Parliamentary Assembly
Assemblée parlementaire
http://assembly.coe.int

COUNCIL OF EUROPE
CONSEIL DE L'EUROPE

Provisional version

Committee on Legal Affairs and Human Rights

Justice by algorithm – the role of artificial intelligence in policing and criminal justice systems

Report*

Rapporteur: Mr Boriss CILEVIČS, Latvia, Socialists, Democrats and Greens Group

A. Draft resolution

1. Artificial intelligence (AI) applications can now be found in many spheres of human activity, from pharmaceutical research to social media, agriculture to on-line shopping, medical diagnosis to finance, and musical composition to criminal justice. They are increasingly powerful and influential, and the public is often unaware of when, where and how they are being used.

2. The criminal justice system represents one of the key areas of the state's responsibilities, ensuring public order and preventing violations of various fundamental rights by detecting, investigating, prosecuting and punishing criminal offences. It gives the authorities significant intrusive and coercive powers including surveillance, arrest, search and seizure, detention, and the use of physical and even lethal force. It is no accident that international human rights law requires judicial oversight of all of these powers: effective, independent, impartial scrutiny of the authorities' exercise of criminal law powers with the potential to interfere profoundly with fundamental human rights. The introduction of non-human elements into decision-making within the criminal justice system may thus create particular risks.

3. If the public is to accept the use of AI and enjoy the potential benefits that AI can bring, it must have confidence that any risks are being properly managed. If AI is to be introduced with the public's informed consent, as one would expect in a democracy, then effective, proportionate regulation is a necessary condition.

4. Regulation of AI, whether voluntary self-regulation or mandatory legal regulation, should be based on universally accepted and applicable core ethical principles. The Assembly considers that these principles can be grouped under the following broad headings:

4.1. Transparency, including accessibility and explicability;

4.2. Justice and fairness, including non-discrimination;

4.3. Human responsibility for decisions, including liability and the availability of remedies;

4.4. Safety and security;

4.5. Privacy and data protection.

5. The Assembly welcomes Committee of Ministers' Recommendation Rec/CM(2020)1 on the human rights impact of algorithmic systems, along with its accompanying guidelines on addressing the human rights impacts of algorithmic systems, and the recommendation of the Council of Europe Commissioner for Human Rights on 'Unboxing Artificial Intelligence: 10 steps to protect Human Rights'. It endorses the general proposals made in these texts for application also in the area of policing and criminal justice systems.

* Draft resolution and draft recommendation unanimously adopted by the committee on 9 September 2020.

6. The Assembly notes that a large number of applications of AI for use by the police and criminal justice systems have been developed around the world. Some of these have been used or their introduction is being considered in Council of Europe member States. They include facial recognition, predictive policing, the identification of potential victims of crime, risk assessment in decision-making on remand, sentencing and parole, and identification of 'cold cases' that could now be solved using modern forensic technology.

7. The Assembly finds that there are many ways in which the use of AI in policing and criminal justice systems may be inconsistent with the above-mentioned core ethical principles. Of particular concern are the following:

7.1. AI systems may be provided by private companies, which may rely on their intellectual property rights to deny access to the source code. The company may even acquire ownership of data being processed by the system, to the detriment of the public body that employs its services. The users and subjects of a system may not be given the information or explanations necessary to have a basic understanding of its operation. Certain processes involved in the operation of an AI system may not be fully penetrable to human understanding. Such considerations raise transparency (and, as a result, responsibility/ accountability) issues.

7.2. AI systems are trained on massive datasets, which may be tainted by historical bias, including through indirect correlation between certain predictor variables and discriminatory practices (such as postcode being a proxy identifier for an ethnic community historically subject to discriminatory treatment). This is a particular concern in relation to policing and criminal justice, because of both the prevalence of discrimination on various grounds in this context and the significance of decisions that may be taken. The apparent mechanical objectivity of AI may obscure this bias ("techwashing"), reinforce and even perpetuate it. Certain AI techniques may not be readily amenable to challenge by subjects of their application. Such considerations raise issues of justice and fairness.

7.3. Resource constraints, time pressure, lack of understanding, and deference to or reluctance to deviate from the recommendations of an AI system may lead police officers and judges to become overly reliant on such systems, in effect abdicating their professional responsibilities. Such considerations raise issues of responsibility for decision-making.

7.4. These considerations also affect one another. Lack of transparency in an AI application reduces the ability of human users to take fully informed decisions. Lack of transparency and uncertain human responsibility undermine the ability of oversight and remedial mechanisms to ensure justice and fairness.

7.5. The application of AI systems in separate but related contexts, especially by different agencies relying sequentially on one another's work, may have unexpected, even unforeseeable cumulative impacts.

7.6. The addition of AI-based elements to existing technology may also have consequences of unforeseen or unintended gravity.

8. The Assembly concludes that, whilst the use of AI in policing and criminal justice systems may have significant benefits if it is properly regulated, it risks having a particularly serious impact on human rights if it is not.

9. The Assembly therefore calls upon member States, in the context of policing and criminal justice systems, to:

9.1. Adopt a national legal framework to regulate the use of AI, based on the core ethical principles mentioned above;

9.2. Maintain a register of all AI applications in use in the public sector and refer to this when considering new applications, so as to identify and evaluate possible cumulative impacts;

9.3. Ensure that AI serves overall policy goals, and that policy goals are not limited to areas where AI can be applied;

9.4. Ensure that there is a sufficient legal basis for every AI application and for the processing of the relevant data;

9.5. Ensure that all public bodies implementing AI applications have internal expertise able to evaluate and advise on the introduction, operation and impact of such systems;

9.6. Meaningfully consult the public, including civil society organisations and community representatives, before introducing AI applications;

9.7. Ensure that every new application of AI is justified, its purpose specified and its effectiveness confirmed before being brought into operation, taking into account the particular operational context;

9.8. Conduct initial and periodic, transparent human rights impact assessments of AI applications, to assess, amongst other things, privacy and data protection issues, risks of bias/ discrimination and the consequences for the individual of decisions based on the AI's operation, with particular attention to the situation of minorities and vulnerable and disadvantaged groups;

9.9. Ensure that the essential decision-making processes of AI applications are explicable to their users and those affected by their operation;

9.10. Only implement AI applications that can be scrutinised and tested from within the place of operation;

9.11. Carefully consider the possible consequences of adding AI-based elements to existing technologies;

9.12. Establish effective, independent ethical oversight mechanisms for the introduction and operation of AI systems;

9.13. Ensure that the introduction, operation and use of AI applications can be subject to effective judicial review.

B. Draft recommendation

1. The Assembly refers to its Resolution … (20…) entitled 'justice by algorithm – the role of artificial intelligence in policing and criminal justice systems'. It notes that this resolution was adopted as work was ongoing within the Council of Europe by the Ad hoc Committee on artificial intelligence (CAHAI).

