COUNCIL OF EUROPE
CONSEIL DE L'EUROPE

Strasbourg, 15.02.2022 PC-CP (2021) 17rev

COUNCIL FOR PENOLOGICAL CO-OPERATION
(PC-CP)

Draft Committee of Ministers Recommendation CM/Rec(2023)XX

Ethical and Organisational Aspects on 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 MIRO
Professor of Criminology and Criminal Law, University 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 related to the use of digital technologies, including artificial intelligence by prison, probation and juvenile justice 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 also 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;

Recognising that the rapid development and use of digital technologies, as well as of artificial intelligence 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 digitalisation of justice is advancing at a great pace in Europe and is gaining more and more place in particular at all stages and in all areas of the criminal justice process;

Noting also that digital literacy needs to be enhanced among key actors in the criminal justice system like police, prison and probation staff 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 their service users.

Drawing attention that these tools need to be designed and developed by carefully selected and vetted private companies in consultation with the prison and probation services and these companies should be made aware that high ethical norms and principles and strict professional rules should be respected, and the profit should not the main driver;

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 rehabilitation and social inclusion of persons in conflict with the law and that the reintegration of offenders should remain central. This use should not undermine the human-centred approach and should avoid discrimination and inequalities;

Endorsing the standards contained in the relevant recommendations of the Committee of Ministers of the Council of Europe and in particular Recommendations: 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.

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 such 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 develop, sell, 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 rapid pace of advancement in including AI digital technologies, as well as of 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 basic freedoms of its users and safety and security of our societies.

Currently AI, algorithmic tools and related technologies are already used in some correctional services across the world, but according to a recent review most jurisdictions still don't use it and almost none have any policies or legislation regarding its use in corrections. Because it is so little used, there's also little research about the results, benefits and risks of its use. (Puolakka & Van De Steene, 2021). The drivers of AI's use in corrections 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 prison and probation systems should explore the possibilities of AI too.

This Recommendation recognises that the rapid pace of scientific and technological change requires reviewing it biannually 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: Artificial intelligence refers to systems which enable computers, robots or other machines to analyse their environment 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 system 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.*

Deep learning, categorising, translation, natural language; optimising techniques; automatic planning, expert systems, recommending systems

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.

Digital technology: This generic term refers to all electronic devices, automatic systems, and technological resources that generate, process or store information.

The difference between analogue and digital technology is that in analogue technology, data is converted into electric rhythms of multiple amplitudes, while in digital technology; information is converted into the binary system, i.e. zero or one, where every bit is the symbol of two amplitudes.

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 experience 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.

Cybersecurity and security techniques (video vigilance; suicide prevention and aggression prevention; prevention of illegal items smuggling):

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.

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.

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.

II. Basic principles

1. Principle of respect for dignity and fundamental rights: Ensure that the design and use of digital technologies and AI tools and services are compatible with human dignity and fundamental rights of all service users.

2. Principle of equality and non-discrimination: Take positive measures, when designing or using new technologies and AI, to prevent or to resolve the creation or intensification of any inequality or discrimination between individuals or groups of individuals.

Social prejudices and stereotypes can turn into algorithms if we don't take care to understand how algorithms are formed and what kind of data they use. This is especially harmful with already vulnerable groups if algorithms start to repeat and validate the biases we have in human thinking. Data itself is not biased but the conclusions extracted by humans from these data can create or deepen biases and the algorithms should rectify this instead of intensifying it.

3. Principle of quality: Certified sources, tangible data and validated scientific methods and values should be used, with models conceived in a multi-disciplinary manner, in a secure technological environment. Technically accurate, reliable and secure AI and related digital technologies should be used.

For the development of AI systems 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 correctional 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;*
- *Impact Assessment;*
- *Governance Assessment;*
- *Mitigation and Evaluation.*

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

4. Principle of legality and legal certainty: ensure that all processes related to the development, provision, operation and maintenance of AI and related digital technologies as well as all decisions and actions taken by the prison and probation services and the private companies acting on their behalf are clearly defined by law and are in compliance with the relevant national and international law and policy.

National policies and regulations should be established regarding the use of AI and related digital technologies also in the prison and probation services. These policies and regulations should follow the principles and recommendations stated on the international level by the Council of Europe.

