COUNCIL OF EUROPE

CONSEIL DE L'EUROPE

Strasbourg, 27.10.2022 PC-CP (2021) 17 rev 7

COUNCIL FOR PENOLOGICAL CO-OPERATION
(PC-CP)

DRAFT COMMITTEE OF MINISTERS RECOMMENDATION CM/REC(2023)XX

ETHICAL AND ORGANISATIONAL ASPECTS OF THE USE OF ARTIFICIAL INTELLIGENCE AND RELATED DIGITAL TECHNOLOGIES BY PRISON AND PROBATION SERVICES[1]

Document prepared by

Håkan KLARIN
CIO IT-Director, Prison and Probation Services, Sweden

Pia PUOLAKKA
Project Manager, Smart Prison Project, Forensic Psychologist, Criminal Sanctions Agency, Finland

Fernando MIRÓ LLINARES
Professor of Criminology and Criminal Law, University Miguel Hernández of Elche, Spain

(Scientific Experts)

[1] The text in blue and italic should go to the commentary

COUNCIL OF EUROPE

COMMITTEE OF MINISTERS

Recommendation Rec(2023)XX
of the Committee of Ministers to member States regarding the Ethical and Organisational Aspects of the Use of Artificial Intelligence and Related Digital Technologies by Prison and Probation Services

(Adopted by the Committee of Ministers on XX 2023 at the XXX meeting of the Ministers' Deputies)

The Committee of Ministers, under the terms of Article 15.*b* of the Statute of the Council of Europe,

Having regard to the European Convention on Human Rights and the case law of the European Court of Human Rights;

Having regard to the Convention for the protection of individuals with regard to automatic processing of personal data ("Convention 108+");

Having regard also to the European Convention for the Prevention of Torture and Inhuman and Degrading Treatment or Punishment and to the work carried out by the European Committee for the Prevention of Torture and Inhuman or Degrading Treatment or Punishment and in particular the standards it has developed in its general reports;

Taking into account the specific conditions under which prison and probation services operate and their role in the execution of penal sanctions and measures by these services which is among the strongest manifestations of public powers imposed on individuals and may interfere deeply with their human dignity, human rights and privacy, including the collection and processing of personal data;

Recognising in this respect that the rapid development and use of digital technologies, as well as of artificial intelligence (AI) in all spheres of social life can bring a number of positive changes in our societies but also raise a number of ethical concerns regarding human rights, respect for private life and data protection;

Noting that the collection of biometric data and the use of algorithms by the criminal justice system are advancing at a great pace in Europe and are gaining more and more place in particular at all its stages and in all its areas;

Noting also that digital and AI literacy needs to be enhanced among key actors in the criminal justice system and urgent measures need to be taken to prepare them to make efficient and ethical use of AI and related digital technologies in their everyday work to the benefit of other service users;

Drawing attention that these tools need to be commissioned for design, development and maintenance to carefully selected and vetted private companies, working in close co-operation with the prison and probation services. These companies should be made aware that high ethical norms

and principles and strict professional rules should be respected, and that the main driver should be rehabilitating offenders and not making profits;

Underlying therefore that it is indispensable to develop rapidly and to regularly review and if necessary, revise principles and norms which should guide the prison and probation services of its member States when using AI and related digital technologies in order to preserve high ethical and professional standards;

Further stressing that AI and related digital technologies should be used not only for safety and security purposes but also for social inclusion of persons in conflict with the law and that their reintegration should remain central. This use should not undermine the human-centred approach and should avoid discrimination and economic and social inequalities;

Endorsing the standards contained in the relevant recommendations of the Committee of Ministers of the Council of Europe and in particular: Rec(2006)2-rev of the Committee of Ministers to member States on the European Prison Rules; CM/Rec (2008) 11 on the European Rules for juvenile offenders subject to sanctions or measures; CM/Rec(2010)1 on the Council of Europe Probation Rules; Rec (2012)5 on the European Code of Ethics for Prison Staff; CM/Rec(2014)3 concerning dangerous offenders, Rec(2014)4 on electronic monitoring and Rec(2017)3 on the European Rules on community sanctions and measures;

Drawing also attention to the EU General Data Protection Regulation (2016/679 and the EU JI-Directive (2016/680) as well as to the OECD Recommendation of the Council on Artificial Intelligence (OECD/LEGAL/0449);

Recommends that governments of member States:

- be guided in their legislation, criminal policy and practice by the rules contained in the appendix to this recommendation;

- ensure that this recommendation and its explanatory report are translated and disseminated as widely as possible and more specifically among judicial authorities, prosecution, police, prison, probation and juvenile justice services, as well as among private companies which design and provide AI and related digital technologies in the framework of the criminal justice system.

Appendix to Recommendation CM/Rec (2023) XX

This Recommendation seeks to provide guidance related to the ethical and organisational aspects of the use of artificial intelligence (AI) and related digital technologies in prisons and by the probation services. The private companies that design, develop, provide, deliver, operate and maintain such technologies to be used by the above services should also follow the ethical and organisational principles and standards contained in the Appendix to this Recommendation.

The juvenile justice services should make use of these rules in a manner adapted to the specific needs of juveniles.

One of the key messages of this Recommendation is that AI and related digital technologies should be used legitimately and proportionately when and if they: (1) contribute to brining positive change in offenders; (2) assist prison and probation staff in their everyday work; (3)help advance the effectiveness of the criminal justice system and in particular the execution of penal sanctions and measures..

The rapid pace of advancement of AI and related digital technologies, as well as the pace and scale of their use by the European jurisdictions means that this Recommendation needs to be reviewed regularly and revised accordingly in order to endeavour to protect the human rights and fundamental freedoms of its users and safety and security of our societies.

Currently AI, algorithmic tools and related digital technologies are already used in some prison and probation services across the world, but according to a recent review most European jurisdictions still do not use these and almost none has any policies or legislation regarding their use by the prison and probation services. Because it is so little used, there's also little research about the results, benefits and risks of their use. (Puolakka & Van De Steene, 2021). The drivers of the use of AI by the penitentiary agencies lie in other agencies of the society where experiences, best practices and ethical principles have been developed more than in corrections so far. Prison and probation services are part of an already digitalized society, so they should explore how to make efficient use of AI and related digital technologies in conformity with the existing national and international human rights standards. Such use should strengthen and not weaken the key role of the human factor.