2. The Assembly recalls that all Council of Europe member States are subject to the same basic standards of human rights and the rule of law, notably those established by the European Convention on Human Rights, as interpreted by the caselaw of the European Court of Human Rights. It considers that regulatory patchworks – varying standards in different countries – may give rise to 'ethics shopping', resulting in the relocation of AI development and use to regions with lower ethical standards.

3. The Assembly therefore calls on the Committee of Ministers to take into account the particularly serious potential impact on human rights of the use of artificial intelligence in policing and criminal justice systems when assessing the necessity and feasibility of a European legal framework for artificial intelligence.

Explanatory memorandum by Mr Cilevičs, rapporteur

1. Introduction

1. The motion underlying this report, which I tabled on 26 September 2018, was referred to the Committee by the Bureau on 12 October 2018, following which the Committee appointed me as rapporteur on 21 January 2019.[1] The Committee held a hearing with experts at its meeting in Berlin, Germany on 14-15 November 2019, with the participation of Dr Michael Veale, Lecturer in Digital Rights & Regulation, University College London, United Kingdom, and Ms Marion Oswald, Vice-Chancellor's Senior Fellow in Law, University of Northumbria, United Kingdom. A fact-finding visit to the West Midland Police, United Kingdom was cancelled on account of the Covid-19 pandemic but was replaced by video-conferences with Tom McNeil of the West Midlands Police and Crime Commissioner's Ethics Committee, and Chief Superintendent Chris Todd and Detective Chief Inspector Nick Dale of the West Midlands Police. I would like to thank all of those concerned for their contributions to this report.

2. Artificial intelligence (AI) is no longer the stuff of science fiction, even if it has not yet fulfilled all of science fiction's predictions. We do not have sentient machines capable of matching or even outperforming human beings across multiple domains (what is known as 'general' or 'strong' AI). We do, however, have systems that are capable of performing specific tasks, such as recognising patterns or categories, or predicting behaviour, with a certain degree of what might be called 'autonomy' ('narrow' or 'weak' AI). These systems can be found in very many spheres of human activity, from pharmaceutical research to social media, agriculture to on-line shopping, medical diagnosis to finance, and musical composition to criminal justice. They are increasingly powerful and influential, and the public is often unaware of when, where and how they are being used. Indeed, the 'better' they become, the less apparent they may be: the research company OpenAI recently announced that it would refrain from releasing a new natural (i.e. human) language processing AI system, as it was capable of producing text that was indistinguishable from human-generated text, and could too easily be abused.

3. As noted in the motion, the criminal justice system represents one of the key areas of the state's responsibilities, ensuring public order and preventing violations of various fundamental rights by detecting, investigating, prosecuting and punishing criminal offences. It gives the authorities significant intrusive and coercive powers including surveillance, arrest, search and seizure, detention, and use of physical and even lethal force. It is no accident that international human rights law requires judicial oversight of all of these powers: effective, independent, impartial scrutiny of the authorities' exercise of criminal law powers with the potential to interfere profoundly with fundamental human rights. The introduction of non-human elements into decision-making within the criminal justice system may thus create particular risks.

4. The Assembly has already touched upon some of the issues relevant to the present report in its Recommendation 2102 (2017) on technological convergence, artificial intelligence and human rights. Recommendation 2102 noted that "it is increasingly difficult for lawmakers to adapt to the speed at which science and technologies evolve and to draw up the required regulations and standards". The Assembly concluded that "safeguarding human dignity in the 21st century implies developing new forms of governance, new forms of open, informed and adversarial public debate, new legislative mechanisms and above all the establishing of international co-operation making it possible to address these new challenges most effectively."

5. Although this is not the first time that the Committee on Legal Affairs and Human Rights has considered AI, I will first explore some basic and general issues, building on the work of the Committee on Culture, Science, Education and Media when preparing Recommendation 2102, before looking at the specific case of the use of artificial intelligence and algorithms in criminal justice systems.

1.1. Key concepts

6. The expression 'artificial intelligence' was first coined in 1955 by John McCarthy (and others) of Dartmouth College, New Hampshire, in a proposal for a research project based on the working hypothesis that "every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it". Unfortunately, there is still no universally agreed definition of AI, and much disagreement on how it should be defined.[2] A central aspect of this problem is the lack of a common

[1] Reference no. 4407.

[2] For example, one definition, elaborated specifically for the purpose of assessing the need for regulation of AI, is that "Artificial Intelligence Is the Ability of a Non-natural Entity to Make Choices by an Evaluative Process" (*Robot Rules: Regulating Artificial Intelligence,* Jacob Turner, Palgrave Macmillan, 2018). Alternatively, AI can be said to exist "when

definition of human intelligence – assuming AI is to be roughly understood as simulating all or certain characteristics of human intelligence, which by definition would not be the case for any future form of 'superintelligent' general/ strong AI.[3] For a general description of AI, please see the appendix to this report..

7. It is sometimes said that a defining characteristic of AI is its 'autonomy'. Great care should be taken with this concept, however, as it may have serious implications concerning accountability and responsibility, up to and including the question of whether or not AI should be considered as a moral agent or even a legal person. This has been well described by the European Group on Ethics in Science and New Technologies. "The term 'autonomy' stems from philosophy and refers to the capacity of human persons to legislate for themselves, to formulate, think and choose norms, rules and laws for themselves to follow… Autonomy in the ethically relevant sense of the word can therefore only be attributed to human beings. It is therefore somewhat of a misnomer to apply the term 'autonomy' to mere artefacts, albeit very advanced complex adaptive or even 'intelligent' systems… Since no artefact or system – however advanced and sophisticated – can in and by itself be called 'autonomous' in the original ethical sense, they cannot be accorded the moral standing of the human person and inherit human dignity… Human beings ought to be able to determine which values are served by technology, what is morally relevant and which final goals and conceptions of the good are worthy to be pursued. This cannot be left to machines, no matter how powerful they are."[4]

1.2. Opportunities and risks in the use of AI – some general considerations

8. It has been said that AI "can redefine work or improve work conditions for humans and reduce the need for human contribution, input and interference during operation. It can help to assist or replace humans with smart technology in difficult, dirty, dull or dangerous work, and even beyond."[5] Indeed, as noted in paragraph 2, AI is already being applied in numerous areas, most often with positive results.

9. The potential power of AI also carries risks. Its speed, complexity and scalability mean that it vastly outperforms human beings at certain tasks. The potential inscrutability of self-generated algorithms means that the method and reasoning employed to produce a particular output may be unknowable, even to the AI's developer. It has been suggested that "AI/AS will be performing tasks that are far more complex and impactful than prior generations of technology, particularly with systems that interact with the physical world, thus raising the potential level of harm that such a system could cause."[6] This argument can be developed further: since "advanced AI could represent a profound change in the history of life on Earth, [it] should be planned for and managed with commensurate care and resources."[7]

1.3. Regulation of AI – some general considerations

10. A huge amount has been written on whether, when and how to regulate artificial intelligence. Some commentators consider that regulation is undesirable, since it would stifle innovation or induce 'ethics shopping' by AI companies; premature, since the technology is still evolving; and even impossible, due to the intrinsic nature of AI. When considering these issues, it is worth first reflecting on past experience, in particular the history of Internet regulation.

