COUNCIL OF EUROPE
75
1949 - 2024
CONSEIL DE L'EUROPE

EXPLANATORY REPORT TO RECOMMENDATION CM/REC(2024)XX OF THE COMMITTEE OF MINISTERS TO MEMBER STATES

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 2024 at the XXX meeting of the Ministers' Deputies)

EXPLANATORY REPORT TO
RECOMMENDATION CM/REC(2024)XX OF THE COMMITTEE OF MINISTERS
TO MEMBER STATES

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 2024 at the XXX meeting of the Ministers' Deputies)

I. <u>General provisions</u>

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.

For the purpose of this Recommendation users include anyone who is using or who may be affected by the use of AI and related digital technologies. Users include prisoners, probationers, prison and probation staff, family members, visitors, lawyers, external organisations, etc.

II. <u>Definitions</u>

This definition is taken from Recommendation CM/Rec (2021)8 of the Committee of Ministers to member States on the protection of individuals with regard to automatic processing of personal data in the context of profiling.

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 also 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.

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.

III. Basic Principles

As the European Convention for the Protection of Human Rights and Fundamental Freedoms states that these rights and freedoms are the foundation of justice and peace in our societies and are best maintained on the one hand by an effective political democracy and on the other by a common understanding and observance of the human rights upon which they depend. Considering Recommendation Rec(2006)2rev of the Committee of Ministers to member States on the European Prison Rules and the European Court of Human Rights rulings, prisoners' human rights must be respected, ensuring that persons deprived of their liberty retain all rights that are not lawfully affected by the decision sentencing them to imprisonment or remanding them in custody. These principles, and the respect of human dignity, shall remain intact when AI and related technologies are used in prisons and by probation services.

Legal frameworks and policies should be established regarding the use of AI and related digital technologies also in the prison and probation services. The use of AI and related technologies in prison and by the probation services shall be governed by a clear legal framework in order to ensure legal certainty and accountability. The requirements of legal certainty and the protection from arbitrariness also flow from the European Convention on Human Rights (see Nuh Uzun and Others v. Türkiye, nos. 49341/18, 29 March 2022).

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.

Given that private companies participate in the AI lifecycle, for example by designing applications or providing data to feed systems, these entities should respect these principles, in view of the fact that that their products will be used in prisons and by the probations services and this use will impact human rights and life of those subjected to their use.

Public authorities should ensure that private sector applications respect these principles by requiring product audits or other compliance mechanisms. Efficient measures should be taken to ensure that civil and/or penal liability is put in place in case of causing intentional or unintentional harm by the use of AI and related digital technologies.

Social prejudices and stereotypes regarding a person or a group can lead to biases and 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. Therefore, AI and its use should be regularly monitored, and efficient and prompt measures should be taken to deal away with biases. Such biases may raise an issue of discrimination under Article 14 of the European Convention on Human Rights (see Basu v. Germany, no. 215/19, 18 October 2022).

AI and its use should strengthen the equality of treatment of persons and groups independent of their sex, gender, sexual orientation, race, colour, language, age, religion, political or other opinion, national or social origin, education, association with a minority, property, birth and state of health. Facial recognition systems and prisoner risk assessment tools have been criticised for this very reason, and their introduction in the prison context should be carefully considered, as discussed further in the text.

AI could deepen existing inequalities between individuals or groups of individuals and therefore in addition proactive measures should be taken to avoid such a danger, like providing digital and AI literacy, engaging stakeholders, examining the likely impact of intended data processing on the rights and fundamental freedoms, implementing human rights by design and privacy by design approaches, digital tools and employment opportunities.

