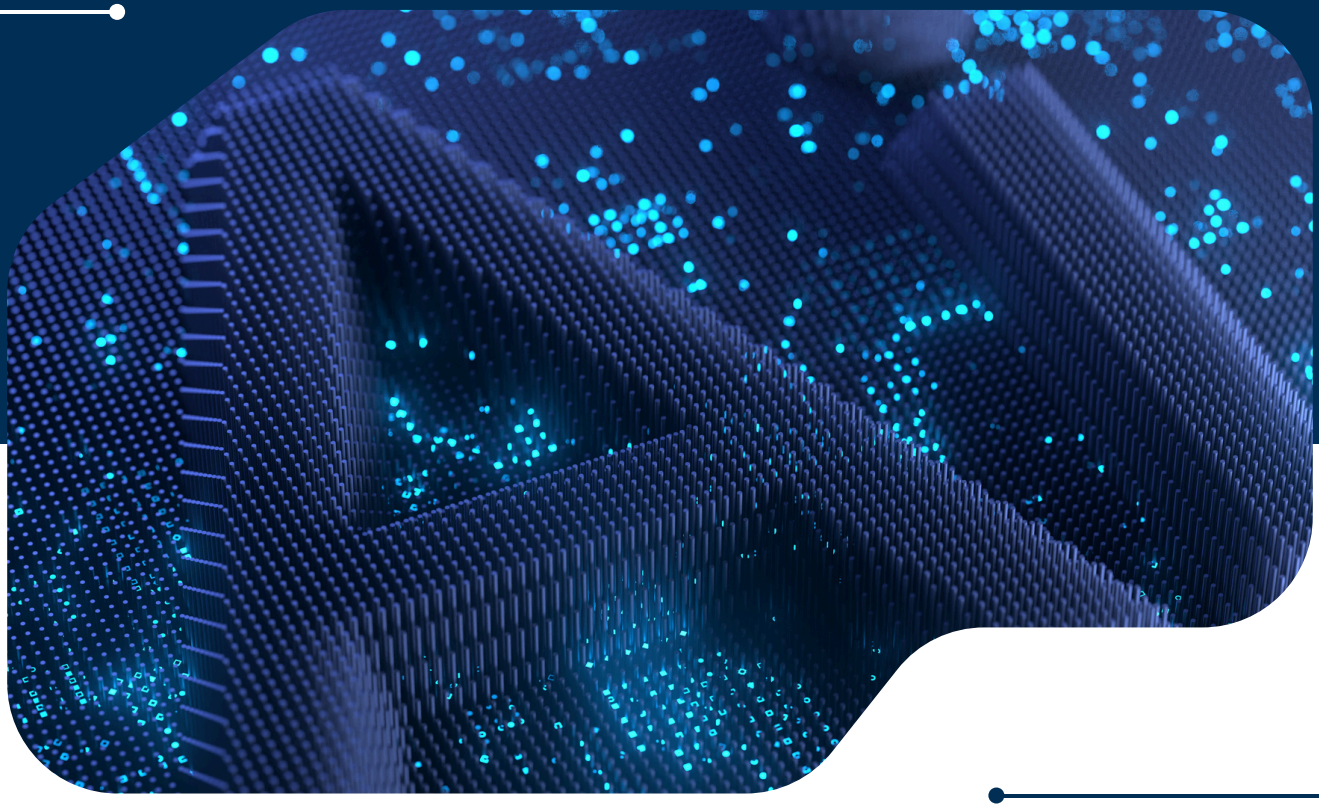

CRIMINAL LIABILITY AND ARTIFICIAL INTELLIGENCE: LEGAL PROFILES AND OUTSTANDING DOCTRINAL QUESTIONS

Position Paper

GEBBIABORTOLOTTO
PENALISTI ASSOCIATI

July 2025
www.gbpenalisti.it

TABLE OF CONTENTS

Introduction....3

Relevance of the Issue in the Current Legal Framework....3

The Decision-Making Autonomy of AI: Machina Delinquere non Potest....5

Aspects of Individual Criminal Liability....7

The Legal Personae of Provider and Deployer....7

Intentional Liability and Data Machine Training....8

Attribution of Criminal Negligence to Providers and Deployers....10

Corporate Administrative Liability under Legislative Decre No. 231/2001....12

Introduction....12

The Organisational, Management and Control Model....14

Risk assessment....16

Codes of Conduct and the Regulation of Artificial Intelligence Technologies....17

INTRODUCTION

Relevance of the Issue in the Current Legal Framework

In accordance with a definition widely recognised within the international legal and academic communities, artificial intelligence is to be understood as a specialised domain of computer science, devoted to the examination of the theoretical foundations, algorithmic methodologies, and engineering techniques underpinning the design of hardware and software systems capable of performing tasks which, on their face, would appear to fall within the exclusive province of human cognitive faculties [1].