11. In 1996, as Internet use was beginning to spread and the giant Internet companies of today were still nascent or yet to be established, John Perry Barlow presented a "Declaration of the Independence of Cyberspace" to the World Economic Forum in Davos, Switzerland. "Governments of the Industrial World, you weary giants of flesh and steel, I come from Cyberspace, the new home of Mind. On behalf of the future, I ask of the past to leave us alone. You are not welcome among us. You have no sovereignty where we gather. […] I declare the global social space we are building to be naturally independent of the tyrannies you seek to impose on us. You have no moral right to rule us nor do you possess any methods of enforcement we have true reason to fear."

---

behaviour comes not purely from the programmer, but some other means, e.g. knowledge bases" (Arvind Narayanan, Princeton University). These definitions may be different, but neither is 'wrong'.

[3] AI has also been defined as something quite distinct from human intelligence: "AI is the continuation of intelligence by other means… It is thanks to this decoupling that AI can colonise tasks whenever this can be achieved without understanding, awareness, sensitivity, hunches, experience or even wisdom. In short, it is precisely when we stop trying to reproduce human intelligence that we can successfully replace it." ("A Fallacy that Will Hinder Advances in Artificial Intelligence", Professor Luciano Floridi, Financial Times, 1 June 2017.) This description highlights a characteristic of AI that is particularly relevant to its potential application in the criminal justice system.

[4] "Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems", European Group on Ethics and Science and New Technologies, European Commission, March 2018.

[5] European Group on Ethics and Science and New Technologies, op. cit.

[6] "General Principles", IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems.

[7] "Asilomar AI Principles", 2017 Asilomar Conference.

12. Although this Declaration was primarily intended to refer to individuals, its cosmopolitan, libertarian attitude has been typical of many Internet companies ever since. Until recently, these companies' behaviour was not seriously challenged by national authorities, who failed or refused to recognise its consequences. As Professor Paul Nemitz has said, "this history of failure to assign and assume responsibility in the Internet age, both by legislators and by tech corporations, [...] led to the fiascos of the Internet, in the form of spreading of mass surveillance, recruitment to terrorism, incitement to racial and religious hate and violence as well as multiple other catastrophes for democracy, the latest being the Cambridge Analytica scandal and the rise of populists, often the most sophisticated beneficiaries of the assistance of Facebook, YouTube, Twitter and co., combining the advertising and network techniques of targeted advertising developed for profit with political propaganda."[8]

13. Apart from now having past experience to learn from, the question of AI regulation can also be distinguished by its state of development. Turning again to Professor Nemitz: "AI is now – in contrast with the Internet – from the outset not an infant innovation brought forward mainly by academics and idealists, but largely developed and deployed under the control of the most powerful Internet technology corporations." In other words, we already know enough about AI in practice to regulate its application – and we also know enough about the companies that are using it to ask whether mandatory, enforceable regulation, as opposed to voluntary, ethics-based self-regulation, may be necessary.

1.4. Regulation and public trust

14. If the public is to accept the use of AI and enjoy the potential benefits it can bring, it must have confidence that any risks are being properly managed. Two leading researchers in the field have noted that "We know that there is no 'formula' for building trust, but we know from experience that technology is, in general, trusted if it brings benefits and is safe and well regulated."[9] Unless AI is to be unwittingly or forcibly imposed on the general public – in other words, if it is to be introduced with the public's consent – then effective, proportionate regulation is a necessary, although not sufficient, condition.

1.5. Regulatory coherency/ harmonisation

15. If regulation is to be introduced, it should be coherent and harmonised on the widest possible scale. The European Group on Ethics in Science and New Technologies, for example, has drawn attention to "the risks inherent to uncoordinated, unbalanced approaches in the regulation of AI and 'autonomous' technologies. Regulatory patchworks may give rise to 'ethics shopping', resulting in the relocation of AI development and use to regions with lower ethical standards. Allowing the debate to be dominated by certain regions, disciplines, demographics or industry actors risks excluding a wider set of societal interests and perspectives."[10]

1.6. Ethical governance

16. Even those who consider that it may be premature to introduce legal regulation of AI would agree that at a minimum, ethical governance is needed. This has been defined as "a set of processes, procedures, cultures and values designed to ensure the highest standards of behaviour. Ethical governance thus goes beyond simply good (i.e. effective) governance, in that it inculcates ethical behaviours in both individual designers and the organizations in which they work."[11]

1.7. Ethical principles

17. Identification of ethical principles is necessary whether they are to form the basis of ethical 'codes of conduct' alone, or also to inform the establishment of legal regulations. Research has shown that there is extensive agreement on the core content of ethical principles that should be applied to AI systems, notably the following:[12]

[8] "Constitutional democracy and technology in the age of artificial intelligence", Phil. Trans. R. Soc. A 378:20180089.
[9] "Ethical governance is essential to building trust in robotics and artificial intelligence systems", Alan F.T. Winfield and Marina Jirotka, Phil. Trans. R. Soc A 376:20180085.
[10] European Group on Ethics and Science and New Technologies, op. cit.
[11] Winfield and Jirotka, op. cit.
[12] See *AI Ethics Guidelines: European and Global Perspectives,* Draft Report commissioned by the Council of Europe Ad Hoc Committee on Artificial Intelligence (CAHAI), Ienca & Vayena, March 2020. Likewise, Winfield and Jirotka (op. cit.) observed that "An informal survey at the end of 2017 discovered that a total of 10 different sets of ethical principles […] had been proposed by December 2017, seven of which appeared in 2017… There is a good deal of commonality across these principles, notably that IAS should (i) do no harm, (ii) respect human rights and freedoms, including dignity and

- *Transparency.* The principle of transparency can be interpreted widely to include accessibility, explainability and explicability of an AI system, in other words the possibilities for an individual to understand how the system works and how it produces its results.
- *Justice and fairness.* This principle includes non-discrimination, impartiality, consistency and respect for diversity and plurality. It further implies the possibility for the subject of an AI system's operation to challenge the results, with the possibility of remedy and redress.
- *Responsibility.* This principle encompasses the requirement that a human being should be responsible for any decision affecting individual rights and freedoms, with defined accountability and legal liability for those decisions. This principle is thus closely related to that of justice and fairness.
- *Safety and security.* This implies that AI systems should be robust, secure against outside interference and safe against performing unintended actions, in accordance with the precautionary principle.
- *Privacy.* Whilst respect for human rights generally might be considered inherent in the principles of justice and fairness and of safety and security, the right to privacy is particularly important wherever an AI system is processing personal or private data. AI systems must therefore respect the binding standards of the EU General Data Protection Regulation (GDPR) and the Council of Europe's data protection convention 108 (and the 'modernised' convention 108+), as applicable.

