorithmic crime was committed in the company's interest, or due to the absence of internal 'digital governance' policies [2].

Holding companies criminally liable (through hefty financial fines, prohibition from practising digital activity, or closure of cloud platforms) represents an effective solution; since companies are best positioned to bear the cost of error and most incentivized to develop 'autonomous deviation resistant' systems to avoid criminal stigma and economic losses [1].

Second introducing 'digital precautionary measures' (Digital Precautionary Measures): to transcend the futility of traditional penalties in cyberspace, qualitative precautionary measures compatible with the nature of the autonomous agent must be innovated [13]. These proposed measures for Iraqi law include:

1. Temporary or permanent programmatic disabling (Algorithmic Quarantine): an injunctive measure by which the deviant algorithm is isolated from the network to immediately cut off its danger [15].
2. Revocation of the artificial intelligence licence (Revocation of Digital License): obligating companies to withdraw the self-learning code from the Iraqi virtual market and invalidate its operating authority [5].
3. The imposition of 'human in the loop mandates' (Human in the loop Mandate): a judicial precautionary measure obligating the company to reengineer the system such that no financial or security decision is executed completely independently, but must pass through a human endorsement window to break the machine's autonomy during critical periods [4].

Organising these measures grants the Iraqi judiciary practical confrontational tools and transitions criminal policy from the phase of 'the impossible after the fact punishment' to the phase of 'effective preventive precaution.'

Discussion

The findings of this study demonstrate that the emergence of autonomous, self-learning systems creates a fundamental challenge for conventional criminal-law doctrines because the traditional connection between human conduct, criminal intent, and harmful consequences becomes increasingly difficult to establish. The shift from technology as a passive instrument to technology as an autonomous agent means that harmful digital conduct may result from decisions generated independently by algorithmic systems rather than from direct human instructions. This condition is further complicated by the Black Box problem, which limits the ability of programmers, users, investigators, and courts to reconstruct the reasoning process that produced a particular algorithmic decision. Consequently, the requirements of **actus reus**, **mens rea**, causation, foreseeability, and criminal participation become difficult to apply when the harmful outcome cannot be directly attributed to the knowledge or will of a human actor. In the Iraqi context, this problem is particularly significant because the Penal Code No. 111 of 1969 and the draft Anti Cybercrime Law remain largely structured around the assumption that criminal conduct is committed by a human actor who consciously uses technology as an instrument. Such an approach is insufficient when an autonomous system independently modifies its behavior, selects targets, or produces criminal consequences that were neither specifically intended nor reasonably foreseeable by its developer or user. The study therefore indicates that relying exclusively on traditional concepts of direct intent, negligence, principal liability, and complicity may generate an impunity gap in cases involving autonomous digital systems. At the same time, granting independent criminal personality to artificial intelligence presents considerable conceptual and practical difficulties, particularly because machines lack moral consciousness, genuine culpability, and an independent capacity to experience punishment. A more suitable approach for Iraqi criminal policy is therefore to focus on risk creation, digital oversight, corporate responsibility, and preventive control rather than attempting to artificially attribute human-like intent to autonomous systems. Presumed liability based on failures of reasonable digital supervision, combined with the criminal liability of developing companies and specialized precautionary measures such as algorithmic quarantine, mandatory human oversight, and restrictions on the operation of high-risk systems, may provide a more realistic legal response. Such reform would preserve the principle of personal criminal responsibility while adapting penal policy to the technological realities of autonomous decision-making and reducing the possibility that serious algorithmic harms remain outside the reach of criminal justice.

CONCLUSION

Fundamental Finding: The fracturing of the physical and mental elements: the research has demonstrated that modern digital space technology has transitioned from

the status of 'tool' to that of 'autonomous agent' through selflearning algorithms, leading to the shift of criminal liability and the dissolution of criminal intent and causal nexus in human terms. **Implication:** Urgent legislative drafting update: we recommend that the Iraqi Parliament incorporate newly developed procedural definitions in the Cybercrime Law draft, encompassing autonomous systems, digital agents, and the black box. Adoption of presumed liability: explicit statutory provision for the criminal liability of the programmer or user based on 'presumed fault in digital oversight and precaution,' and the reversal of the burden of proof in favour of the public prosecution upon the occurrence of automatic criminal results from the machine. Adoption of qualitative cybermeasures: abandonment of traditional penalties inadequate for the virtual space, and their replacement with intelligent precautionary measures such as 'algorithmic quarantine,' 'programmable execution of deviant code,' and extension of financial punishment to legal persons (developing companies). **Limitation:** The incapacity of Iraqi general rules: the Iraqi Penal Code of 1969 suffers from structural rigidity preventing the accommodation of criminal participation concepts (principal and accomplice) or the innocent agent when the machine's autonomous will intervenes. The shortcoming of the Cybercrime Law draft: the study revealed that Iraq's draft AntiCybercrime Law texts were formulated with a traditional perspective, devoid of regulation of algorithmic errors and lacking provisions for 'presumed liability' or 'digital legal person liability.' **Future Research:** The original text does not contain an explicit sentence concerning future research, so no sentence can be placed in this section without adding or modifying the original content.

