COUNCIL OF EUROPE

1949 - 2024

CONSEIL DE L'EUROPE

Strasbourg, 13 November 2024 CDPC (2024)09

EUROPEAN COMMITTEE ON CRIME PROBLEMS (CDPC)

Discussion Paper on Criminal Liability Related to AI systems

- following up Framework Convention on AI & Human Rights, Democracy, the Rule of Law (CAI, CETS No. 225)

Note:

This draft has been prepared for the November 2024 meeting of the CDPC plenary

This paper has been prepared by independent experts. It does not represent the position of the CDPC Secretariat and the Council of Europe.

document prepared by

Prof. Dr. iur. Sabine GLESS, Professor of Criminal Law and Criminal Procedure Law, Faculty of Law, Basel University and Mr Alfonso PERALTA GUTIÉRREZ, Magistrate-Judge of the Court of First Instance and Criminal Investigation, Member of the CEPEJ AI Advisory Board

Sabine Gless / Alfonso Peralta
October 2024

I. Purpose and Scope

Through its Framework Convention on Artificial Intelligence and Human Rights, Democracy, and the Rule of Law (CAI, CETS No. 225), the Council of Europe (CoE) seeks to safeguard human rights, democracy, and the rule of law as AI systems present new opportunities and challenges. This Convention's transversal approach aims at a unified approach of CoE States on the level of principle, providing them with a methodology to assess risk and impact of AI, leaving space for diversity on the legislation level. In the field of criminal justice, CoE legal instruments traditionally promote cooperation, which generally requires a certain level of harmonization for efficient interoperability.

The European Committee on Crime Problems (CDPC), tasked by the Committee of Ministers with overseeing and coordinating CoE activities in crime prevention and control, has been called upon to guide member States on AI-related implications within their field. Specifically, the CDPC has been tasked with drafting a legal instrument on criminal liability related to the use of AI, expected by the end of 2025. With its past activities—particularly the "Feasibility Study on a Future Council of Europe Instrument on Artificial Intelligence and Criminal Law (2020)" and preparatory work by its working group on AI and Criminal Law—the CDPC is well-prepared to fulfill this role.

This discussion paper centers on criminal liability and AI, aiming to help member States navigate their obligations under CAI (CETS No. 225), specifically to:

- implement necessary measures in domestic legislation to uphold the principles, rules, and rights in the CAI (Art. 1);
- protect human rights effectively in relation to AI use (Art. 4);
- adopt or maintain measures ensuring accountability and responsibility for adverse impacts on human rights, democracy, and the rule of law arising from activities within the AI lifecycle (Art. 9);
- adopt or maintain measures ensuring that the privacy rights of individuals and the protection of personal data are respected in activities within the AI lifecycle (Art. 11).
- This position paper also considers relevant obligations under the Convention on Cybercrime (Budapest Convention, ETS No. 185) and its Protocols, along with other CoE conventions on criminal law cooperation where applicable.

The objective is to provide a framework for developing national legislation on criminal liability issues arising within the AI lifecycle and to encourage member States to adapt their criminal laws to address situations where AI usage necessitates it, grounded in shared normative principles. The ultimate aim is

to establish a common minimum dominator that enhances coherence in criminal liability standards and thereby facilitates mutual assistance in criminal matters.

II. Terminology

As a preliminary step in the CDPC's task of drafting an instrument on criminal liability and AI following the adoption of the CAI, this discussion paper, should align with the terminology used in the CAI.

1. Definition of AI system

The legal definition of an **AI system** can be borrowed from Art. 2 CAI:

"artificial intelligence system" means a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations or decisions that may influence physical or virtual environments. Different artificial intelligence systems vary in their levels of autonomy and adaptiveness after deployment.

2. Further definitions

If CoE committees establish further terminology in their work on CAI, such as terms for the providers or users of AI systems, it would be advisable to adopt these terms to enhance conceptual clarity. In this regard, it is also advisable to align the central term of criminal liability with the obligations outlined in CAI (e.g., Articles 9 and 11), and to use the phrase "criminal activities within the lifecycle of AI systems." This terminology is consistent with CAI language and encompasses various forms of punishable conduct, including:

- negligence in designing, training, or using AI systems;
- using AI systems for cyberattacks;
- producing deepfakes and using them for defamation or online sexual grooming;
- failure to update AI systems when flaws and associated risks to critical infrastructure are recognised, among others.