AI has many different possibilities for design and functionality, so prior to implementing AI and related digital technologies, their use and impact should be discussed with the prison and probation management level to ensure that this specific AI tool is necessary, will be fit for the purpose, will improve the quality or efficiency of the prison or probation service and will support the strategical targets of the service. It is also important to evaluate whether this will be done through the least possible intrusion into the rights and private life of all those involved. Such a requirement of proportionality also arises from the perspective of Article 8 of the European Convention on Human Rights (see Van der Graaf v. the Netherlands (dec.), no. 8704/03, 1 June 2004 where the Court found that there has not been violation). Harm risk, security, protection and offender management should be key indicators in decision making. Proportionality in this context means that the interference with human rights created by the use of AI must be “necessary in a democratic society”. An interference will be considered “necessary in a democratic society” for a legitimate aim if it answers a “pressing social need” and, in particular, if it is proportionate to the legitimate aim pursued and if the reasons adduced by the national authorities to justify it are “relevant and sufficient” (see S. and Marper v. the United Kingdom [GC], nos. 30562/04 and 30566/04, 4 December 2008, para 101; and Glukhin v. Russia, no. 11519/20, 4 July 2023, para 78, concerning the use of facial recognition technology).

When the use of AI and related digital technologies provides new and unexpected information, the gathering of which was not intended or related to the original purpose, the use of this information must be done in accordance with law, if strictly necessary, proportionate, and legitimate. The person concerned should have an effective opportunity to challenge such use.

Good governance requires society to be informed and involved as far as possible in the decision regarding the use of AI and related digital technologies and regarding its process of designing and developing.

The information about design, operation and data processing methods should be non-opacite, accessible and understandable to the individuals using these technologies, 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 under their responsibility 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.

The commercial secret behind the design of an AI or related digital technologies should not be an obstacle to public scrutiny which in turn requires accessibility of the users and of those affected by their use in order to ensure traceability. This also requires a reasonably clear explanation of the logic of the algorithms used and of the outcomes reached.

In the context of use of AI and related digital technologies there are basically two scenarios: (1) the AI replaces the human decision-making, or (2) the AI assists the human decision-making. Both these cases are covered by this rule.

Staff may also be affected by the use of AI, for instance one such example is AI face recognition system which detect aggression or suicide attempts, or substance abuse.

There should be provisions for any decision taken by using to a varying extent AI and related digital technologies to be reviewed by a human to ensure respect for all other principles. It should also be possible to file a complaint against such a decision taken by the use of AI to ensure that the decision is taken by a human and that the human centred concerns are of primary importance.

As a minimum there should be provisions on access to an effective remedy before a competent authority (including judicial and data protection supervisory authorities); a 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).

International standards in the area of prison and probation services regarding complaint mechanisms flow from the European Prison Rules (Rule 70.1) and the European Rules on Community Sanctions and Measures (Rule 93). International standards on procedural safeguards and the right to an effective remedy flow from Articles 6 and 13 of the European Convention on Human Rights and Fundamental Freedoms.

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.

AI systems and other related digital technologies used in this field must be safe and reliable, and must have the necessary safeguards to ensure, on the one hand, that there is no misuse and, on the other, that they can be trusted and that the information cannot be accessed except in an authorised manner and also that in case of problems or incidents there are solutions provided so that the tool is not prevented from functioning.

For the good quality design and effective use of AI and related digital technologies, 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.

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 which lead to intelligent machines taking over core professional tasks including cognitive and affective tasks from staff. Examples of such risks are: 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.

Human contact is essential in the prison and probation work and should never be replaced simply because tasks can be done by a machine or because of lack of sufficient numbers of staff, but because this is the only way in which staff can be assisted in better achieving safety, security, good order or improve reintegration prospects of offenders. Staff should be reallocated and retrained to engage into more professional human contacts with offenders aiming at their resocialisation and at protecting society. There are some areas of the prison and probation system which are chronically understaffed, which erodes safety, security and reintegration prospects, so introduction of AI systems for repetitive everyday tasks which can be easily automated may be beneficial, but this should not undermine the regular positive human contacts with offenders.

Prison and probation staff should be consulted and engaged about the coming of AI and related digital technologies and the future shape of their work assisted by these tools. AI and digital 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 than basics as they are involved in decision taking.

The European Code of Ethics for Prison Staff[1] and the Council of Europe Probation Rules[2] contain detailed rules regarding professional ethics and staff training.

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, should be ensured. 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 and digital literacy should be actively promoted to both staff and offenders. Educating both staff and offenders to understand how AI-based processes are going to facilitate offender management in the future will deepen understanding of both key processes and AI and digital literacy. In this way AI can make offender management cycle faster, more cost-effective and optimize compatibility of services and offenders' needs. For example, in Finland all prisoners and probationers can access online basic course on AI from workstations placed in every unit (prisons and probation offices). Finland is also developing a new offender management tool RISE 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.