This Recommendation recognises that the rapid pace of scientific and technological change requires reviewing it regularly and if needed, revising it as this may enable a desirable change of direction. It would also allow for as yet unforeseen shifts in opinion about AI's costs, capabilities and its social impact to be taken account of.

I. Definitions

For the purposes of this Recommendation the following definitions are used:

Artificial Intelligence (AI): Artificial intelligence refers to systems which enable computers, robots or other machines to analyse their environment or big data and to take action, with some degree of autonomy in order to achieve specific goals. AI mimics the perception, learning, problem-solving, and decision-making capabilities of the human mind.

AI is able to simulate human intelligence processes based on the data given to it. Current systems are still on the level of so-called Artificial Narrow Intelligence (ANI), which means their usability is

limited to specific tasks or limited processes compared to the versatility of human intelligence. Artificial General Intelligence (AGI), which would be able to undertake a range of different cognitive and practical tasks, and in that sense mirrors the capabilities of a human person more closely, is in development. Beyond that is the prospect of Artificial Super Intelligence (ASI), purely theoretical for now, beyond our remit, but considered feasible sometime this century (Yampolskiy, 2016). AI is and can be better than humans in specific tasks, but it's up to humans to decide which are these tasks, where AI is most suitable to use and what ethical principles are to be followed to ensure its fair, secure and human-directed use.

AI and algorithmic-based decision making and machine learning (an automated AI statistical and data analytics technique which by using patterns in available data and algorithms and by gradually improving its accuracy in imitating the way humans learn) teaches computers to learn from experience.

The most popular and widespread AI technique to this day is known as machine learning. It can identify patterns in the data and then apply this knowledge to new data, so the AI can learn by itself from the data. The knowledge of the system is in the form of algorithms: a set of rules that describes the relations of different items of the data. AI's computational power enables it to execute certain tasks faster and analyse larger amounts of data more efficiently than humans.

The more developed learning technique is the deep Learning, which is a type of machine learning using artificial neural networks that has many layers and offers greater capabilities of performing complex tasks in which multiple layers of processing are used to extract progressively higher level features from data.

Big data: Constant collection, analysis and accumulation of large amounts of data, including personal data, from different sources and subject to automated processing based on computer algorithms and advanced data processing techniques, using both stored data and data transmitted in continuous flow, in order to generate correlations, trends and patterns.

The huge and constantly increasing amount of collected data and the mixing of data coming from different sources leads to the increased use of AI in order to process it. At the same time, it represents a challenge because it changes the understanding of privacy and of the impact of such information on human rights. Therefore, strict rules and constant supervision are needed to avoid as much as possible the misuse of big data.

Related digital technologies: This generic term refers to all electronic devices, automatic systems, and technological resources that generate, process or store information and data which are being used by AI.

Related digital technologies are for example facial recognition technologies, algorithmic risk assessment tools, wrist bands monitoring biometric data.

Text classification is also another example. It is also known as text tagging or text categorization is the process of categorizing text into organized groups.

AI translators are digital tools that can be used to translate the words and the meaning of not only words, but whole sentences.

Algorithm: A finite suite of formal rules/commands, usually in the form of a mathematical logic, that allows for a result to be obtained from input elements.

The Commissioner for Human Rights (2019:24) uses this definition of an algorithm. AI algorithms can be of two kinds, top down and bottom up. Top-down algorithms control a machine with a pre-determined programme, making its behaviour highly predictable. Bottom-up algorithms also called "stochastic algorithms" allow a machine to learn from past experiences and alter the algorithms with which it was originally programmed in the light of that. This is the so-called "machine learning". Bottom-up algorithms enable machines to function with some degree of autonomy from the humans who originally wrote their programmes, and do not require human intervention to improve their performance.

Biometrics recognition (facial, speech): Automated identification or verification of human identity through measurable physiological and behavioural traits. Major biometrics technologies include fingerprint and iris scanning, facial recognition, hand geometry, and voice recognition.

Some examples of the use of biometric recognition in the criminal justice system include video vigilance, suicide prevention, aggression prevention and prevention of illegal items smuggling (McGoogan, 2016; Houser, 2019).

Electronic monitoring: a general term referring to forms of surveillance with which to monitor the location, movement and specific behaviour of persons in the framework of the criminal justice process. The current forms of electronic monitoring are based on radio wave, biometric or satellite tracking technology. They usually comprise a device attached to a person and are monitored remotely.

This definition comes from Recommendation Rec(2014)4 on electronic monitoring.

Robots: Machines that can substitute for humans and/or can replicate human actions. They may function autonomously within a pre-defined frame of actions or may require user input to operate.

Robots start being used in prisons to assist in some everyday tasks like food distribution, for security, for distribution of medical drugs.

Virtual reality (VR): A computer-generated simulation of a three-dimensional image or environment that can be similar to or completely different from the real world and can be interacted with in a seemingly real or physical way by a person using special electronic equipment.

Chatbots and virtual assistants can be used for customer service in prisons and by probation services.

II. Basic Principles

1. When designing, developing and using AI and related digital technologies, respect for human rights and dignity of all persons impacted by this use should be ensured **(principle of respect for human dignity and fundamental rights).**

Prison and probation services have traditionally been human-centred organizations, although technology has been integral to the very character of imprisonment since its inception, in order to

create a secure institution and to protect society. Digital technologies could be seen as replacing the traditional forms of imprisonment like locks, bars and bolts. However, also in these organizations we are seeing development where machines are able to perform cognitive and practical tasks hitherto associated only with human capabilities. Much work in prison and probation services includes heightened ethical and security questions due to the special nature of these organizations and their persons.

2. All processes related to the design, development and maintenance of AI and related digital technologies to be used by the prison and probation services and the private companies acting on their behalf should be in conformity with the relevant national and international legal norms and should be defined by law **(principle of legality and legal certainty)**.