18. In December 2018 the Council of Europe's European Commission for the Efficiency of Justice (CEPEJ) adopted a European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and their environment, which shows how key elements from the list above apply in this context. CEPEJ's European Ethical Charter sets out five principles: respect for fundamental rights; non-discrimination; quality and security; transparency, impartiality and fairness (including accessibility and understandability – analogous to explainability); and 'under user control'.

1.8. Ethical concerns in relation to the use of AI in criminal justice systems

19. Ms Oswald and others have conducted a detailed review of the operation of HART (an AI-based system used by the Durham Constabulary – see further below), making a number of observations and drawing conclusions of more general applicability.[13] The impact of AI must be considered in its operational context, which in the case of the police includes routines, objectives and decision-making processes. The actions of the police, as public authorities, may be subject to judicial review, but AI does not design an algorithm with the aim of its being amenable to human comprehension. The Durham Constabulary deliberately selected a variant of the HART algorithm that favoured "cautious errors", a trade-off between (more) high-risk false positives and (fewer) false negatives ("dangerous errors"). "Whether the overall benefit to society from a particular value-judgement built into an algorithm justifies the possible negative consequences to single individuals may depend to a large extent on the seriousness of those consequences." The HART algorithm is also fed data from the Durham Constabulary only, not from other local agencies or other police forces. Human decision-makers would have access to more varied sources of information, which is why an algorithm should never have the final say. That said, Ms Oswald and her colleagues could not exclude the possibility that some officers might "(consciously or otherwise) prefer to abdicate responsibility for what are risky decisions to the algorithm, resulting in deskilling and 'judgmental atrophy'." Such a situation would also have negative consequences over time. A human decision-maker can adapt rapidly to a changing context, unlike an algorithm, which would therefore need "careful and constant scrutiny of the predictors used and frequently refreshing of the algorithm with more recent historical data."

20. During our Committee hearing, Dr Veale addressed the use of AI by the police in the wider social and political context. He noted that cuts in social care services, for example, could generate problems that required subsequent intervention by other agencies, including the police, whose own resources were also stretched. AI was presented as having cost-saving advantages and so policing policy could prioritise AI-based approaches. AI developers may seek to avoid discriminatory or other undesirable outcomes but they do not have the wider social context in mind. The solutions they offer thus define the parameters of the problem according to what is technically possible in response. This would amount to privatisation of policy-making and could exacerbate social divisions and reinforce injustices. The public sector, including the police, therefore needed its own, internal expertise in order to assess the actual usefulness of AI-based solutions and be ready to say "no".

---

privacy, while promoting well-being, and (iii) be transparent and dependable while ensuring that the locus of responsibility and accountability remains with their human designers or operators." They also explain how ethics inform standards, which in turn form the basis of regulations.

[13] "Algorithmic risk assessment policing models: lessons from the Durham HART model and 'Experimental' proportionality", Oswald, Grace, Urwin and Barnes, 2018.

21. One of Ms Oswald's colleagues on the West Midlands Police and Crime Commissioner's Ethics Committee (see further below), Jamie Grace, notes that in the UK, "As David Lyon predicted in 2007, the 'safety state' has been pushing out the welfare state from public discourse and public policy, and the 'safety state', concerned with public protection above any consideration of individual autonomy, 'depends extensively on surveillance data'." Mr Grace is of the view that AI will "exacerbate and inflame human rights tensions that already exist in criminal justice settings. These inherent human rights issues include privacy concerns, the chilling of freedom of expression, problems around potential for racial discrimination, and the rights of victims of crime to be treated with dignity."[14]

22. Since I presented my introductory memorandum in April 2019, there has been a widespread, growing backlash against the use of artificial intelligence in policing and criminal justice systems. I will examine certain specific situations further below, but at this point, it is worth considering the main points of a statement issued by a number of senior US researchers and academics in July 2019:

> "Actuarial pretrial risk assessments suffer from serious technical flaws that undermine their accuracy, validity, and effectiveness. They do not accurately measure the risks that judges are required by law to consider. When predicting flight and danger, many tools use inexact and overly broad definitions of those risks. When predicting violence, no tool available today can adequately distinguish one person's risk of violence from another. Misleading risk labels hide the uncertainty of these high-stakes predictions and can lead judges to overestimate the risk and prevalence of pretrial violence. To generate predictions, risk assessments rely on deeply flawed data, such as historical records of arrests, charges, convictions, and sentences. This data is neither a reliable nor a neutral measure of underlying criminal activity. Decades of research have shown that, for the same conduct, African-American and Latinx people are more likely to be arrested, prosecuted, convicted and sentenced to harsher punishments than their white counterparts. Risk assessments that incorporate this distorted data will produce distorted results. These problems cannot be resolved with technical fixes. We strongly recommend turning to other reforms."[15]

2. Real-world applications of algorithms and artificial intelligence in criminal justice systems

23. In this section, I will mainly focus on three examples of the use of AI in criminal justice systems: PredPol, which predicts where crimes may occur and on that basis calculates how best to allocate police resources; HART (Harm Assessment Risk Tool), which predicts the risk of reoffending when deciding whether or not to prosecute; and COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), which also (mainly) predicts whether or not an individual will reoffend. All three are examples of 'black boxes' – proprietary systems whose inner workings are not publicly accessible.

2.1. Predicting crime and allocating police resources – PredPol

24. PredPol, a Californian company that grew out of a project between UCLA and the Los Angeles Police Department, defines "predictive policing [as] the practice of identifying the times and locations where specific crimes are most likely to occur, then patrolling those areas to prevent those crimes from occurring." PredPol uses a client police department's historical data from a two- to five-year period to train a machine-learning algorithm, which is subsequently updated on a daily basis. Only three data-points are used: crime type, location and date/ time. According to PredPol, "No demographic, ethnic or socio-economic information is ever used. This eliminates the possibility for privacy or civil rights violations seen with other intelligence-led or predictive policing models." This latter claim is, however, disputed.

25. This technology does not come cheap. Kent Police in the United Kingdom used PredPol from December 2012 until March 2018, at a cost of £100,000 per year. It was reported that during an initial four-month trial, the use of PredPol resulted in a 6% decrease in street crime.[16] Ultimately, however, Kent Police stated that whilst "Predpol had a good record of predicting where crimes are likely to take place, what is more challenging is to show that we have been able to reduce crime with that information." Kent Police was nevertheless sufficiently impressed by the technology's potential that it now intends to develop its own system.[17]

26. Despite PredPol's claim that there is no possibility of human rights violations, the system has been criticised for perpetuating historical bias in policing practice, whilst at the same time concealing that bias behind a veneer or presumption of mechanical neutrality – known as 'tech-washing'. Even if personal or socio-

[14] "Machine learning technologies and their inherent human rights issues in criminal justice contexts", November 2019.
[15] "Technical Flaws of Pretrial Risk Assessments Raise Grave Concerns", July 2019.
[16] "Predictive Policing statistics from Kent Police", Kent Online, 13 February 2018.
[17] "Kent Police stop using crime predicting software", The Telegraph, 27 November 2018.

economic information is not included in the training dataset, it may nevertheless be inherent in the data, especially relating to the location of offences. If in the past a police force has disproportionately focused its attention on a certain neighbourhood, then crimes committed in that neighbourhood would have been more likely to be detected. This would artificially skew the historical dataset used to train the PredPol algorithm. As a result, PredPol would predict a greater probability of crimes occurring there in future. More resources would then be allocated to policing that area, perpetuating (and, indeed, reinforcing) the historical bias, only now on a purportedly 'objective' basis. If the neighbourhood in question was predominantly inhabited by people of a certain ethnicity or religion – which may have been the very reason why the local police force had historically focused its attention there ('ethnic profiling') – the result could be discriminatory on grounds that are prohibited under international human rights law, including article 14 of the European Convention on Human Rights.