Consideration could be given to the necessity of further concepts used in CoE States' criminal law, such as:

- **"Socially acceptable risk"** as a tolerable risk that permits certain dangerous acts to be carried out with impunity under narrowly defined conditions. This shall include situations where the risks were unknown or totally unforeseeable according to the generally acknowledged state of the art. This also implies that a perpetrator may not be found guilty of a criminal act. For example, the deployment of chatbots that are beneficial for disseminating information, even if, in 1 out of 10,000 uses, the chatbot hallucinates and provides incorrect—possibly defamatory—information.
- **"Novus actus inrerveniens"** meaning that an act or event breaks the causal connection between an act by the defendant and subsequent happenings and therefore relieves the defendant from responsibility for these happenings.
- **"Relevant contributory liability"** which – even though being a tort concept – could be useful to limit criminal liability within the lifecycle of AI systems, meaning that individuals harmed partly due to their own negligence "contributed" to the criminal act. This implies that a perpetrator may not be found (fully) guilty of a criminal act. For instance, this could apply if the driver of a highly automated car does not cooperate diligently in updating the relevant AI systems or does not comply with the maintenance guidance and recommendations.

III. Structured approach and principles

The CDPC is entrusted with handling one of the States' key areas of responsibility:crime prevention and crime control. This also means preventing violations of protected legal interests by detecting and investigating criminal offences, and prosecuting and punishing their perpetrators. The following structured approach and principles could help navigate member States' obligations under CAI:

1. AI systems used as a tool to commit a crime

In light of CAI's principles of accountability, transparency, and proportionality, as well as CDPC's mandate to draft a legal instrument on criminal liability related to AI, existing criminal laws can be evaluated for their capacity to adequately address the potential of AI systems to harm individuals' lives, physical safety, and other well-being, as well as community assets. When AI systems are used merely as tools to intentionally commit established crimes (such as homicide, murder, or theft), existing provisions in domestic criminal justice systems may sufficiently address the criminal conduct. However, if member States wish to modify their criminal laws—such as by considering the use of AI systems as an aggravating circumstance or by expanding liability—they may choose to do so.

2. AI systems used to cause harm in novel ways

When AI systems are used to cause harm in novel ways—either through actions not yet punishable under member States' domestic systems or by leveraging technology to expand the scope and impact of punishable conduct both in scope and impact, specific concerns arise. These include, but are not limited to:

- **Technology-facilitated violence against women and girls (TFVaWG)**: This includes using deepfakes of a person's video, voice, or image for purposes such as sexual exploitation, nudity, advertising, or other commercial uses, with the intent to undermine their moral integrity. Such actions involve the dissemination, display, or transfer of their body image or voice generated, altered, or recreated using AI systems.
- **Online sexual grooming using AI systems**: This involves AI-enabled manipulation or deception to initiate contact with a person under sixteen with the intent to persuade or coerce the individual into producing pornographic material in which a minor is depicted or appears;
- **Distribution of "Dark AI"**: This refers to AI systems specifically engineered for malicious purposes, such as hacking, cracking, or other cyberattacks, as well as AI designed to target critical infrastructure,create situations of serious public security risks or with the intention of facilitating the commission of any crime.

3. Bona fide use of AI resulting in harm

When the bona fide use of AI systems leads to harm due to the inherent risk of unforeseeable actions, such as in cases where:

- a chatbot slanders an individual due to an AI hallucination,

- an accident is caused by a self-driving car failing to identify a sled accidentally crossing a highway, or
- high-frequency trading systems inadvertently engage in market manipulation,

member States may wish to discuss a harmonised approach in criminal law to establish a threshold for defining criminal negligence or, alternatively, a narrowly defined exemption from liability without prejudice to possible administrative or civil liabilities.