Legal frameworks and policies should be established regarding the use of AI and related digital technologies also in the prison and probation services. These rules and policies should follow the principles and recommendations stated on the international level by the Council of Europe. Adequate and effective guarantees against arbitrary and abusive practices due to the application of AI and related digital technologies in the public sector should be afforded by the national law.

The term used is "national law" rather than "national legislation", as it is recognised that law making may take different forms in the member States of the Council of Europe. The term "national law" is designed to include not only primary legislation passed by a national parliament but also other binding regulations and orders, as well as the law that is made by courts and tribunals, in as far as these forms of creating law are recognised by national legal systems.

In general, it can be said that trustworthy AI is (1) technically robust and reliable, (2) legally regulated and (3) ethically defensible. All AI applications must comply with the Convention for the protection of individuals with regard to automatic processing of personal data ("Convention 108").

The European Ethical Charter on the use of AI in judicial systems and their environment (European Commission, 2018) has defined the following five principles (1. principle of respect for human rights; 2. principle of non-discrimination; 3. principle of quality and security; 4. principle of transparency, impartiality and intellectual integrity; 5. principle "under user control") which are to be taken into account also in the prison and probation services.

Most of the AI applications and software are developed by private sector. Private sector organisations are not necessarily aware of the special circumstances of correctional space which should be taken into account when designing AI applications for these settings. There must be collaboration in this development with private sector and corrections.

Human-centred and safety clauses and considerations should be clearly defined and negotiated. There should be clearly set goals when signing contracts, defining why and for what purposes AI is needed and how it is to be used taking into consideration data protection and ethical principles. Maintenance and revision of AI tools is important to ensure these tools don't start to produce biased outputs. There's also a need to coordinate collection and use of data within prison between the different systems used and companies providing services to ensure collaboration between technical and field-specific experts.

3. When designing, developing or using AI and related digital technologies, discrimination and bias should be avoided. Proactive measures to prevent or to resolve the creation or intensification of

any inequality between individuals or groups of individuals should be taken. **(principle of equality and non-discrimination).**

Social prejudices and stereotypes can turn into algorithms if those designing, developing and using AI and related digital technologies do not understand how algorithms are formed and what kind of data they use, how and for what purpose. This is especially harmful with already vulnerable groups if algorithms start to repeat and validate the biases we have in human thinking and thus perpetuate these. Examples of such possible biases are racial, or gender biased algorithms or algorithms used for security or money laundering purposes or for labour selection or insurances.

AI could also deepen existing inequalities between individuals or groups of individuals and therefore proactive measures should be taken to avoid such a danger, like providing digital and AI literacy, digital tools and employment opportunities.

4. AI should be used when it is strictly necessary, only in a manner that implies the least negative impact on human rights of its users and if its intensity corresponds to the purpose and the expected results **(principle of necessity, proportionality and efficacy of AI)**.

Before implementing AI and related digital technologies, their use and impact should be discussed with the prison and probation management level in order to ensure that AI will be fit for the purpose and will support the strategical targets of the organization. Risk of causing harm, safety, security and offender management should be key indicators in decision taking.

5. The process of designing, developing and use of AI and related digital technologies should be transparent to public scrutiny and monitored on a regular basis **(principle of good governance, transparency, traceability and explicability)**.

Good governance requires society to be informed and involved as far as possible in the process of designing, developing and use of AI.

The information about design, operation and data processing methods should be non-opacite, accessible and understandable to service users, external public scrutiny should be ensured as this brings effective responsibility and accountability.

The establishment of public registers listing AI used in the public sector, containing essential information about the system such as, its purpose, actors involved in its development and deployment, basic information about the model, and performance metrics should be addressed in the context of a legally binding or non-legally binding instrument on AI in the public sector.

Prison and probation staff and persons should be informed about the coming of AI and the future shape it will have on them. They should be informed when and how AI assisted decision making or surveillance is involved in their case. In the offender management process, they should understand how particular AI assisted conclusions are made, and the recommendations of such systems should be shared with them.

Add on explicability.

6. When a decision is based on the use of AI and related technologies, effective appeal procedures should be put in place, ensuring that such a decision is reviewed by a human **(principle of the right to a human review of decisions)**.

A decision cannot not be taken based solely on the use of AI and related technologies. Any decision taken by using to a varying extent AI and related digital technologies can be appealed in order to ensure its revision by a human.

Final decisions based on AI and related digital technologies should always be taken by humans and the human-centred concerns should be of primordial importance.[2]

As a minimum there should be provisions on access to an effective remedy before a competent authority (including judicial authorities); a mandatory right to human review of decisions taken or informed by AI and related digital technologies; and an obligation for public authorities to implement adequate human review for processes which are informed or supported by AI and related digital technologies and to provide relevant individuals or legal persons with meaningful information concerning the role of AI in taking or informing decisions relating to them (except where competing legitimate overriding grounds exclude or limit such review or disclosure).

7. Reliable and accurate AI and related digital technologies should be obtained by using certified sources, tangible data and validated scientific methods and values. The design and use of AI and related digital technologies should be done in a secure and audited technological environment **(principle of quality, trustworthiness and security)**.

For the development of AI and related digital technologies an interdisciplinary team dedicated to maintenance, development and continuous improvement of AI-solutions should be established. This team should include both engineers, mathematicians and business developers as well as social researchers and scientists, data security and data protection experts who are familiar with the prison and probation systems and who ensure constant coordination with the prison and probation services in order to ensure the solutions meet the organizational targets, based on the expert knowledge professional ethics in all the relevant fields.

The following main steps should be taken for an initial review of an AI and related digital technologies, where relevant:

- *Risk Identification means what are the risks regarding use of AI in the particular topic / solution regarding both correctional staff's and offenders' rights etc.;*
- *Impact Assessment means how will AI change processes in the particular field of application, what is the expectance for it to make things better / faster / more optimal etc.;*
- *Governance Assessment means how do we regulate it, does it need some new legislative or regulative policies;*
- *Mitigation and Evaluation means how to mitigate risks, how to monitor in real time its use and how to evaluate the results – what is the process of correcting possible flaws it makes etc.*

[2] For information, although not carryng an identical idea, Articlze 22 of the GDPR states: "The data subject shall have the right not to be subject to a decision based solely on an automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her."