27. Other United Kingdom police forces are also now experimenting with another form of predictive policing. Nine forces, led by West Midlands Police and including also London's Metropolitan Police and Greater Manchester Police, are developing the National Data Analytics Solution (NDAS – see further below). This will use a combination of machine learning AI and statistics to assess, for example, the risk of someone committing or becoming a victim of gun or knife crime, as well as the likelihood of someone falling victim to modern slavery. The system is based on data relating to around five million individuals, from which it identified almost 1400 indicators that could predict crime, of which 30 are particularly significant. West Midlands Police will work with the UK's data protection authority to ensure the system complies with privacy regulations. The officer responsible for the project has acknowledged that it is partly a response to significant cuts to police budgets in recent years and the resulting pressure to prioritise attention on persons who need interventions most urgently.[18]

28. I will examine the ethical regulatory framework that has been developed around the West Midlands Police AI-related projects in greater detail below.

2.2. Predicting reoffending and preventing recidivism – HART

29. The HART system was developed by Durham Police, working with researchers from Cambridge University, using training data from 104,000 persons who had been arrested over five years. It uses 'predictor values', most of which focus on the suspect's offending history, as well as age, gender and geographical area, to categorise an offender as being at low-, medium- or high-risk of committing new serious offences over the following two years. Those in the medium-risk category – those "likely to commit a non-serious offence" – may then be included in the police force's 'Checkpoint' programme, which "offers eligible offenders a 4-month long contract to engage as an alternative to prosecution. The contract offers interventions to address the underlying reasons why they committed the crime to prevent them from doing it again". HART was developed with the aim of reducing the number of people who are incarcerated despite being susceptible to other forms of intervention that would be as or even more effective at reducing the risk that they reoffend.

30. The lead police officer in the HART project, along with academics and an Australian police officer, have published a detailed article analysing HART, taking "the necessity, proportionality and foreseeability principles set out in European human rights law [as] a starting point".[19] Recognising that AI represents "a different type of decision-making, not an enhanced human brain", the authors pose a series of challenging questions. "Will therefore judicial review and human rights principles stand the test of time? How much opacity are we prepared to accept? How much error? How much uncertainty in terms of future benefits?" These are big questions, relevant also for every application of AI in the field of criminal justice.

31. The authors of the article themselves recognise one important risk or critique, noted also above in relation to PredPol. "Some of the predictors used in the model… (such as postcode[20]) could be viewed as indirectly related to measures of community deprivation. […] [One] could argue that [using postcode as a] variable risks a kind of feedback loop that may perpetuate or amplify existing patterns of offending. If the police respond to forecasts by targeting their efforts on the highest-risk postcode areas, then more people from these areas will come to police attention and be arrested than those living in lower-risk, untargeted neighbourhoods. These arrests then become outcomes that are used to generate later iterations of the same model, leading to an ever-deepening cycle of increased police attention." The Durham Police chief constable, Michael Barton, has said that people should be concerned if HART were used in the courts, but that it was used to reduce reoffending, not for sentencing.[21] This overlooks the fact that someone incorrectly identified as high risk

[18] Exclusive: UK police wants AI to stop violent crime before it happens", New Scientist, 26 November 2018.
[19] "Algorithmic risk assessment policing models: lessons from the Durham HART model and 'Experimental' proportionality", Oswald, Grace, Urwin and Barnes, 3 April 2018.
[20] I.e. a codified indicator of a specific neighbourhood, used to facilitate the delivery of post.
[21] "UK police test if computer can predict criminal behaviour", Financial Times, 6 February 2019.

because of bias in the training dataset would be denied access to the 'Checkpoint' programme and could instead end up being placed in pre-trial detention.

32. In 2017, HART was 'refreshed' with new data, with the aim of reducing reliance on postcode predictors. A dataset called 'Mosaic', developed by a company called Experian primarily as a commercial product for marketing purposes, was used. Mosaic is based on profiles of all 50 million adults resident in the UK, assembled using data gathered from public sources, including the Internet. Mosaic explicitly defines categories by reference to, for example, age group or ethnicity (e.g. "disconnected youth", "dependent greys" or "Asian heritage"). The NGO Big Brother Watch has argued that "For a credit checking company to collect millions of pieces of information about us and sell profiles to the highest bidder is chilling. But for police to feed these crude and offensive profiles through artificial intelligence to make decisions on freedom and justice in the UK is truly dystopian."[22] Durham Police stated that it "worked with Experian to improve its understanding of local communities",[23] and in 2018, it stopped using Mosaic, although reportedly for financial rather than ethical reasons.[24]

33. Whilst Durham Police has stressed that HART is used only for advisory purposes and that individual decisions are the responsibility of trained police officers, some have been sceptical about how things will work in practice. As with Kent Police's use of PredPol, Durham chief constable Barton has revealed that repeated cuts to his force's budget have motivated increasing recourse to new technologies.[25] These same cuts may have consequences for the availability of officers' time and attention, which is a significant factor in ensuring effective human responsibility for decisions made using HART. Andrew Wooff of Edinburgh University has observed that in the "time-pressured, resource intensive" world of policing, "I can imagine a situation where a police officer may rely more on the system than their own decision-making processes."[26] In addition, Big Brother Watch has argued that "given that the algorithm has been designed to detect the cases that the police might miss, or is reluctant to deem high-risk, is it really feasible to expect that police officers would consistently make judgments against the AI result? In order to function as a decision-making aid, it needs to alert the police to potential offenders they might not have considered. Therefore, it is questionable whether police would comfortably ignore these suggestions."[27]

34. Big Brother Watch has also drawn attention to explainability and accountability issues, noting that **"**AI decisions cannot be challenged because it may not be possible to even explain the conclusions reached**".** Durham Police responded that it "would be prepared to reveal the HART algorithm and the associated personal data and custody event datasets to an algorithmic regulator."[28] Not an unreasonable position to take, but at the same time a strong argument in favour of creating such bodies.