Such systems, by virtue of autonomous data processing, machine learning capabilities, and the progressive refinement of outputs, become apt to simulate behaviours that are intelligently directed toward specific, predetermined objectives [2].

The exponential evolution of intelligent technologies, coupled with their pervasive penetration across the most diverse sectors of human activity, necessitates a profound, critical, and systematic reflection on the impact such innovations exert upon the domain of legal regulation.

In the present analysis, particular attention shall be devoted to the interaction between artificial intelligence and substantive criminal law – a field wherein the emergence of artificial agents endowed with increasing operational autonomy has produced a manifest tension with the traditional dogmatic constructs of subjective attribution and individual criminal responsibility.

Among the most debated issues – beyond the controversial question of whether artificial entities may conceivably be regarded as potential perpetrators of criminal offences – lie further inquiries of systematic relevance. These concern the extent to which criminal liability may be attributed to traditional legal subjects, namely natural or legal persons who, in various capacities, participate in the chain of conception, development, distribution, and deployment of artificial intelligence systems.

[1] M. Somalvico, *Intelligenza artificiale*, Milan, 1987.
[2] Vd. il report *ID R&D Human or Machine: AI Proves Best at Spotting Biometric Attacks*, 2022.

The difficulty, in such instances, resides in the precise identification of the technological segment wherein a criminally relevant act may be situated, as well as in the determination of whether such conduct fulfils the elements of a statutory offence that is culpable and injurious, in accordance with the foundational principles of legality, material harm, and the personal nature of criminal liability.

The reconstruction of criminal liability is rendered even more complex by certain intrinsic features of artificial intelligence systems, which pose unprecedented challenges to traditional frameworks of criminal causality. In particular, the causal chain in instances of malfunction arising from intelligent systems is often markedly intricate, and the identification of the individual to whom the harmful event may be imputed necessitates a detailed analysis of the role assumed by each actor within the processes of design, training, and deployment of the system.

This task is further complicated by the alleged decision-making autonomy of machine learning-based systems, which - owing to their capacity for self-learning and adaptive behaviour [3] - are, in the view of some of the more audacious strands of legal doctrine, potentially capable of severing the causal nexus between the programmer's initial choices and the ensuing harmful event.

Nevertheless, it is well established in criminal law that only events which are truly unforeseeable and unavoidable are capable of breaking the chain of causation. Conversely, the so-called "generically foreseeable unpredictability" [4] of algorithmic conduct imposes upon the human agent a heightened duty of prevention and risk management, thereby rendering increasingly tenuous the distinction between mere technical error and legally relevant negligence.

[3] S. RUSSELL- P, NORVIG, *Artificial Intelligence: A Modern Approach*, Pearson College Div., 4th ed., 2020, p. 651 ss.
[4] C. PIERGALLINI, *Intelligenza artificiale: da "mezzo" ad "autore" del reato?* cit., p. 1762.

The Decision-Making Autonomy of AI: *machina delinquere non potest*

Within the contemporary legal discourse, the decision-making autonomy of artificial intelligence systems is frequently invoked as a potential basis for positing the direct liability of the machine, in relation to unlawful acts to which it contributes, whether wholly or in part [5].

However, such an approach risks engendering a perilous conceptual ambiguity - particularly within the realm of criminal law - wherein the notion of liability is inextricably bound to anthropocentric categories such as volition, awareness, and culpability.

Criminal law, more so than any other branch of the legal order, was originally conceived for natural persons, insofar as it is only the human being who is deemed capable of self-determination, moral blameworthiness, and susceptibility to just punishment. From this perspective, the proposition that a machine might be held criminally liable for its "actions" stands in direct contradiction to a principle which, while perhaps implicit, remains deeply entrenched within the criminal tradition: *machina delinquere (et puniri) non potest*.