Risk management and mitigation frameworks set up in previous phases should be evaluated, adapted and maintained during the deployment phase.

AI should be continuously evaluated and studied in order to ensure both that they function properly and that they really produce the expected benefits too. A constant and preferably real-time assessment is necessary to prevent biased use or misuse of these systems and possible harms that they could produce. Any detected harms should be analysed immediately and taken responsibility to correct the harms, and if necessary, cease to use these systems if harms can't be prevented. AI itself should not be blamed or made responsible for harm - it is human's responsibility to control, develop and govern these systems.

8. AI and related digital technologies should be used in a manner which preserves and promotes positive and beneficial human relations between staff and offenders as this is instrumental in changing behaviour and in ensuring social reintegration **(principle of human-centred use of AI and related digital technologies)**.

AI need to be human-centred. While offering great opportunities, AI also give rise to certain risks that must be handled appropriately. The socio-technical environments need to be trustworthy, and designers and manufacturers of AI and related digital technologies need to be aware and need to strive not only to make profits but also to seek to maximise the benefits of AI while at the same time preventing and minimising their risks. (EU High Level Expert Group 2019:4).

Risks should be avoided of using AI and related digital technologies for creating unemployment by intelligent machines taking over core professional tasks including cognitive and affective tasks from human workers: the atrophying of certain human skills when AI replaces or augments human workers; the withering away of certain occupational practices and "embodied knowledge" when machines can do this in lieu of people; the instrumentalising or degrading of staff-offender relationships if, instead of dealing with them on a genuinely interpersonal basis, the contact is more and more mediated via machines, which collect and codify data on them in the course of every encounter (or even constantly, if they are monitored with tracking devices); the monitoring of employee's performance and productivity in workplaces can be massively augmented if sensors (wearable and/or embedded in buildings and equipment) and software systems - not necessarily full AIs - are used to gather, analyse and compare data with an unprecedented degree of granularity.

9. All AI and related digital technologies users should understand the basics regarding what this use implies, including how and for what purpose and what are the ethical rules to be respected **(principle of "AI and digital literacy").**

Prison and probation staff should be consulted and engaged about the coming of AI and the future shape of their work assisted by AI. AI literacy should be actively promoted by the prison and probation services. All staff should have the opportunity to learn basics of AI and ethics of use of AI and have proper training to be able to use the planned AI in their work. Managers and senior staff members should know more as they are involved in decision taking.

Reference to the European Code of Ethics for Prison Staff and to the Council of Europr Probation Rules to be added in.

Investment in capacity building (initial and continuous training and education) of staff and awareness raising about the benefits, risks, capabilities and limitations of AI and related digital

technologies, and through enabling public interest research. Such skills should encompass theoretical as well as practical knowledge on the interplay between the design, development and application of AI on the one hand, and human rights, democracy and the rule of law on the other hand.

AI literacy has also been actively promoted in Finland where all prisoners and probationers can access online basic course on AI from joint use workstations placed in every unit (prisons and probation offices). This course is recommended both to staff and offenders. Finland is also developing a new offender management tool RISE[3] AI, which will be an AI-based component in the new offender management system to help with assessing offenders and orienting them to most suitable services and units during their sentence. Educating both staff and offenders to understand how this AI-based process is going to facilitate offender management in the future will deepen understanding of both key processes and AI literacy. Besides this RISE AI will help to make the whole offender management cycle faster, more cost-effective and optimize compatibility of services and offenders' needs.

III. Data Protection

10. Offenders continue to enjoy their fundamental rights and freedoms, including the right to respect for private life and the right to data protection when AI and related digital technologies are used. Limitations to these rights and freedoms should only be allowed when they are in accordance with law, pursue a legitimate aim, and are necessary in a democratic society and are proportionate.

The use of artificial intelligence or related technologies in the field of execution of penal sanctions and measures may require massive processing of different types of data, particularly personal data, both for the use of AI to be effective and for it to avoid biases or errors. The difficulty in predicting which elements of the data should be selected as relevant for the objective of the AI should be balanced against the need to minimise or limit access to data in order to respect private life of service users as much as possible.

The use of any AI or related technologies in the field of prisons and probation must respect the ethical principles laid down in the European Convention on Human Rights and its additional Protocols, the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (Convention 108+) and the UN Universal Declaration of Human Rights. The legal imperative imposed by The General Data Protection Regulation (EU)2016/679, Directive 2016/680/EU[4] and Council Framework Decision 2008/977/JHA[5] must also be respected by the EU member States.

11. All key actors, regardless of whether public or private, participating in the design, development, maintenance, and use of AI and related digital technologies should comply with data protection rules, and the offender's free and informed consent for accessing and processing their personal data should be obtained if required by law.

---

[3] RISE is short for Rikosseuraamuslaitos, which is the Finnish name of the agency

[4] Directive 2016/680/EU – Directive of the European Parliament and of the council on the protection of individuals with regard to the processing of personal data by competent authorities for the purposes of prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and the free movement of such data.

[5] Council Framework Decision 2008/977/JHA of 27 November 2008 on the protection of personal data processed in the framework of police and judicial co-operation in criminal matters.

Data protection regulations must be complied with from the very moment of designing AI until its use in the field of criminal justice system.

AI can be very intrusive for the private life of the persons concerned as they collect and process a lot of personal data. Therefore, the access and the use of such data should be strictly regulated and should in the majority of cases be done after obtaining the free and informed consent of the person.

In prisons other considerations may lead to exceptions to this rule for security, safety and good order reasons which may justify the need for interventions without obtaining such a consent.

12. Data should be stored in a form that allows a personal identification for no longer than is strictly necessary to fulfil the purposes for which it was initially collected. Data managers should adopt security measures to ensure the integrity and confidentiality of the stored data, avoiding unauthorised access, alteration, retention, destruction, or disclosure of personal data.

Situations where attempts to save data for longer periods than allowed in case it becomes necessary in the future should be avoided and sanctioned. On the other hand, data security, in general, and the adoption of cybersecurity measures, in particular, are essential to prevent improper and illicit access to data, like data related to health and medical care, finances and salary, to HR, data related to Offender Management System (OMS), data related to incident reporting, or to planning and transportation of inmates

13. Data managers should ensure that the data they use to train the AI and related digital technologies is accurate and the samples are sufficiently representative of the key characteristics of the general population and minority groups, including the target groups that might be affected.