2.3. Predicting reoffending and deciding on remand, sentencing and parole – COMPAS

35. COMPAS, a system now owned by the company Equivant, is used in several US state jurisdictions to assess an individual's risk of reoffending. It has been described by Northpointe, now a division of Equivant, as a "fourth generation risk and needs assessment instrument. It is a web-based tool designed to assess offenders' criminogenic needs and risk of recidivism", using three 'scales': 'pretrial release risk' (i.e. risk of failure to appear and new felony arrest); 'general recidivism' (commission of new misdemeanour or felony offences within two years); and 'violent recidivism' (commission of violent offences).[29] New York State, for example, began by using COMPAS to assess people serving probationary sentences; it is now used there also by judges during sentencing. In Florida, it is used when deciding whether to remand an accused in custody or on bail. In Wisconsin, it is used at each step through the prison system, from sentencing to parole.

36. In 2014, then U.S. Attorney General Eric Holder warned that "Although [COMPAS-type risk scores] were crafted with the best of intentions, I am concerned that they inadvertently undermine our efforts to ensure individualized and equal justice. […] They may exacerbate unwarranted and unjust disparities that are already far too common in our criminal justice system and in our society." In response to these concerns, the investigative website ProPublica studied the use of COMPAS and reached two particularly critical conclusions. On its effectiveness, ProPublica found that "when a full range of crimes were taken into account, […] the

[22] "Police use Experian Marketing Data for AI Custody Decisions", Big Brother Watch, 6 April 2018.
[23] "Durham police criticised over 'crude' profiling", BBC News, 9 April 2018.
[24] Op. cit., Financial Times, 6 February 2019.
[25] Op. cit., Financial Times, 6 February 2019.
[26] Quoted in "UK police are using AI to inform custodial decisions – but it could be discriminating against the poor", Wired, 1 March 2018.
[27] "A Closer Look at Experian Data and Artificial Intelligence in Durham Police", Big Brother Watch, 6 April 2018.
[28] Wired, op. cit.
[29] "Practitioners Guide to COMPAS", Northpointe, 17 August 2012.

algorithm was somewhat more accurate than a coin flip", and it was "remarkably unreliable in predicting violent crime". On its neutrality, ProPublica found that whilst COMPAS made mistakes at roughly the same rate for both white and black individuals, it was far more likely to produce false positives (i.e. a mistaken 'high risk' prediction) for black people and more likely to produce false negatives (i.e. a mistaken 'low risk' prediction) for white people.[30]

37. It should be noted that ProPublica's findings have been criticised, both on technical grounds relating to the validity of the statistical analysis and in relation to how the findings were reported.[31] Yet even an author of one of these critical rebuttals has suggested that "what looks to be bias is not in the tool — it's in the system".[32] This reintroduces the notion of 'techwashing' – and the damaging, discriminatory consequences for individuals are much the same however the problem is described.

38. The controversy over COMPAS has also led commentators to underline the need for public debate, transparency and accountability. "Democratic societies should be working now to determine how much transparency they expect from ADM [automated decision-making] systems. Do we need new regulations of the software to ensure it can be properly inspected? Lawmakers, judges, and the public should have a say in which measures of fairness get prioritized by algorithms. But if the algorithms don't actually reflect these value judgments, who will be held accountable?"[33] Some have argued that such transparency should include access to the algorithm: "The thorny issues raised by software such as [COMPAS] is a compelling reason to make the formulas public, or at least to subject them to rigorous, independent review."[34]

2.4. Identifying solvable historical cases – the 'Cold Case Calendar'

39. Police in the Netherlands have developed a machine learning AI-based system to help them identify old, unsolved, serious cases ('cold cases') that may now have good prospects of being solved. This system was based on the insight that more than half the time, reopened cold cases are solved because of new technology that was not available at the time of the first investigation. (Almost half of the time, it is due to new tips from witnesses; the term 'Cold Case Calendar' originally referred to a method of canvassing prison inmates on 'cold cases'.) Once the 'cold case' files are digitised, they are fed into the AI system, which identifies those containing promising evidence that could be re-examined using new forensic techniques. Similar work conducted manually by police officers could take weeks of work per case, despite the likelihood of success being remote. The officers responsible for the project hope that it may be extended to identify 'cold cases' that could be solved using non-forensic data, such as social science, social networks and witness statements. It might even prove capable of improving the police's ability to solve ongoing investigations into offences.[35]

40. One problem with this approach is that for police data that had not been used in the original investigation, the legally-specified retention periods were too short for cold case investigations. Destruction of such data could thus remove a possible basis for solving the case through new forensic techniques. The Police Chief therefore decided not to destroy such old data. Instead, it is now subject to restricted access, with only a limited number of so-called 'gatekeepers' having access. The government has accepted this decision, stating that "it would be preferable to accept this deficiency in compliance with the law and to settle for the steps taken by the Police Chief to restrict access to the data to that which is strictly necessary. Compliance with the letter of the law could only be achieved by means of a coarse selection method, which would also destroy data that could contribute to detection in cold cases. This would seriously impede the resolution of these cases. This is now being prevented."[36] It is not clear, however, whether this approach is consistent with the general principle in data protection law that personal data should be processed (including being retained) only where there is a basis in law or the consent of the data subject.[37]

2.5. Other examples

41. Not all algorithmic systems that are intended to assist decision-making involve artificial intelligence. A US research project has found that courts in 46 out of 52 US states, along with the District of Colombia, use

[30] "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks", ProPublica, 23 May 2016.
[31] "False Positives, False Negatives, and False Analyses: A Rejoinder to 'Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks'", Flores, Lowenkamp and Bechtel.
[32] Anthony Flores, quoted in "The machines that could rid courtrooms of racism", Washington Post, 18 August 2016.
[33] "Inspecting Algorithms for Bias", Matthias Spielkamp, MIT Technology Review, 12 June 2017.
[34] Op. cit., Washington Post.
[35] "How the Dutch police are using AI to unravel cold cases", The Next Web, 23 May 2018.
[36] "Volunteers and artificial intelligence used in cold cases", Government of the Netherlands, 4 February 2019.
[37] See e.g. article 5 of Council of Europe data protection convention 108+.

some form of risk assessment tool during the pre-trial detention decision-making process.[38] Most of these are "non-learning algorithmic tools".[39] The most common tools were Public Safety Assessment (PSA), used in at least five states and 59 counties (administrative divisions of states), covering 56.3 million people; the Virginia Pretrial Risk Assessment Instrument (VPRAI), used in at least 43 counties covering 19.9 million people and its Revised version (VPRAI-R, supposedly corrected against race and gender bias), used in at least one state and 16 counties covering 14.3 million people; the Ohio Risk Assessment System Pretrial Assessment Tool (ORAS-PAT), used in at least five states and 48 counties; and COMPAS (see above), used in at least 11 counties covering 4.3 million people.