III. <u>Data Protection and Privacy</u>

Article 6 of the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (CETS 108+) states the following: "Personal data revealing racial origin, political opinions or religious or other beliefs, as well as personal data concerning health or sexual life, may not be processed automatically unless domestic law provides appropriate safeguards. The same shall apply to personal data relating to criminal convictions."

The European Court of Human Rights has held that there is no question that a prisoner forfeits his or her rights under the European Convention on Human Rights merely because of his or her status as a person detained following conviction. Indeed, prisoners in general continue to enjoy all the fundamental rights and freedoms guaranteed under the Convention save for the right to liberty, where lawfully

[1] CM/Rec (2012)5
[2] CM/Rec (2010)1

deprived. Any restrictions on these other rights – including the right to respect for private life and data protection under Article 8 – must be justified, although such justification may well be found in the considerations of security, in particular the prevention of crime and disorder, which inevitably flow from the circumstances of imprisonment (see Hirst v. the United Kingdom (no. 2) [GC], no. 74025/01, 6 October 2005, paras 69-70).

The use of AI 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, adequate and not excessive for the objective of the AI should be balanced against the need to minimise or limit access to and processing of data in order to respect private life of individuals as much as possible.

The use of any AI or related technologies in the field of prisons and probation must respect the standards laid down in the European Convention on Human Rights, 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[3] and Council Framework Decision 2008/977/JHA[4] must also be respected by the EU member States.

Individuals' right to human dignity has to be secured even when their personal data are processed and they must be informed of their data protection rights (right not to be subject to automated decision making, right to information and access to personal data, right to object, right to rectification and erasure, right to remedy and the right to benefit from the assistance of a data protection supervisory authority) and any limitations and contexts in which limitations may apply.

Proportionality of data processing and data protection principles and obligations 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 by national law. This also applies in the context of data protection. The European Court of Human Rights has held that the processing of personal data of prisoners must be done in accordance with the law, the processing of data must pursue a specific legitimate aim, and must be proportionate, namely necessary in order to achieve the aim pursued (Nuh Uzun and Others v. Türkiye, nos. 49341/18, 29 March 2022, para 83).

Prior to the use of AI for processing personal data, it is important to explicitly specify the legitimate and permitted purposes. The data processing shall be fair, lawful and proportionate in relation to the specified and legitimate purpose pursued and reflect at all stages of the processing a fair balance between all interests concerned, whether public or private, and the rights and freedoms at stake.

Considering the imbalance of power between prison and probation services and data subjects such as offenders and inmates, consent could not be considered, in principle, as an appropriate legal basis. However, where the processing of personal data in individual cases is based on consent as provided by Article 5(2) of Convention 108+, such consent should be obtained taking also into consideration the need to protect society.

Data controllers include any natural or legal person, public authority, service, agency or any other body which, alone or jointly with others, has decision-making power with respect to data processing.

[3] 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.

[4] 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 should be deleted or only preserved in a form that permits identification of an individual for no longer than it is necessary for the specific purpose for which the data are processed. Situations where attempts to store data for longer periods than allowed in case it becomes necessary in the future should be avoided. 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.

Security measures should take into account the current state of the art of data-security methods and techniques in the field of data processing. Their cost should be commensurate with the seriousness and probability of the potential risks. Security measures should be kept under review and updated where necessary.

Appropriate security measures include but are not limited to adopting and implementing policies and procedures to investigate and address security weaknesses and data breaches that may have adverse impacts for individuals and to report such incidents to individuals and data protection supervisory authorities.

Before collecting any personal data, there should be a clear definition of the purpose of its use, manner of its collecting, storing and processing, in order to avoid violation of human rights of the individual concerned. Such data should not be used for other purposes than its initially intended use or initially intended users. As far as possible after the collection, such data should be anonymised or pseudonymised in order to make these data unidentifiable.

It is essential that measures are adopted to ensure the accuracy of any personal data processed, and that inaccurate personal data can be corrected or deleted in an efficient and timely manner. Data quality must form part of a cycle of continuing assessment and evaluation.

