Association International de Droit Pénal - International Association of Penal Law

XXI International Congress of Penal Law

"Artificial Intelligence and Criminal Justice"

International Colloquium of Section I (Criminal Law- general part):

"Traditional Criminal Law Categories and AI:

Crisis or *Palingenesis*?"

Draft Resolution

Preamble

Considering that

- the advent of Artificial Intelligence (AI) technology and autonomous or artificial agents (AA) support and replace many human activities and represent a real benefit for the society;

- AI systems and artificial agents are becoming increasingly autonomous, and their behaviors may be unpredictable to those who design, program, produce, distribute and use them;

Observing that

- the areas of application of AI technologies is considerably wide, and illicit facts related to their implementation might harm different interests, legal goods and fundamental rights;

- AI systems may also play an increasingly role in the perpetration of criminal acts as "instrument" to commit several existing criminal offences, and it might become the facilitator factor of the emergence of new criminal behaviors;

Paying particular attention to

- the increasing delegation of decisions to AI systems and AAs in different areas of activities, cases where their autonomous functioning causes harms will likely to be more frequent, raising the question of which person can be held liable for them;

- the autonomy of AI systems, that also created a debate within academia related to the possibility to consider them as the "subject" of a crime;

Bearing in mind

- the importance of appropriate reactions that Criminal Law is required to provide in preventing and punishing offences committed by, through or against AI systems and artificial autonomous agents;

- the fundamentals principles that must be guaranteed in establishing and applying criminal sanctions (including also punitive sanctions in a broader sense, which could be applied to legal persons), such as the principle of legality and the principle of culpability, which is a ~~as~~ necessary expression of the personality of penal responsibility;

Aware that traditional Criminal Law categories and criminal responsibility models need therefore to be considered and, if necessary, adapted to the emerging protection requirements;

Taking into account

- the "Ethics Guidelines for trustworthy AI" presented to the European Commission on 8 April 2019 by the High-Level Expert Group, and other significant recommendations of other international bodies (the "Feasibility study on a future council of Europe instrument on Artificial Intelligence and Criminal Law" of the European committee on crime problems of the Council of Europe, 4 September 2020)

- the recommendations of the XIV International Congress (Vienna, 1989) on the legal and practical problems posed by the difference between criminal law and administrative penal law, those of the XVIII International Congress (Istanbul, 2009), about the incrimination of preparation and participation in a crime, and those of the XIX International Congress (Rio de Janeiro, 2014), on Information Society and Penal Law,

The participants of the International Colloquium of the Section I of the XXI International Congress of Penal Law (Criminal Law: general part): "Traditional Criminal Law Categories and AI: Crisis or Palingenesis?" have adopted what follows

Recommendations

I. On the notion of Artificial Intelligence and the attribution of legal personality to AI systems and autonomous agents

1. At present, it is neither possible nor useful to identify a unitary and legally binding definition of Artificial Intelligence, which rather expresses a metaphor for multiple algorithmic and robotic systems that use Machine Learning (ML) techniques and that are in continuous technical evolution, becoming increasingly sophisticated and self-adaptive.

2. It is preferred to based criminalization and regulation interventions on the specific features of the different 'AI systems' and Artificial Agents, which for Criminal Law purposes must be considered and legally defined within the specific sectors where criminal sanctions are to be established and/or applied, also by referring to the respective definitions and disciplines contained in extra-criminal sources, if sufficiently detailed (e.g. self-driving cars, different robotic systems in medicine, artificial agents for stock exchange trading or logistics management).

3. There is no legal basis or utility, for Criminal Law purposes, in recognizing legal personhood to the Artificial Intelligence systems and Autonomous Agents known today.

4. On the one hand, there is an ontological distinction from human agents. AI systems lack the moral freedom to choose and evaluate the possible solutions to a problem or dilemma, considering also the context of social and ethical relationships and opportunities, with the necessary flexibility and capacity to adapt to even contingent or supervening situations and conditions.

5. On the other hand, punitive sanctions to such technological systems and agents would not respond to the purposes and functions of criminal punishment, because the effect of the threat of the penalty and its application would be emptied by the absence of self-awareness of their own existence in the past, present and future, and, above all, by the absence of voluntary self-determination, so that even excluding the retributive function, not even those of special and general prevention would be feasible.

II. On the Need for the Criminal Protection of Legal Goods harmed by or through Artificial Intelligence Systems and Autonomous Agents

6. It is necessary to recognize the essential importance of a reasonable and proportionate intervention of Criminal Law in preventing and punishing harms to interests, legal goods and fundamental rights that AI systems and Artificial Agents might cause, given that the same offences, if realized by natural and legal persons, according to the traditional categories of Criminal Law, might constitute a criminal offence crime. Therefore, they cannot go unpunished simply because they are carried out by or through the aforementioned systems.