42. The Pretrial Justice Institute (PJI), a US non-profit organisation, had in the past promoted the use of algorithmic risk assessment tools as a means of reducing recourse to cash bail (conditional release on payment of a sum of money), which often led to less wealthy defendants being remanded in custody. In July 2020, however, it reversed its position, having observed that whilst pretrial detention rates fell, the ethnic profile of detainees remained around 50% black and 30% white. Its announcement of the new policy stated that:

> "We now see that pretrial risk assessment tools, designed to predict an individual's appearance in court without a new arrest, can no longer be a part of our solution for building equitable pretrial justice systems. Regardless of their science, brand, or age, these tools are derived from data reflecting structural racism and institutional inequity that impact our court and law enforcement policies and practices. Use of that data then deepens the inequity."[40]

43. Criticism has also come from within the public administration. A report for the State of Massachusetts general court found that "However well intentioned, risk assessment tools may have their own limitations due to their reliance on data of questionable correlation to predictability and the rigidity of application that restricts a judge's decision. […] Risk assessment tools depend on historical data to process and determine a likely outcome when applied to an individual defendant. The quality of the prediction largely depends on the quality of the data put into the tool. A significant flaw in many risk assessment tools is that the algorithm relies on arrest data as the starting point of analysis to predict a future event, and this data may incorporate implicit bias. […] Furthermore, many commercially available risk assessment tools do not reveal their algorithms or methodologies, which fosters mistrust in the system, makes it difficult for defense counsel to challenge their results, and inhibits the ability of a jurisdiction to address inconsistent results quickly and accurately. The reliance on arrest data as a predictive factor has led to biased results counterproductive to the purpose of using a tool in the first place. […] The drawbacks of implementing a currently available risk assessment tool would likely outweigh any incremental improvement in bail decisions."[41]

44. Judges on the Supreme Court of Ohio decided against recommending the use of pretrial risk assessment tools as part of a reform of the state's bail system, despite the previous findings of a special task force. The judges were said to have been notably influenced by arguments presented by the American Civil Liberties Union, which particularly emphasised the risk of racial bias.[42]

3. Oversight and regulation

3.1. The UK West Midlands Police and Crime Commissioner's Ethics Committee

45. In this section, I will focus on the situation of the West Midlands Police (WMP), which is well-developed and about which considerable information is available (including through my own contacts with the persons concerned). WMP has a 'Data-Driven Insights' project intended to make better use of the large amounts of data already held in its various datasets which exist as separate information 'silos' whose contents are not easily combined or cross-referenced. Realising that the same data-processing tools were easily transferrable to the other 43 police forces in the United Kingdom, this project was expanded to become the National Data Analytics Solution (NDAS). The NDAS is being developed in partnership with eight other UK police forces or agencies and with technical support from a private company, Accenture. The project consists of three main elements: Insight Search, a search engine covering nine databases, available to officers on mobile devices; Business Insights, which focuses on internal matters (such as officer well-being); and Insights Lab, which

[38] Mapping Pretrial Justice, a collaboration between the Movement Alliance Project and MediaJustice – see pretrialrisk.com.
[39] "Artificial Intelligence in Adjudication and Administration. A Status Report on Government Use of Technology in the United States", Coglianese and Ben Dor, 2019.
[40] "Updated Position on Pretrial Risk Assessment Tools", 2 July 2020.
[41] "Final Report of the Special Commission to Evaluate Policies and Procedures Related to the Current Bail System", General Court of Massachusetts, 31 December 2019.
[42] "Ohio Supreme Court proposes bail reforms that don't include risk assessments", cleveland.com, 24 January 2020.

develops innovative data analytics tools specifically for policing. The Data-Driven Insights project had a £17 million budget, with Insight Search estimated to have generated £23 million-worth of benefits in its first three years of operation alone. So far, its AI-based systems have mainly been used to identify individuals who may be likely to go on and commit knife or firearms offences ('Most Serious Violence') and to identify persons who may be victims of human trafficking ('Modern Slavery').

46. The West Midlands Police and Crime Commissioner (PCC) is an elected, independent official tasked with insuring the efficiency and effectiveness of the WMP and the accountability of its Chief Constable. The West Midlands PCC has established an Ethics Committee to advise the PCC and the Chief Constable on data science projects being proposed by the WMP's Data Analytics Lab. This committee's stated goal is to "ensure that ethics and people's rights are put at the heart of the Lab's work". Its terms of reference require it to consider a wide range of ethical principles; despite initial reticence on the part of the WMP, there is explicit reference to human rights, both generally and more specifically in relation to non-discrimination and privacy.[43] The committee is deliberately diverse, with a 50:50 gender balance and recruitment handled by an agency specialising in ethnic diversity, instructed to bring vacancies to the attention of "ethnic minority, single parent, disabled, religious, sexual orientation and gender groups". Most members have some form of specialist technical background,[44] although there are also non-specialist members. The committee includes a senior representative of another force, which Mr McNeil told me helps to "ground the committee in reality". Its activities are highly transparent and involve community outreach and consultation activities, albeit within a limited budget. It was decided that although the NDAS was a collaboration between several police forces, the West Midlands PCC's Ethics Committee would also advise on NDAS proposals, since there was no alternative national ethical structure. Nevertheless, one potential structural weakness of the Ethics Committee is that its existence depends on the West Midlands PCC; there is no guarantee that any future PCC would maintain it.

47. When making proposals to the Ethics Committee, the Data Analytics Lab uses the 'ALGO-CARE' matrix. This was developed by Ms Oswald and others following review of the Durham Police HART programme (see above)[45] and has been accepted by the National Police Chiefs' Council as a model for best practice in self-regulation. It is intended to prompt those developing algorithmic tools for use by the police to consider a series of essential practical and ethical concerns. ALGO-CARE is an acronym used to group a series of questions addressing these concerns, including the following:

- Advisory (is the algorithmic output used in an advisory capacity, with a human being responsible for decisions?)
- Lawful (does the use of the algorithm pursue a lawful purpose, is it proportionate, and was the data used lawfully obtained and processed?)
- Granularity (are the algorithm's suggestions sufficiently detailed, and are the input data reliable and corrected against possible bias?)
- Ownership (who owns the algorithm and the data, do the police have all the necessary rights to use them, and will the system be maintained, updated and secured as necessary?)
- Challengeable (what are the oversight and audit mechanisms? Are subjects notified and informed of its use?)
- Accuracy (does the specified level of accuracy match the policy goal and can it be verified periodically? Can the rate of false positives/ negatives be justified, what are the consequences of false results, and is the resulting risk acceptable? Do the users of the tool have the necessary expertise?)
- Responsible (is use of the tool objectively fair, transparent and accountable? Would it be considered in the public interest and ethical?)
- Explainable (is appropriate information available on the rules and weighting given to different factors? Would the police be able to have a data science expert explain and justify the tool?)

48. The Ethics Committee does not deliver binding decisions but rather gives advice to the PCC and Chief Constable. The advice is not about the law or compliance with legal standards, but about ethics more widely. In principle, the Chief Constable could ignore this advice, but s/he would remain accountable to the PCC who could, in theory, dismiss him/her. In practice, the committee's advice has always been followed. For example, a proposal for an 'integrated offender management' application went through three iterations before obtaining Ethics Committee approval. Mr McNeil told me that a number of proposals had been rejected, including a predictive model following feedback from diverse grassroots community bodies.