4. AI systems designed, trained, or deployed in violation of CAI obligations

When AI systems are designed, trained, or deployed in violation of CAI obligations, member States may wish to consider a harmonised approach in criminal law to penalise non-compliance with these obligations, particularly if the non-compliance with requirements lead to incorrect, incomplete, hidden or misleading information that might affect human rights, democracy and the rule of law. Some situations may be those related to:

- accountability, transparency, and non-discrimination in the use of training data; and
- respect for copyright and privacy rights in obtaining training data and training AI systems
-

Additionally, the putting into service, produces, acquires for their use, imports or, in any way, provides to third parties or the use of an AI system that is prohibited by its national or European legislation could be considered as an offence.

5. Principle of legality

Clear legislation on criminal activities within the lifecycle of AI systems should, on one hand, respect Article 7 of the ECHR, which states that no one shall be held guilty of a criminal offense for any act or omission that did not constitute a criminal offense under national or international law at the time it was committed, and, on the other hand, protect the fundamental rights, values, and freedoms enshrined in the European Convention on Human Rights.

6. Principle of proportionality

To address the specific challenges of criminalising activities within the AI system lifecycle (see Section V below), member States may wish to discuss whether, or rather how, the principle of proportionality should guide a cautious approach that:

- focuses on the most serious crimes,
- prioritises high-risk AI systems and applications, or
- employs another threshold to limit criminal prosecution of activities within the AI system lifecycle
- defines certain exemptions from criminal law (see also III.3, as well as V.2 and 4).

IV. Legal approaches/frameworks for establishing criminal liability

The working group's feasibility study (see *supra* I.) as well as groundwork of comparative criminal law illustrates that member States' criminal justice systems allow for different legal approaches and frameworks to establish criminal liability in the different situations where activities within the lifecycle of AI systems ought to be punished (see *supra* III., 1.-4.)

First, member States must make a general decision regarding the type of framework they desire for developing national legislation to address punishable activities within the lifecycle of AI systems. Subsequently, they can determine which legal approach would best achieve the ultimate objective of establishing a framework for adequate national legislation.

V. Challenges

To adequately address punishable conduct within the lifecycle of AI systems, various challenges must be met. The list below does not exclude additional issues but focuses on questions pertinent to criminal liability arising in this area.

1. AI Systems as Possible Agents in Criminal Law

AI systems have great potential because they act autonomously, processing large pools of data 24/7. However, for various reasons, they can malfunction and cause serious harm. As it stands today, AI systems are not suitable recipients of criminal punishment. From a criminal law perspective, they are not responsible actors because they lack the capacity to perceive themselves as morally responsible agents and cannot comprehend the concept of retributive punishment. In other words, a chatbot or self-driving car cannot morally assess its actions (and, thus is no moral agent), nor can it be imprisoned. Criminal law in member States is generally tailored to the conduct and intentions of humans, whether natural persons or individuals acting on behalf of entities (corporate liability). At present, it does not seem necessary to create new forms of legal personality for AI systems.

2. Bona Fide Use of AI Systems That Cause Harm

AI systems have significant potential to benefit the individual as well as society as a whole.

a) The Foreseeable Unforeseen

However, they can malfunction and cause serious, often unforeseeable harm. Notable examples from the everyday life include:

- chatbots or search engines slandering individuals;
- self-driving cars causing accidents and resulting in fatalities;
- vacuum robots injuring humans.

This raises the question: under what conditions should individuals or corporations be held criminally responsible for producing, programming, or using intelligent machines that cause harm, even if all relevant actions in the lifecycle have adhered to state-of-the-art standards? For instance, should the:

- deployer of a search engine be liable for slander?
- engineers and software experts who developed a self-driving car be charged with manslaughter?
- producers of vacuum cleaners be held accountable for bodily harm?

These questions persist even if the “legal agents behind an AI system” complied with the prevailing standards in their respective fields. AI systems, particularly those that learn in a "data-driven" manner (as opposed to "model-driven" or "rule-based") and function using complex methods (such as neural networks or generative AI), cannot have their actions fully predicted.