Paraphrasing the classical formulation once employed to deny criminal liability to legal persons - *societas delinquere non potest* - it may now be affirmed that the criminal justice system does not, either in *abstracto* or *in concreto*, recognise any autonomous criminal legal subjectivity in respect of artificial intelligences, robots, or algorithmic systems.

Accordingly, even in circumstances where an artificial intelligence system engages in conduct that may, in abstract terms, be subsumed under the definition of criminal offence, the legal order does not envisage any form of direct criminal liability attributable to the machine itself. Rather, such an occurrence may, at most, give rise to a form of mediated or vicarious liability on the part of a human agent, in accordance with an imputative model firmly rooted in criminal law doctrine. Indeed, within such a framework, the artificial system is to be regarded as a mere instrument in the hands of the true perpetrator of the offence [6].

[5] The principal proponent of the theoretical framework supporting the potential attribution of directi criminal liability to artificial intelligence systems is the Israeli criminal law scholar Gabriel Hallevy, see G. HALLEVY, *Liability for Crimes Involving Artificial Intelligence Systams*, Springer, 2015.

[6] Per tutti, cfr. PAGALLO, *The adventures of Picciotto Roboto*, p. 352-353.

Therefore, no legal system has, to date, recognised the existence of an autonomous form of criminal liability attributable to artificial intelligence as such. On the contrary, any criminally relevant reconstruction of unlawful conduct committed through the act, or the omission of an intelligent system necessarily presupposes, and indeed requires, the antecedent act or omission of a human subject – one who is imputable, aware, and blameworthy.

Against this conceptual backdrop, the decision-making autonomy of artificial intelligence – while undoubtedly significant from a technical and operational standpoint – cannot, in its current form, be regarded as a juridically relevant source of autonomous liability. At least for the time being, such autonomy remains legally inconsequential, unless and until it is accompanied by a profound conceptual revolution capable of redefining the very foundations of criminal law, the structure of which remains, as of today, inextricably anthropocentric.

ASPECTS OF INDIVIDUAL CRIMINAL LIABILITY

The Legal Personae of Provider and Deployer

In an effort to delineate the contours of criminal liability connected to the use of AI systems, the roles of the provider and the deployer – as defined by Regulation (EU) 1689/2024, also known as AI Act – assume primary significance [7]. The provider is the entity that designs, develops, or commissions the development of an AI system under its own responsibility, thereby assuming the obligation to ensure its compliance from the very inception phase [8]. Such a role is thus vested with obligations of a predominantly technical and design-related nature, which impact the system's architecture, purpose, and safety assurances even prior to its deployment. Conversely, the developer is the party who within the scope of their professional or institutional activity, activates the system, concretely determining its modalities of application, assuming operational control, and directly influencing its interaction with the factual environment in which it is utilized [9].

The conceptual and normative distinction between these figures does not merely hold classificatory significance, but rather constitutes a crucial hermeneutic nexus in the analysis of subjective attribution profiles and in the determination of criteria for assigning criminal liability. Indeed, the structurally distinct nature of the activities respectively performed by the provider and the deployer entails a differing degree of causal involvement in the genesis of the harmful event, as well as a heterogeneity in their respective spheres of operational control, in accordance with the principles of culpability and offensiveness.

[7] Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence and amending certain Union legislative acts, COM/2021/206 final, 21 April 2021 (so-called AI Act).

[8] European Parliament and Council of the European Union (2024). Regulation (EU) 2024/1689 on Artificial Intelligence (AI Act), Article 3(3). Official Journal of the European Union.

[9] European Parliament and Council of the European Union (2024). Regulation (EU) 2024/1689 on Artificial Intelligence (AI Act), Article 3(4). Official Journal of the European Union.

Intentional Liability and Data *Machine Training*

Amidst the framework of criminally relevant liabilities connected to the use of artificial intelligence systems, the hypothesis of malicious use of AI raises particularly salient issues, especially concerning the process of machine learning training, namely that essential phase during which the algorithm is instructed through the processing of vast amounts of data.

In this context, the quality, completeness, and veracity of the data employed assume a decisive role in determining the reliability and accuracy of the results generated by the system. Indeed, any intentionally manipulative interference with such datasets constitutes wilful conduct whereby the machine, far from being an autonomous subject, reveals itself as an operative instrumentality of the human agent.