As a minimum there should be provisions on access to effective remedy, a mandatory right to human review of decisions taken or informed by AI and related digital technologies except where competing legitimate overriding grounds exclude this, and an obligation for public authorities to implement adequate human review for processes which are informed or supported by AI sand related digital technologies and to provide *relevant individuals or legal persons with meaningful information concerning the role of AI systems in taking or informing decisions relating to them, except where competing legitimate overriding grounds exclude or limit such review or disclosure. Adequate and effective guarantees against arbitrary and abusive practices due to the application of an AI and relate digital technologies in the public sector should be afforded by the national law.*

These safeguards should, at least, include the following: the right to an effective remedy before a national authority (including judicial authorities) against such decisions; the right to be informed about the application of AI and related digital technologies in the decision-making process; and the right to choose interaction with a human in addition to or instead of an AI system, and the right to know that one is interacting with AI and related digital technologies rather than with a human.

5. Principle of necessity, adequacy and efficacy: Before implementing AI and related digital technology tools, their use and impact should be discussed with the prison and probation management level in order to ensure that AI systems will be fit for the purpose and will support the strategical targets of the organization. Safety, security and offender management should be key indicators in decision taking.

How to prove necessity, the rationale behind the use of AI tools, they should be fit for the purpose

6. Principle of good governance, transparency and traceability: Society should participate in the process of developing and use of AI and related digital technologies and should be kept informed and engaged in decision-taking. The information about design, operation and data processing methods should be non-opacite, accessible and understandable, external scrutiny should be ensured as this brings effective responsibility and accountability.

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 fulfil the General Data Protection Regulations (GDPR) of the European Parliament and the Council of the European Union (2016). The European Ethical Charter on the use of AI in judicial systems and their environment (European Commission, 2018) has defined the following five principles which are to be taken into account also in the prison and probation services. Good governance requires society to be informed and involved as far as possible in the process of developing and use of AI.

AI systems 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.

The establishment of public registers listing AI systems 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.

7. Principle "under user control": A prescriptive approach should be avoided and users of AI and related digital technologies should are informed actors and in control of their choices.

Service users include staff, offenders, family members as well as any other person impacted by the use of AI tools.

Prison and probation staff and clients should be informed about the coming of AI and the future shape it will have on them. Their AI literacy should be actively promoted. 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.

8. Principle of digital technology being "human-centric": Positive human relations are instrumental in changing offending behaviour as they help enhance rehabilitation and social inclusion of offenders. In addition, 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.

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

Risks should be avoided of using AI tools 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. Principle of "digital and AI literacy" – staff and users in general should be trained on the basics regarding how and for what purpose to use AI and related digital technologies and on the ethical rules to be respected in relation to this.

Prison and probation workers should be consulted and engaged about the coming of AI and the future shape of their work assisted by AI systems. 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 AI and have proper training to be able to use the planned AI systems in their work. Managers and senior staff members should know more as they are involved in decision taking.

A legally binding or non-legally binding instrument on AI in the public sector could address measures to increase digital literacy and skills among both civil servants and the general public, notably through investment in capacity building (initial and continuous training and education) of public officials and awareness raising about the benefits, risks, capabilities and limitations of AI systems, 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 systems on the one hand, and human rights, democracy and the rule of law on the other hand.

III. Use of AI and related digital technologies in prisons

A) Use for safety, security and good order purposes (safeguards)

a. Proportionality, data protection, privacy, legality, human factor, video vigilance

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 smart 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).

When using AI systems in security and monitoring tasks, the intrusive nature of monitoring censors should be taken into account. It's not the purpose to accelerate and intensify control and monitoring in a way that produces more harm than benefits. Neither should it lead to more extensive and constant monitoring than currently or produce even more punitive and penal interventions. People's privacy should not be violated more than necessary in order to ensure security.

b. Repetitive safety and security tasks of little value can be done by AI

If some AI systems would lessen the need for human work force in routine tasks / manual work, impacts of these systems should be discussed with prison and probation workers before implementing these systems. This change should be taken as an opportunity to free them up to do other tasks, and the organizations should be able to show concretely what these other tasks are.

Robots, chatbots and AI-based behavioural change programmes have the possibility to perform cognitively more challenging tasks than the routine work described above. 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.