b) The Negligence Dilemma

This creates a dilemma for traditional negligence concepts: individuals who deploy an AI system for interaction with humans can foresee that the AI system might malfunction and cause harm, but they cannot predict how this might occur. The fact that individuals involved in the lifecycle of an AI system cannot eliminate the possibility that such systems may cause harm leads to two mutually exclusive conclusions regarding their liability for negligence: One could argue that they cannot be held responsible because AI systems act “on their own.” A robot that would adopt behaviours not initially foreseen by its programmers, and not only that, but that would even adopt a fully autonomous behaviour of free will, disobeying even those orders that had been incorporated into it, would have to be shut down or deactivated and withdrawn from the market. Alternatively, it could be claimed that one should foresee any and all potential harm caused and, thus, face de facto strict liability. The first argument is unconvincing: the inherent unpredictability of AI systems cannot absolve individuals engaged in the lifecycle of an AI system from liability because it is this unpredictability that creates a duty of care. Similarly, if a zoo manager releases a tiger out of the cage and the tiger subsequently kills people on the street, the zoo director cannot successfully argue that tigers are wild animals and therefore uncontrollable. However, imposing strict criminal liability on anyone involved in the lifecycle of AI systems would be disproportionate, as individuals and society benefit from AI, and also the CAI recognises the opportunities presented by this new technology, yet seeks to mitigate associated risks.

c) Balancing Interests

Member States may wish to carefully balance the conflicting interests at play. On one hand, there is a compelling interest in obtaining redress for harm suffered due to AI system malfunctions. On the other hand, there is a legitimate interest in utiliszing AI systems. Given the significant social benefits associated with the use of AI systems, member States might consider limiting criminal liability to situations where actions in the lifecycle of an AI system do not adhere to state-of-the-art standards or lack reasonable measures to control associated risks.

3. Need for specific legislation

Regarding the various legislative response options, several issues arise, particularly whether member States consider it necessary to adopt specific legislation in light of the new obligations arising from the CAI.

a) Need for legislation due to bona fide use of AI Systems

One question is whether member States want to adopt specific measures to address the "negligence dilemma" resulting from the autonomous actions of AI systems (as explained in section V.2), in particular considering the principle of proportionality. This could involve measures that ensure that all parties with duties in the lifecycle of an AI system can be held adequately liable, such as:

- providing for enhanced criminal liability where features of AI systems, particularly automation or contributory operations, foster the commission of a criminal offense, or
- granting a narrowly defined impunity in cases of relevant contributory liability or socially accepted risks, or
- other measures seen as appropriate.

b) Need for Legislation due to tech-facilitation of high-scale crime ("Dark AI systems")

A similar question arises regarding the capacity of AI systems to act autonomously 24/7, creating the potential to replace entire criminal organisations with a single software architecture that continuously deepfakes or hacks on demand (Crime-as-a-Service as a business model). This increases the risk for individuals who may fall victim to fraud, privacy violations, unauthorized use of images, infringements on intellectual or industrial property, defamation, manipulation of the electoral process, or misleading courts in criminal trials.

The question is whether such tech-facilitation requires a specific response from criminal law that targets criminals providing or acquiring these services, or whether "Dark AI" can be addressed using traditional legal tools, such as imposing harsher penalties during sentencing.

c) Need for Legislation with regard to the principle of legality

As a general issue, the question arises whether member States should adopt new legislation in order to uphold the principle of legality as enshrined in Article 7 of the ECHR and in their respective domestic laws.

d) Continue without specific legislation

Alternatively, member States could choose to continue without specific legislation and rely on their traditional body of law—statutes and case law built on rules and cases addressing the (negligent) use of machines by humans, often found in criminal product liability. However, in doing so, they must ensure that they effectively protect human rights related to the use of AI (Article 4) and have measures in place to ensure accountability and responsibility for adverse impacts on human rights, democracy, and the rule of law resulting from activities within the lifecycle of AI systems (Article 9). They must also ensure

that privacy rights of individuals and their personal data are protected in relation to activities within the lifecycle of AI systems (Article 11).

4. Specific exemptions from criminal liability, such as "Dual use technology"

As pointed out above, AI systems hold great potential that can benefit individuals and society as a whole but also can cause harm. In the light of the benefits of AI systems and the principle of proportionality, member States, before establishing new criminal liability, may want to consider specific exemptions from criminal liability, also avoiding the risk of widespread criminalization that could stifle AI development. Dual use technology could be a situation that merits such an exemption.