Hence, the most discerning criminal law scholarship does not hesitate to acknowledge that, within the scope of intentional offenses, the subjective element of imputability encounters no substantial theoretical impediments, insofar as the consciously pursued unlawful intent remains the central pivot of the wrongdoing, rendering immaterial the nature – whether traditional or technologically advanced – of the means employed to effectuate the crime. Under this perspective, the intelligent system continues to be regarded as *intrumentum scleris*, an executive vehicle for the criminal input formulated upstream [10].

However, the increasing sophistication of intelligent systems, as well as the technical insidiousness they may exhibit by virtue of their modes of operation, has prompted the legislator to intervene normatively through the introduction of Senate Bill No. 1146, which remains pending approval. Specifically, Article 26, paragraph 1(a) of the Bill envisages the introduction of a common aggravating circumstance, applicable to any criminal offence, where such offence is committed through the use of artificial intelligence systems that, by their nature or operational modalities, have constituted an insidious instrument, impeded public or private defence, or aggravated the consequences of the offence.

[10] P. SEVERINO, *Le implicazioni dell'intelligenza artificiale nel campo del diritto con particolare riferimento al diritto penale*, in P. SEVERINO (a cura di), *Intelligenza artificiale, politica, economia, diritto, tecnologia*, Luiss Uni Press, 2022.

This legislative provision not only represents a significant indication of the growing recognition of the criminal relevance of the intentional use of artificial intelligence systems but also demands renewed scrutiny of the subjective element of the offence, as well as the objective dangerousness of the means employed.

Attribution of Criminal Negligence to Providers and Deployers

Within the contours of contemporary legal regulation, the issue of negligent criminal liability arising from the use of artificial intelligence-based tools demands a structurally rigorous dogmatic analysis, centred on the roles of the provider and the deployer [11].

With regard to the provider, criminal liability arises, in the first instance, within a conceptual structure referable to the categories of product-related negligence and fault in the exercise of inherently hazardous activities - paradigms well-established in criminal doctrine and recognised within European legal systems.

Artificial intelligence, particularly in its more autonomous functional manifestations, by its very nature constitutes a high-risk technological activity. Consequently, within this context, the provider occupies a legally qualified position of guarantor, the source of which is to be found not only in the general principles of criminal law, but also in the more recent provisions of the AI Act, which sets forth specific obligations relating to the design, validation, and ongoing monitoring of AI systems on the part of the provider.

However, the hyper-complex, stratified, and fragmented structure of the artificial intelligence development chain – frequently distributed across numerous actors and segmented into iterative and interdependent phases – renders the establishment of a linear and reliable causal nexus particularly arduous, especially in terms of tracing negligent conduct back to a specific harmful outcome [12].

Indeed, the lack of algorithmic transparency, the opacity of machine learning architectures, and the emergent unpredictability inherent in the adaptive behaviours of AI systems give rise to a veritable decision-making black box, which stands in stark contrast to the principles of legality and personal culpability.

[11] European Parliament and Council of the European Union (2024). Regulation (EU) 2024/1689 on Artificial Intelligence (AI Act), Article 25. Official Journal of the European Union.
[12] I. SALVADORI, *Agenti artificiali, opacità tecnologica e distribuzione della responsabilità penale*, 2021. P. 83.

In fact, from this perspective, the negligent liability of the provider may only be deemed to subsist where the breach of specific precautionary obligations can be established with evidentiary rigour – such obligations being grounded in sector – specific regulations, technical guidelines, recognised industry standards, and consolidated best practices.

In other words, criminally relevant negligence must coincide with a serious and unjustifiable underestimation of residual risk - that is, with conduct marked by carelessness or imprudence, which exceeds the threshold of permissible risk. The boundaries of such a threshold are to be delineated by technical and regulatory precautionary norms, which serve as concrete sources of the duty of technological diligence. In the absence of conduct exhibiting such a degree of blameworthiness, no finding of subjective reproach consistent with the principles of harmfulness, proportionality, and culpability may be sustained.

Distinct - though no less problematic - is the matter of the deployer's negligent criminal liability, whose duty of care and position of guarantor, even more so than that of the provider, remains contingent upon future legislative determinations concerning the nature and scope of the supervisory obligations that may be imposed. Indeed, the rationale underpinning the European regulatory framework generally requires that any activity potentially harmful to fundamental rights be exercised under conditions of meaningful human oversight.