7. It is necessary to identify and define specific models for attributing criminal liability to the persons (both natural and legal persons) who are 'behind' the AI systems, starting with the owners and those who decide on their concrete use, based on their interest and their benefit, and who must therefore be held legally liable, also from a “punitive” – not only a criminal Law - perspective.

8. The responsibility of the subjects described in the previous point does not exclude but must be compared and possibly added to that of other subjects (either natural or legal persons) who contribute to the causal chain of the harm: from the end user to the seller, distributor, producer, programmer, designer of the systems themselves.

9. In particular, a distinction must be made whether they are:

a. AI systems used in illicit activities: in this area, there will mainly be malicious conducts, which pose fewer problems in terms of imputation of criminal liability, given that AI systems are conceptually no different from other instruments and means of committing a crime.

Two questions, however, require to be answered:

a.1. in case of deviant behavior of the system from the intended illicit activity, the traditional principles of *aberratio ictus* and *aberratio delicti* must be applied, i.e. the mere material diversity of the harmed object must not represent an excuse if its characteristics are not relevant for the configuration of the criminal offence (e.g. killing one person instead of another is not relevant for the realization of the crime of murder, when it is intended by the agent). Instead it should be preferred to base criminal liability for a crime other than the one intended based on the possibility of concretely foreseeing such a different development of the action put in place by the AI system, by applying the principles of negligence based liability (as set out in paragraph III below);

a.2. since AI systems can be used for particularly harmful or dangerous conducts, given that they can amplify and aggravate the harm caused (as happen with ICT), since the consequences can be very distant from the actions that gave origin to them, making it more difficult to intervene *post factum* to prevent or at least to stop or reduce the harmful consequences, it should be considered the incrimination, as autonomous preparatory offences, of activities of programming, production, distribution, sale of algorithms, software, 'malicious' AI systems. This criminal policy should be limited to high-risk AI systems (which can cause harm to life, body, or liberty of other human beings) and only in case of clear and present danger (on the conditions required in incriminating preparatory acts, see the resolution of the Section I of the XVIII AIDP Congress in Istanbul, 2009).

b. AI systems used in lawful activities: This case rises the most delicate issues in reference to the area of 'permitted risk', which should be delimitated through the hoped-for regulation of specific security obligations and precautionary rules to be applied to the activities of planning, development, production, distribution, sale, as well as use, of AI systems. The adjustment of the models of criminal liability in this area must address the friction that can be created between forms of responsibility for negligent behavior and the technical features of AI systems, namely: (1) their autonomy; (2) the concrete unpredictability of their decisions and behaviors; (3) the opacity of their regulatory mechanisms; (4) the complexity of their programming, development, production, updating and maintenance process.

III. On the adaptation of the models of imputation of liability to the features of Artificial Intelligence systems, specifically to their degree of autonomy

10. First of all, a distinction must be made, also according to the already recognized graduated automation and autonomy of AI applications in several areas, between the different levels of decision-making and operational autonomy of AI systems, which go from those where the 'automatic' behavior allows the human agent to have significant

control over the system, to those that are truly 'autonomous', where human intervention can only be distant, in time and in space, from the functioning of the AI system, which “decides” based on the information collected and on algorithms that adapt to its experience, so that there is a structural margin of unpredictability of the concrete outcomes.

11. In relation to the different types of AI systems, the definition of specific rules and standards of behavior, as foreshadowed in the proposal for a European regulation on Artificial Intelligence, is of fundamental importance.

12. The most pressing need for adaptation of the traditional categories of Criminal Law concerns the area of AI systems with a greater degree of autonomy, which are also the result to which current technological development and experimentation in many fields is tending, so that they will undoubtedly be even more important in the near future.

13. Under this perspective, the field of corporate criminal liability might provide a useful reference (even though legal persons are entities formed by human subjects), as well as the fields of product liability criminal law and criminal liability for the protection of health and safety in the workplace.

14. In these legally regulated fields, often harmonized at a European level, existing principles might be extended, with the necessary adaptations, to AI-related crime regulation, since they are based on the preventive assessment of the risks inherent in the specific activities performed, which have margins of permitted risk and correlated obligations of risk prevention and containment, according to models of organization and management with specific regard to the sources of danger of commission of criminal offences. Duties to act, especially in the face of adverse signals or events, are imposed to the relevant categories of persons (human beings), operating according to their respective competences, that assume the position of guarantors, in order to promptly adapt the regulatory and security measures of their activity, to the point of stopping it, if necessary.