Processing of certain types of data for the sensitive information it reveals, may lead to encroachments on interests, rights and freedoms. This can, for instance, be the case where there is a potential risk of discrimination or injury to an individual's dignity or physical integrity, where the data subject's most intimate sphere, such as his or her sex life or sexual orientation, is being affected.

It is also important to consider that once compromised (stolen for example) biometric data cannot be replaced. Therefore, processing of special categories of data should only be permitted where appropriate safeguards (which are adapted to the risks at stake and the interests, rights and freedoms to be protected), which complement the other protective provisions of Convention 108+, are provided for by law.

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 companies, 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.[5]

[5] Article 6 (Special categories of data), Convention 108+ stipulates the following: "Personal data revealing racial origin, political opinions or religious or other beliefs, as well as personal data concerning health or sexual life, may not be processed automatically unless domestic law provides appropriate safeguards. The same shall apply to personal data relating to criminal convictions."

IV. Use of AI and Related Digital Technologies

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

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.

When implementing digital or AI-driven surveillance applications, the prison administration shall define the levels of privacy regarding the use of rooms and areas which potentially may be subjected to such surveillance. Unless in a situation of acute danger of self-harm or violence against others, the general accommodation cells in prison should not be subjected to AI-driven or related digital technologies surveillance.

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-centred 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 undue 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. A similar example is the use of AI in gathering intelligence from other forms of digital communication by offenders, for example e-mails and electronic requests in cell-device systems.

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 human rights (see the European Court of Human Rights, Gorlov and Others v. Russia, nos. 27057/06 et al, 2 July 2019). 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 prison and probation services must be consulted in order to identify and evaluate their needs regarding safety, security and good order tasks and how to best assist them in executing these tasks by using AI and related digital technologies. The procedure for consulting them depends on the national law and practice.

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. 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.

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 related digital technologies. Moreover, the requirement of proportionality under Article 8 of the European Convention on Human Rights must be observed (see, for instance, Aycaguer v. France, no. 8806/12, 22 June 2017).

During probation, electronic monitoring systems can use AI techniques to facilitate the management of supervision of offenders and the decision making. 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 to ensure the reviewing of 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.

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.

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

AI tools have already been used to some extent also in the offender management systems (OMS) and processes in some jurisdictions. For example, AI tools are already 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 each individual case is different. AI should not replace humans in decision-making processes, but should work as a tool for humans, 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 compared to data collected with traditional methods. AI's role in the offender management systems (OMS) should be advisory and evidence based. The operation of the system should also be subject to the requirements of legality and protection from abuse as required under Article 8 of the European Convention on Human Rights (see Nuh Uzun and Others v. Türkiye, nos. 49341/18, 29 March 2022).

Experts should be well acquainted with both AI and criminological research in order to develop reliable and valid AI for the use by offender management systems (OMS). Prison and probation services are dealing with already stigmatized and vulnerable persons in the majority of the cases. There is a risk for stereotypical, discriminative, and ex post facto type of conclusions that can be repeated by AI if this fact is not taken into consideration in the algorithms used. AI algorithms can easily be biased and can start to repeat the same mistakes humans make. Designing an algorithm for use in the prison and probation context requires being exact about what we want to achieve and understanding the typical biases in human thinking which AI is supposed to be replacing (Fry, 2018). At best algorithms could overcome the harmful effects of cognitive biases (Sunstein, 2018) instead of repeating them.

The first and still most common applications of AI technology are various 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). Many jurisdictions have developed standardized instruments for risk and needs assessment based on this model during the last 20 years in offender management (Raynor, 2019). A recent project in the Finnish Prison and Probation Service is developing an AI application, named RISE AI, for offender management. RISE AI will be a recommender system that recommends rehabilitative services to offenders during their sentences based on the available offenders' background information. 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).

Risk assessments and especially AI based risk assessments should be regarded as dynamic processes rather than as final statements. They should be reviewed in due time and adjusted in accordance with the developments and changes that have taken place in the subject's life, behaviour, abilities, social relations, insights etc. Their use should not be limited to justifying restrictions but rather should aim to identify the procedures needed to reduce the detected risks and to develop effective plans for support and care.