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.

B) Use for offender management purposes – RA, rehabilitation and reintegration (safeguards against biases)

a. AI systems should not be used as purely automated decision-making. AI systems 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 or managers to make final decisions on what is hopefully better information that they would otherwise have had. AI's role in the offender management systems (OMS) should be advisory. Evidence-based

AI applications have been used to some extent also in the offender management systems (OMS) and processes. Examples of these include for example automated risk analysis and service orienting. In the rehabilitative practices AI offers possibilities for the use on Virtual Reality (VR) for rehabilitative purposes and behaviour modification.

AI systems should not replace face-to-face contact in the rehabilitation work with offenders. It is vital to keep the active human presence in our daily and rehabilitative work in both prisons and probations. AI systems can *assist rehabilitative processes, and programs or individual therapeutic work can include AI-based methods like Virtual Reality (VR), but no rehabilitative work should be solely based on AI without human in the process.*

Experts responsible for the development of AI systems specifically tailored for offender management should be aware of the most common and possible risks of these AI systems like algorithmic bias and unrepresentative data sets. Experts should be enough acquainted with both AI systems and criminological research in order to develop reliable and valid AI systems for the use Offender Management Systems (OMS). The experts should understand that prison and probation services are dealing with already stigmatized and vulnerable group. This means there's a risk for stereotypical, discriminative, and ex post facto (definition?) type of conclusions that can be repeated in AI systems.

b. RA and crime prediction purposes
c. Education and training
d. Treatment
e. Preparation for release, rehabilitation and resocialisation

C) Use for management and HR purposes

a. Selection and recruitment process
b. Staff training
c. Budget and financing
d. Staff support and debriefing, balance between professional and family life

A third likely use of AI is in the management level of prison and probation services themselves to optimize human and managerial processes and predict future organizational capacity. Realtime information provided by AI systems 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 systems 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.

IV. Use of AI and related digital technologies by probation services

a. AI related EM and use of data to supervise and help
b. RA, crime prediction, offender management: education, training, treatment
c. Offenders' family
d. Selection and recruitment process
e. Staff training
f. Budget and financing
g. Staff support and debriefing, balance between professional and family life

Contemporary probation supervision, in its essence, is based on interactive processes of assessment, implementation and review, all undertaken by variously trained personnel, within a management structure. All stages of the process require some degree of conversation and dialogue. Risk/need assessment is becoming progressively more automated and "assisted decision support systems" are coming on the market *to take this further. The judgements made by a human probation officer about a client's risk level can already be informed by a machine. But what about the encounter itself? The more structured, focused and formulaic these conversations are required to be (as in customer service) the more easily a chatbot could routinely undertake them. It could counsel a client individually (with cognitive behavioural techniques), or lead them though a scripted offending behaviour programme, or a scripted restorative justice encounter (if the victim consented). It could award points and rewards to incentivise compliance. With machine learning, chatbots would improve over time, passing a client on to a human supervisor only when a certain threshold of concern had been passed, e.g. a breach decision. Would a client care one way or the other whether the voice that was advising or questioning him/her was a machine or a human? Should the client always be told? Why? AI's have already been trained to write short pieces of journalism, football reports, for example, culling data from internet. How long before an AI is required to write a pre-sentence report - or several dozen - simultaneously?*

VI. Private companies

a. Human-centred clauses and considerations clearly defined and negotiated
b. Safety clauses
c. Clearly set goals when signing contracts, defining why and what purpose AI tools are needed, how to be used, data protection, etc.
d. Maintenance and updating of the AI tools, etc.
e. A need to coordinate collection and use of data within prison between the different systems used and companies providing services, as there may be a problem

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.

VI. Ethical Guidelines and Data Protection Rules

Prison and probation services have traditionally been person-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 clients. These questions present an extra challenge to prison and probation services.

VII. Research, Development and Evaluation

Subject to certain limitations, the development and design of, as well as the research in, AI and related digital technologies should be carried out freely, with due consideration for safety and security, and in full compliance with the Council of Europe standards on human rights. To ensure prevention of unlawful harm potentially stemming from the development, design, and application of AI systems, including clarifying the concept of "unlawful harm" for the purpose of the transversal instrument on AI, human rights, democracy and the rule of law.