Yet this approach risks engendering a criminal control dilemma, insofar as the deployer - though formally vested with a preventive authority - may in practice confront autonomous systems whose internal functioning, driven by machine learning, adaptivity, and self-optimisation, is opaque, unpredictable, and in extreme cases, effectively uncontrollable. Thus, one risks attributing liability based on formal position alone, in the absence of any real capacity to intervene - thereby undermining both the principle of personal culpability and the protective function of criminal law.

Accordingly, to avoid reducing the human operator to a mere scapegoat, the attribution of negligent liability must be confined to cases of gross negligence or qualified omission, where a clear breach of duty can be substantiated [13].

[13] On the principle of the legality of negligence, see . F. GIUNTA, *La legalità della colpa*, In *Criminalia*, 2008, p.149.

CORPORATE ADMINISTRATIVE LIABILITY UNDER LEGISLATIVE DECRE NO. 231/2001

Introduction

The legal framework introduced by Legislative Decree No. 231/2001 marks a decisive departure from the traditionally individualistic approach that has historically characterised criminal law, establishing an autonomous form of liability attributable to collective entities.

Indeed, liability is no longer conceived as an automatic derivative of the conduct of the individual perpetrator, but rather as the juridical recognition of an institutional deficiency inherent to the entity itself - embedded within its organisational and managerial choices.

Accordingly, the offence attributable to the legal person does not merely constitute a projection of the crime committed by the natural person, but takes the form of structural liability, grounded in the failure to adopt - or the ineffective implementation of - organisational models suitable for preventing the commission of criminal offences. This is the so-called organisational fault, the defining feature of a compliance-based system that prioritises prevention as a means of anticipatory penal protection.

From this perspective, the core of the Decree lies in the entity's ability to establish and maintain a structural and procedural apparatus capable of preventing criminal conduct by its representatives. Indeed, it is not sufficient that the offence be committed by a senior manager or subordinate acting in the interest or to the benefit of the entity [14] or by a subordinate acting in the interest or to the benefit of the entity – the so-called *objective criterion*. It is also necessary, however, to ascertain whether the organisational structure in place was adequately designed to detect or prevent the unlawful conduct – this being the *subjective criterion*.

[14] More precisely, this refers to individuals vested with powers of representation or management of the entity, or those in charge of organisational units enjoying functional and financial autonomy. E. N. Mazzacuva, *Diritto penale dell'economia*, Milano, 2023, p. 47.

To this end, the adoption and effective implementation of an organisational, management, and control model may operate as a ground for exemption [15] from liability for the entity, but only where it is demonstrably embedded within the corporate structure, continuously updated and monitored by an autonomous and independent supervisory body [16].

This regulatory framework, is currently under considerable strain due to the disruptive integration of artificial intelligence systems into corporate decision-making processes. Indeed, the use of algorithms - often characterised by opaque logic and unpredictable outcomes - threatens to destabilise traditional dynamics of accountability and oversight, introducing elements of technological delegation that elude conventional organisational safeguards. Consequently, there arises a compelling need to adapt the compliance models under Legislative Decree No. 231/2001, so as to explicitly govern the deployment of intelligent systems and to extend compliance obligations to encompass the risks associated with decision-making automation. In this respect, the adequacy of the model can no longer disregard the necessity of a thorough digital risk assessment, capable of identifying exposure areas arising from the integration of AI into corporate processes, nor can it overlook the imperative to update control measures designed to mitigate these emerging dimensions of criminal risk.

[15] E. AMATI, N. MAZZACUVA, *Diritto penale dell'economia*, p. 58.
[16] L. Parodi, *Illecito penale dell'ente e colpa di organizzazione. Una recente conferma della traiettoria garantista tracciata dalla giurisprudenza di legittimità*, in *Sistema Penale*, *2023*.

The Organisational, Management and Control Model

The organisational model, as envisaged by Legislative Decree no. 231/2001, does not constitute a mere procedural framework but rather a dynamic [17] safeguard aimed at preventing the commission of offences through the rational organisation of corporate activities. Its effectiveness, in terms of exoneration or mitigation of the entity's liability, is contingent upon the fulfilment of substantive requirements expressly laid down in Articles 6 and 7 of the Decree. Among these, relevance attaches to the mapping of risk areas, the adoption of protocols to regulate decision-making processes, the transparent management of financial resources, the establishment of reporting obligations towards the Supervisory Body, as well as the introduction of an internal sanctioning mechanism to repress violations.