To prevent discrimination and bias, it should be explicitly defined which criteria of personal data are relevant and necessary to determine the individual risk in question. While the criminal record, sex and age might be criteria that have a high value to determine an individual's risk of e.g. recidivism, other sensitive criteria mentioned in Article 14 of the European Convention on Human Rights [Prohibition of discrimination] like "race, colour, language, religion, political or other opinion, national or social origin, association with a national minority, property, birth or other status" or in Article 6 of the Convention 108+ [Special Categories of Data] bear a high risk for biased results; as a rule they can be avoided by criteria that have a closer link to an individual's learning and behaviour.

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. However, each purpose requires its own assessments and procedures, and the risk assessment is not the only assessment on which to base decisions in the justice system and corrections. The decisions deriving from such use should not be automated but should be taken by professionals. The aim of using AI in this way is to improve decision-making related to finding the best trajectory for the offenders regarding their needs and minimizing their risks. This purpose requires the use of various information, not only information regarding risk level of offenders.

AI can be used in various treatment, educational and training platforms, systems and procedures. 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). The Hong Kong Prison department is also actively developing AI technologies for offenders' self-management in order to enhance the efficiency of penal operations and the effectiveness of rehabilitation programmes (Houser, 2019).

The use of robotic systems for rehabilitative tasks is another example. The possibility to use AI to address the solitary confinement problems by employing digital assistants, similar to Amazon's Alexa, as a form of 'confinement companions' for prisoners has been discussed in the US. Even if these 'companions' could alleviate some of the psychological stress of 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.

There are also concerns whether occupations can disappear while AI is taking over the job humans used to do in a faster and more accurate way. AI can assist rehabilitative processes, and programs or individual therapeutic work can include AI-based methods like VR or robotics, but it should be stressed in this respect that no rehabilitative work should be solely based on AI without human in the process. AI can also bring new occupations. One such example is shown in a pilot in Finnish prisons, where prisoners are 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.

Rule 25, Recommendation CM/Rec (2018)5 concerning children with imprisoned parents lists the different information and communication options for prisoners to maintain contacts with their child remotely while stressing that such options should never replace the face-to-face contacts.

An example of this is the use of chatbots as virtual assistants in health care. Chatbots can offer preliminary information, guidance and suggestions to the patients. However, their role is only advisory and can't replace the care, treatment and decisions done by health care professionals, who are responsible for the consequences of all procedures suggested and done to the patients.

AI offers the possibility to alleviate the routine tasks staff are responsible for on an everyday basis like assisting offenders in escorting them to visiting areas and in establishing contacts with lawyers, possible employers, psychologists, social workers and other professionals as well as with their families. This is not only the case in prisons but also for probationers. Nevertheless, the use of AI and related digital technologies in such cases should be done carefully because of different reasons: language and technological inabilities; psychological difficulties; young or old age and other.

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

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 instead of benefiting only organisational, material and financial purposes.

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.

After implementation of the predictive machine learning models, their predictions 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.

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.

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.

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.

When using AI and related digital technologies in the field of human resources management such as selection, recruitment and professional development, it is important to consider the principle of quality and the principle of equality and non-discrimination. The main reason for this is that the nature of AI and related digital technologies can use complex and non-transparent internal algorithms for the decisions in the process. If there exist biases in the training data of AI, the trained algorithm will likely exhibit those biases in its predictions. Therefore, to the extent possible, algorithms should be explainable, meaning that its reasoning should be transparent to human observers and decision-makers. It is paramount to train algorithms on data that is representative and of the highest possible quality. This relates to the principle of good governance, transparency and traceability. This creates the possibility that a person can be informed of the reasons for the decisions being made.

This also should involve the possibility to request a revision by a human professional of any decision taken regarding filtering of employment requests or of requests for professional training or development.

VI. Research, Development, Evaluation and Regular Revision

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. Machine learning systems still need ongoing evaluation and revision by humans.

AI systems should be continuously evaluated and studied in order to ensure 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.

Risk management and mitigation frameworks set up in previous phases should be evaluated, adapted and maintained during the deployment phase. Especially should focus be on how to identify risks in the particular solution, assess the impact on the change processes and expectance of benefits and from a governance perspective what is needed to regulate either with legislative or regulative policies.

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.

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.