REFERENCES

- [1] M. Schérer, "Shift of liability in the digital space: Re-evaluating the 'independent agent' dilemma," *European Journal of Crime, Criminal Law and Criminal Justice*, vol. 32, no. 1, pp. 12-38, 2024.
- [2] K. M. Mansour, *The General Theory of Digital Crime: The Shift of Traditional Concepts of Criminal Intent and the Principal Actor*. Cairo, Egypt: Dar Al-Nahda Al-Arabiya, 2024.
- [3] A. Bartha, *Criminal Responsibility of Autonomous Systems: A New Paradigm in Cyber-Criminal Law*. Routledge, 2021.
- [4] S. H. Al-'Amiri, "Challenges facing the Iraqi criminal judiciary in proving cybercrime," *Journal of Legal Sciences, University of Baghdad*, vol. 38, no. 1, pp. 70-130, 2023.
- [5] V. Radutskiy, "Criminal law dimensions of autonomous cyber-physical systems actions," *Journal of Digital Forensic Practice & Law*, vol. 19, no. 2, pp. 89-107, 2025.
- [6] G. Hallevy, *When Robots Kill: Artificial Intelligence under Criminal Law*. Boston, MA, USA: Northeastern University Press, 2013.
- [7] D. J. Al-'Ukayli, "Criminal liability for artificial intelligence crimes: A comparative study," *Al-Muhaqqiq Al-Hilli Journal for Legal and Political Sciences*, vol. 13, no. 2, pp. 110-175, 2021.
- [8] B. Casey, "Challenging the doctrine of mens rea in the age of self-learning algorithms," *Southern California Law Review*, vol. 98, no. 3, pp. 501-534, 2025.

- [9] T. C. King, N. Aggarwal, M. Taddeo, and L. Floridi, "Artificial intelligence crime: An overview of malicious uses and countermeasures," *Journal of Cyber Policy*, vol. 5, no. 2, pp. 231-255, 2020.
- [10] T. S. Al-Shammaa, "Information technology crimes and their effect on criminal attribution rules," *Al-Rafidain Journal of Law, University of Mosul*, vol. 24, no. 76, pp. 40-82, 2022.
- [11] D. C. Vladeck, "Machines without principals: Liability rules and artificial intelligence," *Washington Law Review*, vol. 97, no. 1, pp. 115-152, 2022.
- [12] S. Yanisky-Ravid and S. Hallisey, "Equality and AI criminal liability: Scaling the corporate model to autonomous systems," *Fordham Intellectual Property, Media and Entertainment Law Journal*, vol. 31, no. 3, pp. 677-724, 2021.
- [13] H. F. Al-Janabi, *Civil and Criminal Liability for Errors of Autonomous Intelligent Systems: A Study in Light of Iraqi and French Law*. Beirut, Lebanon: Zain Legal Publishing, 2023.
- [14] U. Pagallo, *The Laws of Robots: Crimes, Contracts, and Torts*. Springer, 2013.
- [15] C. Duffy and N. McDonnell, "Malicious autonomous agents: The shift from tool to actor in cyberattacks," *Computers & Security*, vol. 114, Art. no. 102118, 2022.
- [16] Y. Bathaee, "The artificial intelligence black box and the failure of intent and causation," *Harvard Journal of Law & Technology*, vol. 31, no. 2, pp. 889-938, 2018.
- [17] G. Sartor and A. Omicini, "The autonomy of technological agents and the law: From soft computing to hard responsibility," *Artificial Intelligence and Law*, vol. 30, no. 1, pp. 45-73, 2022.
- [18] S. Gless, E. Silverman, and T. Weigend, "If robots do the crime, who pays the price? Crucial questions of criminal liability for autonomous systems," *Criminal Law Forum*, vol. 27, no. 2, pp. 213-246, 2016.
- [19] M. M. Al-Bashir, "The technical and legal problem of the presumed actor in cyber crimes: An analytical study in light of Arab legislation," *International Journal of Law, Qatar University*, no. 3, pp. 95-121, 2022.
- [20] G. Hallevy, "The criminal liability of artificial intelligence entities – From science fiction to legal social control," *Akron Law Review*, vol. 47, no. 2, pp. 5-35, 2021.
- [21] Iraqi Penal Code No. 111 of 1969, as amended, *Iraqi Gazette*, Baghdad, Iraq, 1969.
- [22] S. Chesterman, "Artificial intelligence and the limits of legal personality," *International & Comparative Law Quarterly*, vol. 69, no. 4, pp. 819-844, 2020.
- [23] European Parliament, "Report with recommendations to the Commission on a civil liability regime for artificial intelligence (2020/2014(INL))," *European Parliament*, 2020.
- [24] M. Zalnieriute, "The automated state and criminal justice: Shifting the responsibility burden," *Law, Innovation and Technology*, vol. 15, no. 1, pp. 143-169, 2023.

* Nadhim Jawad Rashid Al-Mamore (Corresponding Author)