Such a structure must also provide for periodic verification procedures to ensure its ongoing relevance and effectiveness, as well as instruments enabling its systemic adaptation to regulatory, organisational, or technological changes. It is precisely on this latter front that one of the most pressing interpretative and practical challenges arises today: the integration of artificial intelligence systems within corporate processes.

The structured incorporation of AI into business operations engenders the emergence of novel profiles of criminal risk, sometimes entirely unprecedented in nature. In this context, organisational models are called upon to evolve, incorporating specific safeguards aimed at regulating the responsible and compliant use of intelligent technologies. The ongoing legislative developments - notably the proposed Artificial Intelligence Act - mandate a profound reconsideration of the compliance architecture, both from a preventive and repressive standpoint, requiring the delegated legislator to redefine, substantively and procedurally, the criteria for imputing administrative liability of entities for offences committed through AI. Such redefinition must take into account the actual degree of control exercisable over the utilised systems.

[17] E. AMATI, N. MAZZACUVA, *Diritto penale dell'economia*, p. 60.

Consequently, this entails an extension of organisational liability to behaviours which, although materially carried out by non-attributable entities - such as autonomous software or predictive models - can nevertheless be ascribed, in terms of organisational fault, to deficiencies in corporate governance.

The digital transition and the advent of artificial intelligence must therefore be regarded as factors capable of reshaping corporate liability. As a consequence, there arises the need to structure and/or restructure organizational models in such a manner that they may be deemed suitable not only for managing traditional risks, but also for addressing emerging risks scenarios stemming from the adoption of autonomous and self-learning technologies, within a framework of anticipatory and adaptive compliance.

Risk assessment

The drafting of the organisational model demands a profound and nuanced understanding of the entity, encompassing its organic structure, functional framework, prerogatives of governing bodies, as well as the corporate purpose in its entirety. Only upon such comprehensive awareness can a precise and rigorous mapping of the risk areas for unlawful conduct be undertaken, which translates into a statistical-probabilistic assessment grounded in the systematic processing and analysis of pertinent data.

The ongoing digital transformation and evolution necessarily require entities to review and strengthen their risk assessment activities, so as to explicitly account for the critical issues and risks associated with the use of artificial intelligence.

The crux of the issue lies in the examination of both the advantages and challenges inherent in digital compliance, as well as the implications that the adoption of such technological instruments entails regarding the attribution of administrative liability of entities. Specifically, the introduction and deployment of artificial intelligence systems necessitate an expansion of the traditional evaluative framework, insofar as these systems may intensify the complexity of risk by introducing elements of opacity, decision-making autonomy, and the potential amplification of unlawful consequences.

Accordingly, the risk assessment process must necessarily encompass a digital dimension of analysis, aimed at identifying and weighing the specific risks arising from the use of such technologies, in order t ensure that the organisational model preserves its preventive efficacy within the new frontier of automation and algorithmic intelligence.

Codes of Conduct and the Regulation of Artificial Intelligence Technologies

The strengthening of the compliance measures pursuant to Legislative Decree No. 231/2001 also extends to the Code of Ethics under article 6, paragraph 3 of the Decree, which, as is well known, constitutes the value-based and behavioural expression of the entity's identity, serving as an internal standard of legality and a cultural safeguard of compliance. Indeed, in the era of digital transformation, it is called upon to explicitly and systematically incorporate principles of integrity, transparency, and responsibility applicable to the use of intelligent technologies.

From this perspective, the inclusion of specific provisions regarding the use of artificial intelligence within the Code of Ethics constitutes not only a preventive measure consistent with the decree's rationale but also a manifestation of the entity's ethical commitment to consciously managing the risks arising from automation.

Thus updated, the Code assumes a proactive role in fostering a corporate culture rooted in the respect for substantive legality, even within the new dimensions of digital conduct.

GEBBIABORTOLOTTO

PENALISTI ASSOCIATI

Corso Vittorio Emanuele II, 68
10121 – Torino
Telefono +39 011 4546389
segreteria@gbpenalisti.it