15. From these recognized principles, the following recommendations can be elaborated to structure criminal liability for AI-related harms:

i. **Criminal liability of natural persons.** It must be based on the identification of personal positions of guarantee, in relation to the competences and functions performed in using AI systems, normally headed by top management in complex organizations. In each case, the formalization of positive obligations, of a technical, organizational and control nature, shall be addressed.

Criminal liability for negligent behaviors must comply with the general principles of Criminal Law, namely, the principle of personal culpability, since the objective connection between the causal contribution of the human agent and the commission of the offence by the AI system does not suffice, given that the foreseeability and evitability of the illicit fact are also necessary. However, criminal responsibility for negligence, for not having acted differently from what would have been possible, must be correlated not so much to the specific and concrete event or fact that occurred, as to the scheme of 'organizational fault', referring to the way the artificial agent is

structured and operates. The assessment of the risks arising from the AI system's activities must also include the awareness of outcomes that are concretely "unforeseeable" in individual cases, which is the basis of the obligation to prepare adequate and always up-to-date surveillance and containment measures, for which the natural person in charge remains responsible, being accountable (accountability), as the owner or top representative of the organization that uses the AI system in its own interest or to its own advantage.

ii. **Criminal liability of legal persons.** Considering that a large part of AI systems is produced or used by legal persons, it is of necessary to hold them accountable for the offences committed by or through such systems.

In this respect, assuming that precise public standards of conduct and compliance are to be introduced (cf. supra § 11), criminal punishment of the legal person, proportionate to the offences committed by or through AI systems and to the degree of fault of the organization, might be related to a model of liability based on organizational guilt, which leads to imputing responsibility subjectively, as the object of culpable reprehensibility, to the legal person in case of offences caused by the lack, deficiency or inadequacy of organizational and prevention measures, to be implemented and updated on the basis of the assessment of the specific risks deriving from the activities entrusted to and, in any case, carried out by the AI systems, in their interest or to their advantage.

A new model of corporate autonomous liability, not based on the liability of the individual natural person, should be promoted, since the legal person can be held liable even if the natural person who realized the harm is not individually punishable due to particular conditions or circumstances or if he/she is not specifically identified. Indeed, it is enough to ascertain the commission of an objectively typical and unlawful act in the interest or to the advantage of the organization (under Italian law, according to Article 8 of Legislative Decree No. 231/2001, which, however, is not a case of autonomous corporate liability, since it is based on the realization of a crime committed by a natural person, even just at an abstract level).

IV. On the types of sanctions applicable to natural and legal persons "behind" Artificial Intelligence systems and autonomous agents

16. The criminal sanctions applicable to natural persons, including imprisonment, and to legal persons, possibly of an administrative nature, according to the various legal systems, but in any case of a punitive nature, including fines and suspension of the activity by which the offence was committed, should in principle correspond to those applied for the type of offence realized, in accordance with the principles of the single legal systems, namely the principles of proportionality and individualization of the sanctions. When it comes to legal persons, these sanctions, in addition to pecuniary measures, should also include the injunction to modify the corporation's compliance and internal control system, as well as the possibility of ordering a period of public monitoring of the corporation to ensure that it complies with the imposed standards.

17. The important role of non-pecuniary penalties, such as the sanction of disqualification from exercising specific activities and confiscation, should be emphasized. Specifically, confiscation allows direct action to be taken against the AI system with or by which the offence was committed, without the need to recognize it as a legal entity or as having criminal capacity (see para. I above).

V. Alternatives to criminal sanctioning: standards and obligations

18. Given the problematic and foreseeable difficulty of implementing an effective criminal liability system for natural persons and legal entities 'behind' AI systems for offences committed by or through them, before or, at the very least, in parallel with criminal law reforms, it would be necessary for European and national legislations to fully define the regulation of the several sectors in which AI systems are implemented (such as those mentioned above of self-driving cars, health and surgical robots, autonomous weapons, etc.). Technical standards, structural characteristics and operating conditions of AI systems and their components should be regulated.

19. Such regulation, which must operate from the planning, production, distribution and sales phases to the actual use of AI systems, should also provide for concrete requirements concerning adaptation in case of adverse events or warning signals, with injunctional procedures, such as those already provided in areas of complex risks (e.g., health and safety in the workplace and environment protection), the violation or non-compliance of which may be punished with penal and/or punitive sanctions.

20. Alongside or as an alternative of criminal sanctioning, alternative models of compliance might be regulated, based on administrative regulations and on the need of restoration and prevention, as well as of reparation, which are the basis of the aforementioned injunctional system, or of future restorative justice interventions.