For the good quality design and effective use of AI a big amount of variety of data samples should be fed in the algorithm to the extent that it is in line with the applicable law. It is important to highlight that the quality of the data not only depends on a general representativeness, but also on the fact that they correspond to the existence of different minority groups so that ultimately, they are not discriminatory due to their lack of representation.

14. The collection and processing of special categories of personal data should be generally avoided unless it is strictly necessary, and data managers should adopt additional safeguards regarding their storage and disclosure. AI and related digital technologies that are based on sensitive data, such as biometric data, should be used in controlled environments to avoid false positives and indiscriminate in gathering the data.

(Add in a reference to the Commentary of Article 6 of the Convention 108+)

Despite the existence of legal frameworks regulating the processing of personal data, it is important to note that these frameworks remain not well defined when it comes to regulating such processing by public authorities including security services. This is because, both in the regulations that are oriented towards processing by private entities, as well as those that directly regulate the processing of personal data by public authorities including security services, there are exceptions that legitimise non-respect of the right to privacy when the protection of the public interests or public safety so require.

IV. Use of AI and Related Digital Technologies

A. Use for the purpose of safety, security and good order

15. The use of AI and related digital technologies for maintaining safety, security and good order should allow for better risk and crisis management. Its use should be strictly necessary, proportionate to the purpose and should avoid any negative effects its intrusive nature may have on the privacy and well-being of offenders and staff.

Safety and security via surveillance is one of the most important functions within prison and probation services and there are many AI-technologies that can be used to support staff in this area. With AI, it becomes possible to automate tasks that have formerly required human capabilities, opening not only to increased efficiency, but also to increased quality and effectiveness.

The use of AI and related digital technologies for the purpose of safety, security and good order is an important function within prisons and probation services and it requires close attention to the principle of "human-centric" in relation to the risk of decreasing meaningful human contact while implementing AI. Security processes should allow alleviating staff from habitual repetitive tasks like opening and closing doors, monitoring movements and behaviour, etc. and this should be used to help staff develop and maintain positive human relations thus enhancing rehabilitation and social inclusion of offenders.

Image recognition AI technology can be coupled with the CCTV systems and can be used to recognize unduly behaviours such as violence, smuggling contraband, handling drugs and other forbidden objects or harmful behaviour such as suicide attempts. This would allow for new levels of surveillance where many deviant behaviours could be detected and prevented. Such an AI could monitor cameras and alert staff if suspicious situation is noticed. This technique could be further developed with facial recognition techniques capable of identifying inmates and staff, tying them to certain events or incidents.

Audio recognition capabilities could be coupled with telephones or microphones in prisons. In this case unduly talks and behaviour can be detected, but it may also be possible to gather intelligence about offenders and their interlocutors which could be used to inform investigations.

Movement analysis is yet another technique in which an offender's position and behavioural patterns can be tracked and analysed for purposes of surveillance and intelligence. AI may also be used in different kinds of predictive analysis. With machine learning it is possible to analyse vast datasets to reveal novel patterns and perform complex statistical tasks. This can be used to optimize operational functions like occupation and transports but may also be used to predict certain behaviours like violence or attempts to escape from prison or to escape justice.

The different techniques mentioned above could also be used in combination to create a surveillance system which allows for more complex analysis based on multiple data sources. When using AI in security and monitoring tasks, the intrusive nature of heightened surveillance should be considered. With AI, the level of effectiveness may become significantly higher and lead to a state of control that is unwanted. Constant and ubiquitous surveillance may have unintended consequences and stand in violation to prevailing laws or the human rights. Collection of data on a massive scale may also be considered intrusive and infringe on data privacy laws if left unchecked. Increased levels of surveillance and a feeling of being monitored at all times may also cause psychological stress among

offenders and staff and could lead to detriment in their wellbeing. With new possibilities for detection and intervention it may also be possible to design processes that restrict or control the behaviour of offenders or staff. This may seem an attractive or tempting proposition but can lead to serious infringements on human rights or prevailing legislation. Extensive automation and technification of person processes may also lead to a decrease in meaningful human contact. This could be seen as depriving persons of their dignity and may also be an impediment to rehabilitation.

While implementing AI in the area of security and surveillance it is of great importance to consider the principle of necessity, proportionality and efficacy. The level of monitoring and control should not be excessive and should stand in proportion to the intended purpose. It is not the purpose to accelerate and intensify control and monitoring in a way that produces more harm than benefits. People's privacy and integrity should not be violated more than necessary to ensure security.

It should be noted in this context that it is not permissible, for example, to use AI to control access to the Internet or to limit in any other way activities or rights that probationers are not restricted explicitly from doing by the competent body's decision.

The majority of AI applications in the prison context relate to security. They raise the question of whether digital security technologies, managed by AI, are more or less intrusive than traditional means of security, and whether they can be operated in a way that is more supportive of rehabilitative practices within an institution. Examples of these include surveillance systems using facial identification and movement analysis or other methods to detect suspicious behaviour. Also in probation, the most AI application is for security and surveillance purposes like electronic monitoring (EM).

The principle of legality and legal certainty should also be considered so that AI is used in a way that respect existing legislation within the used area. Extended collection and analysis of data may have unintended consequences that can lead to violation of prevailing laws or personal rights. Neither should. The principle of respect for human dignity and fundamental rights must be respected.

16. Prison and probation services should be consulted in order to identify and evaluate the needs regarding replacing manpower by AI and related digital technologies in the execution of everyday routine tasks related to safety, security and good order. This will allow to design and use well adapted AI and related digital technologies in a manner allowing the redeployment of staff to other tasks which contribute to better preparing offenders for their social reintegration.

Who should consult them? (explain)

Robots, chatbots and AI-based behavioural change programmes have the possibility to perform cognitively more challenging tasks than the routine work only. However, also they function still in a limited way and are supposed to assist but not replace human interventions. However, in the future the development may bring more possibilities also with these more challenging tasks. It's up to humans to analyse which tasks can or which can't be replaced by AI applications. In any event the human control over the applications should always be there.