5. Corporate criminal liability

AI systems are typically produced by corporations, not by individuals. Thus, the question of corporate criminal liability arises. However, this issue is not specific to AI and criminal liability; the arguments raised in previous debates also apply here.

6. Practical challenges

Overall, it may be advisable to also consider the most significant practical challenges when working on the instrument. These challenges include, for instance, evidentiary issues. Novel challenges may for example arise, when AI systems autonomously generate evidence—such as when they monitor users (like Fitbits or highly automated cars) and store motion profiles or alerts. It can be very difficult for courts to assess the reliability of such evidence when presented in a criminal trial. The CAI's principles of transparency and accountability could assist to build a grid or framework for a meaningful vetting of such evidence.

Further practical challenges are related to enforcement shortcomings in cybercrime, with one hurdle being the anonymity of information and communication in cyberspace. Again, the CAI's principles of transparency and accountability can facilitate more efficient criminal investigation and prosecution, in particular as member States are in an excellent position for cooperation through the Council of Europe Convention on Cybercrime (Budapest Convention, ETS No. 185) and other instruments for efficient mutual assistance in criminal matters.

VI. Specific provisions

First, member States must decide what kind of framework they wish to establish for the development of national legislation addressing punishable activities within the lifecycle of AI systems (see *supra* III. 1.-4.). Subsequently, a meaningful discussion can take place regarding what specific provisions may be necessary for an adequate criminal law approach. Possible specific provisions could include:

- Adaptations of homicide or injury laws to address crimes caused by AI systems, such as accidents involving self-driving vehicles.

- New provisions to address violations of data protection or privacy rights committed through AI systems, including the use of spyware and hacking tools.
- New provisions addressing deepfake crimes executed by AI systems, such as the technology-facilitated production and dissemination of a person's image for sexual use without that person's consent.

VII. International Cooperation

The CAI compels member States to engage in international cooperation and encourages novel forms of collaboration (Art. 25). Regarding criminal liability connected to cyberspace, the Convention on Cybercrime (Budapest Convention, ETS No. 185) and its Protocols already provide a gold standard that is widely recognised around the world, as do other CoE conventions on Mutual Assistance in Criminal Matters and similar treaties between member States and other countries. The combined effect of these instruments offers opportunities to conduct successful criminal investigations and proceedings concerning criminal offenses occurring during the lifecycle of AI systems.

A core issue for cooperation between states is dual criminality, as referenced in the Committee of Ministers' Recommendations No. R (85)10 regarding the practical application of the European Convention on Mutual Assistance in Criminal Matters concerning letters rogatory for the interception of telecommunications, No. R (88)2 on piracy in the field of copyright and neighboring rights, No. R (87)15 regulating the use of personal data in the police sector, No. R (95)4 on the protection of personal data in the area of telecommunications, particularly telephone services, and No. R (89)9 on computer-related crime, which provides guidelines for national legislatures regarding the definition of certain computer crimes, as well as No. R (95)13 concerning issues of criminal procedural law related to information technology.

> Consideration could be given to the necessity of a list of offenses for which dual criminality checks can be abolished. This would help to guarantee cooperation and could foster mutual trust

VIII. Further developments beyond the ‘criminal liability’ approach.

The CAI, with its principles of accountability, transparency, and proportionality sets a standard for criminal justice system far beyond criminal liability in the lifecycle of an AI system,

For instance, CAI can determine how profiling can be used, which might shape the tools of predictive policing. n area that is inevitably connected to criminal liability and AI, but merits its own examination, is legal automation, specifically the various efforts to automate legal processes made possible by advancements in natural language processing (NLP). If a member State were to consider automating the processing of minor criminal charges, the overall system of cooperation in criminal matters would require a new assessment. It is a fundamental question in the era of vastly and rapidly developing AI

systems whether only humans should judge humans. Or if it is conceivable to have a criminal trial in which a software architecture autonomously applies the law to a specific case of a very minor nature. Today, we still view human judges as guarantors of the separation of powers and the democratic legitimacy of jurisprudence, and we are reluctant to risk undetectable errors in judgment that may arise from using AI systems that are prone to hallucination. However, in the near or distant future, procedural safeguards such as the right to a fair hearing and the opportunity for meaningful dialogue could be adapted for use with AI systems.