The notion that one of AI-based automation's most important achievements is, or will be, the shift of employees' energies away from "routine tasks" towards more important, "non-routine" tasks is commonplace in much of the literature that straightforwardly champions AI, including that from the European Institutions. It is a rather dubious argument. Much depends on what is defined as "a

routine task". It is useful for AI's champions to promote AI as a benign and limited measure that will merely automate dull, routine, back-office tasks but leave the recognisably core tasks of a profession, the human expertise which give it its identity, intact. But that may not be so: fully professional expertise is already within AI's purview. Much will depend on the economic and political value which is attached to these traditionally human/professional tasks.

A danger that needs to be avoided is replacing staff by AI not only assisting them. Positive, meaningful human contact with inmates should never be replaced by a machine and staff should be retrained and developed to use their intellectual and emotional capacities and qualities to invest in helping offenders desist from future offending.

17. The use of AI in electronic monitoring, including biometric recognition technologies should be proportionate to the propose and used only when strictly necessary. It should be carried out under regular human control and should be human-centred. It should be oriented to favour the reintegration of offenders and should be respectful of all the principles and guarantees associated with the use of electronic monitoring and of the current recommendation.

Rec(2014)4 on electronic monitoring contains very detailed rules, including ethical rules on the use of EM. The current rules apply in addition to it in case of use of AI and elated digital technologies.

During probation, electronic monitoring systems can use AI techniques to facilitate management the supervision of offenders and the decision taking. AI can either store and forward or do a real time supervision and collection of data, based on the automation of some functions such as the generation of automatic alarms in the event of non-compliance. In these cases, the automation of functions should be limited as much as possible to ensure the possibility of reviewing incorrect automated decisions and always incorporating a human perspective. It is recommended that simplification in the use of these systems should not lead to an increase in the use of electronic monitoring beyond what is necessary. It is also recommended not to authorize the use of remote immobilization systems for persons on probation due to the incompatibility between their needs and rehabilitation aim; and, in any case, the automation of such acts should be absolutely forbidden.

Particular caution must be taken with the use of "remote immobilization of offenders" systems that allow the inclusion in electronic monitoring bracelets or anklets of a "conducted energy device" (CEDs), which would perform the function of electric pistols and which, by means of AI, would supposedly be capable of detecting a breach of the law and of detecting the presence of an offender. Once the breach is detected, these systems would supposedly be able to trigger an alarm for the manager to make the decision to activate the alarm or be automatically activated by the AI through "a signal sent over the Internet" immobilizing a person until law enforcement officers can come to arrest him or her. From an ethical point of view this possible use of the technology generates too many risks to be applied in probation cases and exceeds the objectives of traditional electronic monitoring systems, not only because of the risks of potential misuse by law enforcement agencies, but also because of the damage that can be caused by its faulty or negligent use.

We should also take into account the risk that the incorporation of AI may lead to a simplification of their use, extending electronic monitoring beyond what is strictly necessary, leading to forms of authoritarianism and hyper-control, as well as unnecessarily increasing the cost of the penal system. Monitoring, in general, should avoid automation in any decision making that may involve harm or restriction of rights both to the persons subject to monitoring and to others who relate to them. Of course, constant review and auditing, both legal and ethical, is necessary when using these tools.

B. Use for Offender Management purposes: Risk Assessment, Rehabilitation and Reintegration

18. AI and related digital technologies should be used with care to automate contacts, manage appointments, assign electronic devices and ensure control processes in order to facilitate communication of offenders with family, professionals and relevant services.

AI offers the possibility to alleviate the routine tasks staff are responsible for on an everyday basis like assisting offenders in their contacts with lawyers, possible employers, psychologists, social workers and others as well as with their families. This is not only the case in prisons but also for probationers.

Nevertheless, the use of AI in such cases should be done with care because of different reasons: language and technological inabilities; psychological difficulties; young or old age and other.

19. Where AI is used to manage offenders' files and cases and generate automatic alerts in cases of non-compliance, it should be used if it improves monitoring and decision-taking, the final responsibility for which remains with the professionals. The human-centred approach should remain a key element in decision making.

In some jurisdictions, mostly outside Europe currently, such AI tools are already in use which is aimed at improving file management and offender management. Nevertheless, staff should take the final decision regarding how to manage a case in situations of non-compliance as the reasons behind differ in each individual case.

AI should not replace humans in decision making processes, but work in tandem with them, supplying precedents, recommendations or options for a particular course of action, leaving human professionals and managers to take final decisions based on more accurate, comprehensive and objective data and information than collected with traditional methods. AI's role in the offender management systems (OMS) should be advisory. Use of AI should be evidence-based.

AI applications have already been used to some extent also in the offender management systems (OMS) and processes. Examples of these include for example automated risk assessment and service orienting.

AI has the possibility to support decision making during the entire offender management cycle including assessment and classification of offenders and planning, executing, evaluating and adjusting services for offenders. The ambition of using AI in the context of offender management lies in the desire to improve decision making related to finding the best trajectory for the offenders regarding their needs and minimizing their risks. For example, the Hong Kong Prison department states they are actively developing AI technologies for persons in custody self-management in order to enhance the efficiency of penal operations and even the effectiveness of rehabilitation programmes (Houser, 2019). A recent project in the Finnish Criminal Sanctions Agency is developing an AI application for offender management. RISE AI will be a recommender system that recommends rehabilitative services to offenders during their sentences based on the available offender background information, like various criminogenic risk factors. This application will complement the risk and needs assessment tools currently in use, thereby improving the accuracy of service

recommendations made to offenders. Here 'accuracy' is referring to meeting offenders needs and reducing their risk for re-offending (Puolakka, 2020).

20. When developing AI and related digital technologies in order to increase the accuracy and objectivity of risk assessment, the challenges of algorithmic biases and unrepresentative data sets should be addressed. AI and related risk assessment tools should not replace professional decisions and regular face-to-face human contact in the rehabilitation work with offenders.

Move the accuracy bit of previous commentary here

The first and still most common applications of AI technology are evident in the context of risk assessment tools (Pereira, 2020). Most of these models are based on the original and still dominant risk-need-responsivity (RNR) model of risk assessment (Andrews, Bonta, & Hoge, 1990). To support this model the adoption in many jurisdictions of standardized instruments for risk and needs assessment is one of the most important, widespread, and continuing developments of the last 20 years in offender management (Raynor, 2019).

Experts should be enough acquainted with both AI and criminological research in order to develop reliable and valid AI for the use Offender Management Systems (OMS). The experts should understand that prison and probation services are dealing with already stigmatized and vulnerable persons in the majority of the cases. This means there's a risk for stereotypical, discriminative, and ex post facto type of conclusions that can be repeated in AI if this fact is not taken into consideration in the algorithms used.

Many ethical questions are related to the of AI in offender management. At best algorithms can overcome the harmful effects of cognitive biases (Sunstein, 2018), but they can also easily be biased and start to repeat the same mistakes humans make. Designing an algorithm for use in the prison context requires thinking deliberately about what it is that we exactly want to achieve and a solid understanding of the human failings they're supposed to be replacing (Fry, 2018).

AI can be used in education and training platforms / systems, in various treatment procedures and in preparation for release, rehabilitation and resocialisation. In the rehabilitative practices AI offers possibilities for the use on Virtual Reality (VR) for rehabilitative purposes and behaviour modification (Teng & Gordon, 2021 and Pires et al., 2021). AI can assist rehabilitative processes, and programs or individual therapeutic work can include AI-based methods like VR, but no rehabilitative work should be solely based on AI without human in the process.

The use of robotic systems for rehabilitative tasks besides security tasks is another ethical question. Some are discussing the possibility of using AI to address the solitary confinement crisis in the US by employing digital assistants, similar to Amazon's Alexa, as a form of 'confinement companions' for prisoners. Even if these 'companions' could alleviate some of the psychological stress for some prisoners, these companions might actually contribute to the legitimization of solitary confinement penal policy instead of questioning it (Završnik, 2020). Considering that AI chatbots and virtual assistants are already used to some extent in civil health care, it is a relevant question to ask if and how these solutions could be used in a meaningful and rehabilitative way in the prison setting.

This brings another concern relate to the use of AI and robotics in society: occupations can disappear while AI is taking over the job humans used to do in a faster and more accurate way. However, AI can also bring new occupations. One such example is shown in a pilot in Finnish prisons, where prisoners

were training AI algorithms (Newcomb, 2019), which also shows the possibility to provide prisoners with new job-related and digital skills to help them successfully re-enter into the modern society and labour market.

C. The use of AI and related digital technologies for Staff Selection, Management, Training and Development

21. The use of AI and related digital technologies in the selection, management, training and development of staff should be used to optimize human and managerial capacities and processes and predict future organizational capacity. It should be focused towards supporting the staff's professional qualities and development and should help balance their professional and family life.

Possible uses of AI in human and managerial processes can include selection and recruitment process, staff training and budget and financing. However, cost-effective use of resources should support staff well-being and help balance their professional and family life instead of benefiting only organisational, material and financial purposes. AI can help detect especially problematic areas in staff resourcing and expensive processes and practices. Decisions should not violate staff rights or lead to discrimination, inequalities and unfairness.

Real time information provided by AI can help optimize the use of resources and understand how the organization and staff are performing. All this can assist better decision making on the organizations' management level.

When using AI to assist decision making and managerial processes, there should be a clear understanding of what kind of data the particular system is using. The problems in the data itself mean lack of enough clean, accurate or enough well documented data. Biased data can lead to biased algorithms which can mislead decision making and proper managing of resources. It is true that AI-driven analysis and decision making can correct the biases that are present in human decision making without allowing heuristics, stereotypes, emotions and other irrelevant but "humane" factors interfere with objective analysis. However, evidence has shown that also algorithms can easily be biased and start to repeat the same mistakes humans are prone to. This shouldn't be surprising considering that AI is only using the data and weighing defined by humans and can't do much more then simulating human (statistical) decision making.

Within management and HR, AI can be used to create decision-support systems capable of handling complex tasks and data-analysis in a way that is not possible with conventional software. AI capable of natural language processing can make meaningful analysis of unstructured textual data, meaning that it can read and understand text written by humans and make predictions or decisions based on it. With machine learning it becomes possible to analyse and mine the enormous data sets that exist within HR to find novel patterns and make predictions. AI can also be used in chat bots to communicate with inmates, employees, job applicants or third persons.

In the recruitment process, AI can be designed to analyse CVs and motivation letters and make initial selections or evaluations. This will not only result in great time savings but will also lead to a greater chance of hiring the right person. Research shows that most applicants are not included in the recruitment process since recruiters do not have time to consider their application files. With AI, every single file would be taken into consideration in the initial selection phase. Furthermore, with this procedure, recruiters can concentrate on insightful evaluation of the applications selected by the AI rather than having to sift through a large number of unsuited applicants.

AI can also be used to check the social media profiles and online presence of the chosen candidates. This can be done to see if information in the application is consistent or look for signs of activity inconsistent with the authority's policy, such as racist or xenophobic comments.

As mentioned above, AI can also be used to mine and analyse the vast data sets collected and maintained by management and HR departments to find novel patterns and make predictions. Such analyses can be used to create applications that support functions such as internal mobility, employee retention, and employee health and satisfaction. AI could for instance make recommendations based on personal data about the suitability of employees for certain positions. It may also be possible to create smart surveys for employees to evaluate level of satisfaction and wellbeing or detect declining mental health.

Relating to this area it is important to know that AI based on machine learning technology is prone to become biased. There are an abundant number of examples where machine learning systems have exhibited unintentional reproduction and reinforcement of social biases. In fact, it is often hard for AI-engineers to counteract bias in AI even as the problem is well understood. If there is bias present in the input data that is subsequently used to train the AI, the bias will probably persist and may even become amplified. This could in practicality result in the AI taking into consideration traits such as gender or skin colour when making selections or recommendations, or in other ways functioning in a discriminatory or non-democratic way.

The problem with bias in AI is exacerbated by the fact that machine learning models are often opaque. Opacity in machine learning means that it is difficult to fully understand how the model works, even for experts. It may be perceived as a black box were input goes in and output comes out, but the patterns and attributes used by the model to make its predictions are unobservable by humans. This means that we cannot explicitly see how the model is reasoning and must therefore do extensive and continuous testing and verification while working with machine learning. However, there is also a growing field of research called XAI or Explainable AI, striving to make the decision processes of AI easy to perceive, detect and understand.

This being said, AI-technology can also be used to reduce unconscious human bias in HR-processes by replacing human decision-making in certain steps. A well trained and highly functional AI with low bias will outperform any human in long term consistency and objectivity. Paradoxically, AI could become the safeguard of fair and democratic processes in HR management.

While working with machine learning it is important to consider the principle of quality and the principle of equality and non-discrimination. It is paramount to train machine learning models on data that is representative and of the highest possible quality. If there exist bias in the training data, the trained model will likely exhibit that bias in its predictions.

Furthermore, machine learning applications must be subject to extensive testing, verification and logical proofing before implemented in any ethically sensitive environment to ensure that it behaves as intended.

To the extent possible, machine learning models should also be explainable, meaning that its reasoning should be transparent to human observers. This relates to the principle of good governance, transparency and traceability.

After implementation, the predictions of machine learning models can never be trusted blindly but must be continuously evaluated and tested by trained staff. Therefore, it is important to consider the principle of AI and digital literacy. Knowledge about AI and awareness of the risks should be promoted among staff working in close vicinity to the system and awareness promoted among those who are affected by the systems output.

VI. Research, Development, Evaluation and Regular Revision

22. Subject to certain limitations, the development and design of, as well as the research in, AI and related digital technologies should be sufficiently well funded and supported. It should be carried out with due consideration of data protection rules, with anonymised data published and should help develop further the proper and efficient use of AI and related digital technologies and help prevent causing harm.

Research is important to evaluate and monitor whether AI produce supposed benefits and whether they can support effective practices in security, offender management and human resources. Development of AI should be evidence-based. Regular revision of AI is important to ensure that they function in a proper and ethical way and don't produce biased results. The maintenance, development, evaluation and revision of AI should be done by experts in the specific field and should be evidence based. Self-learning (machine learning) systems still need ongoing evaluation and revision by humans.

23. AI and related digital technologies and their use should be evaluated at regular intervals by independent and competent evaluators concerning their performance, their intended and unintended outcomes and the need for adaptations. The initial funding should include or take into account the follow-up costs for implementation and evaluation.

References

Andrews, D. A., Bonta, J., & Hoge, R. D. (1990). Classification for effective rehabilitation: Rediscovering psychology. *Criminal Justice and Behavior*, 17, 19-52.

Fry, H. (2018). *Hello World - Being human in the Age of Algorithms*. W.W. Norton & Company Ltd.

European Commission (2018). *Coordinated Plan on Artificial Intelligence, Communication from the commission to the European Parliament, The European Council, The Council, The European Economic and social committee and the Committee of the Regions*. EC, COM (2018) 795 final.

Houser K. (2019, February 4). *China is Installing 'AI Guards' in Prison Cells. They'll make escape impossible - but the trade-off might be inmates' mental health*. Futurism. https://futurism.com/chinese-prison-ai-guards-cells

McGoogan, C. (2016, December 6). Liverpool prison is using AI to stop smuggling drugs and weapons. *The Telegraph.* https://www.telegraph.co.uk/technology/2016/12/06/liverpool-prison-using-ai-stop-drugs-weapons-smuggling/

Newcomb, A. (2019, March 28). *Finland Is Using Inmates to Help a Start-Up Train Its Artificial Intelligence Algorithms.* Fortune. http://fortune.com/2019/03/28/finland-prison-inmates-train-ai-artificial-intelligence-algorithms-vainu/

Pereira, A. (2020). *Artificial Intelligence, Offender Rehabilitation & Restorative Justice. The "Good" Algorithm? Artificial Intelligence: Ethics, Law, Health.* International Workshop organized by the Pontificia Academia Pro Vita. Date: 2020/02/26 - 2020/02/28. New Hall of the Synod, Vatican City. https://limo.libis.be/primo-explore/fulldisplay?docid=LIRIAS2960856&context=L&vid=Lirias&search_scope=Lirias&tab=default_tab&lang=en_US&fromSitemap=1

Pires, A.R., Fernandes, A., Estalella, G., Zisiadou, M., Carrolaggi, P., Loja, S., & Leitão, T. (2021). The Potential for Virtual Reality for Education and Training in Prisons. Available at: https://www.researchgate.net/publication/357752509_THE_POTENTIAL_OF_VIRTUAL_REALITY_FOR_EDUCATION_AND_TRAINING_IN_PRISONS

Puolakka P. (2020). *RISE AI: Reducing the Risk of Recidivism with AI.* Aalto Executive Education: Diploma in Artificial Intelligence. Unpublished.

Puolakka, P., & Van De Steene, S. (2021). Artificial Intelligence in Prisons in 2030. An exploration on the future of Artificial Intelligence in Prisons. *Advancing Corrections Journal*, Edition # 11, ICPA.

Sunstein, C.R. (2019, January 23). Algorithms, Correcting Biases. *Oxford Business Law Blog*. https://www.law.ox.ac.uk/business-law-blog/blog/2019/01/algorithms-correcting-biases

Raynor, P. (2019). Development, critics and a realist approach. In Ugwudike, P., Graham, H., McNeill, F., Raynor, P., Taxman, F. S., & Trotter, C. (Eds.). *The Routledge Companion to Rehabilitative Work in Criminal Justice.* ProQuest Ebook Central.

Sunstein, C.R. (2019, January 23). Algorithms, Correcting Biases. *Oxford Business Law Blog*. https://www.law.ox.ac.uk/business-law-blog/blog/2019/01/algorithms-correcting-biases
Yampolskiy, 2016

Teng, M.Q., & Gordon, E. (2021). Therapeutic virtual reality in prison: Participatory design with incarcerated women. New Media & Society, 23(8), 2210–2229.

Završnik, A. (2020). Criminal Justice, artificial intelligence systems, and human rights. *Academy of European Law - ERA Forum,* 20, 567-583.