[43] See further at www.westmidlands-pcc.gov.uk/wp-content/uploads/2019/07/Ethics-Committee-Terms-of-Reference-as-at-1-April-2019.pdf?x83908.
[44] Marion Oswald, who participated in our committee's hearing in Berlin in November 2019, is a member.
[45] Oswald et al, 2018, op. cit.

49. From my conversation with the two senior officers, Chief Superintendent Todd and Detective Chief Inspector Dale, it was apparent that they had a good understanding of the ethical issues involved in the NDAS. We discussed issues of data protection, public trust, transparency, explicability, accountability and responsibility and oversight; as well as the need for technological solutions to prove their effectiveness, which would have to be assessed in order to determine the proportionality of any interference with protected rights. The officers underlined that there would always be a 'human in the loop' to take (and take responsibility for) the final decisions – the algorithmic output would only ever inform, not determine the decision. They also described how potential problems of historical bias in datasets had been raised and addressed during interactions between the police and the Ethics Committee.

50. Chief Supt. Todd and DCI Dale were aware, as was Mr McNeil (and Ms Oswald, during our committee hearing), that the Ethics Committee was not a perfect solution. They also highlighted the fact that there was no specific legislation in the UK on the use of AI by the police. The College of Policing was currently developing Authorised Professional Practice, a national common standard, on the use of data analytics but this too would not be mandatory; failure to comply could, however, be a relevant consideration in legal proceedings.

3.2. Elements of possible regulation of AI used in policing and criminal justice systems

51. Dr Veale and Ms Oswald also made suggestions for how AI should be regulated. The use of algorithmic systems in human decision-making processes should be justified on a case-by-case basis, taking into account the operational context. This should include consideration of the nature of resulting interventions and their impact, especially on marginalised communities. There should be a national register of AI systems deployed in the public sector. Particular attention should be paid to the cumulative effect of algorithmic systems: one, in isolation, might have only proportionate adverse effects, but multiple systems used at different stages of data processing might have much greater, hard to foresee cumulative impacts. Systems should be explicable to their immediate users and to those affected by the decision-making process. It should be possible to scrutinise and test the AI system from within the place of operation. Proprietary systems that do not allow for this should be excluded. The consequences of introducing AI elements into existing technology should be carefully examined: for example, facial recognition AI applied to CCTV networks had already created an extremely powerful form of mass surveillance going far beyond original expectations when CCTV was first introduced, and had the potential also to allow automated lip-reading.

3.3. CAHAI and a possible legal framework for artificial intelligence

52. In September 2019, the Committee of Ministers established the inter-governmental Ad Hoc Committee on Artificial Intelligence (CAHAI). The CAHAI has been instructed to examine the feasibility and potential elements of a legal framework for the development, design and application of artificial intelligence. Its work is based on Council of Europe standards of democracy, human rights and the rule of law, as well as other relevant international legal instruments and ongoing work in other international and regional organisations. Along with the usual participants representing Council of Europe member and observer States and other Council of Europe bodies (including the Assembly), the CAHAI has an exceptionally high level of engagement with private sector bodies, civil society, and research and academic institutions.

53. The CAHAI held its first meeting on 18-20 November 2019. Amongst other things, it decided that a key element of the future feasibility study would be a "mapping of risks and opportunities arising from the development, design and application of artificial intelligence, including the impact of the latter on human rights, rule of law and democracy". The CAHAI currently expects to adopt the feasibility study at its third meeting, scheduled for December 2020.

54. This is the institutional context within which the Assembly will debate the present and the various other AI-related reports currently under preparation in different committees. The Assembly has chosen to approach the topic on a contextual basis, examining the impact of AI in different areas. Within the Committee on legal affairs and human rights, for example, there are also reports on brain-computer interface technology, autonomous vehicles and (in the early stages of preparation) lethal autonomous weapons systems. The recommendations that the Assembly may adopt on the basis of these reports will thus provide important guidance for the CAHAI when mapping the risks and opportunities of AI and its impact on human rights, rule of law and democracy, and subsequently determining the need for a binding international legal framework.

55. The use of artificial intelligence tools in policing and criminal justice systems presents particular human rights risks. On the one hand, this is because of the significance of the decisions that might be taken on the basis of algorithmic output – decisions on surveillance, search and seizure, arrest, detention, sentencing, release on bail or parole etc, all of which have a particularly strong human rights impact. On the other, it is

because the historical police data used by AI to train algorithms are often tainted by historical bias, and because the introduction of AI into such a delicate decision-making process may undermine human responsibility and accountability and render the decisions non-transparent and difficult to challenge. My discussions with the West Midlands Police in the UK showed that the police themselves are aware of the need for legal regulation of the use of AI in policing, and in the absence of legal regulation are already taking steps of their own to ensure ethical oversight and standardized good practice. The stringent general requirements of lawfulness and due process set out in articles 5 and 6 of the Convention – the rights to liberty and security and to a fair trial, which are specifically intended to apply to the police and criminal justice system – confirm the need for legal regulation.

56. On the level of international standards, applicable norms already exist, notably in the European Convention on Human Rights and the case-law of the Court, as well as specialized instruments such as Convention 108+. Due to the sui generis nature of artificial intelligence and the novel ways in which it might be applied, however, the relevant standards would best be further defined in a specialized, binding instrument.

4. Conclusions and recommendations

57. Whilst research on digital computer-based AI dates back to the middle of the last century, its development and application have taken dramatic steps forward in recent years, thanks in large part to advances in machine learning techniques that have become possible through increased computer processing power and the availability of massive training datasets. AI is now used in a wide and ever-growing range of situations, some of them with the potential to have profound implications for democracy, human rights and the rule of law – including the criminal justice system.

58. This report has explored some of the general issues and concerns relating to AI and examined four concrete applications of AI in different aspects of the criminal justice system. It has also considered some of the specific concerns raised in relation to certain applications, along with a system of ethical oversight for the use of AI by the police. My analysis indicates that the general concerns about AI are equally applicable to the specific applications of AI in the policing and criminal justice systems that I studied. It also suggests that the general arguments in favour of stronger regulation of AI may be particularly relevant to its application in criminal justice systems. My detailed proposals for responding to this situation are set out in the attached preliminary draft resolution and recommendation.

Appendix

Artificial Intelligence – description and ethical principles

There have been many attempts to define the term "artificial intelligence" since it was first used in 1955. These efforts are intensifying as standard-setting bodies, including the Council of Europe, respond to the increasing power and ubiquity of AI by working towards its legal regulation. Nevertheless, there is still no single, universally accepted 'technical' or 'legal' definition.[46] For the purposes of this report, however, it will be necessary to describe the concept.

The term "artificial intelligence" is generally used nowadays to describe computer-based systems that can perceive and derive data from their environment, and then use statistical algorithms to process that data in order to produce results intended to achieve pre-determined goals. The algorithms consist of rules that may be established by human input, or set by the computer itself, which "trains" the algorithm by analysing massive datasets and continues to refine the rules as new data is received. The latter approach is known as "machine learning" (or "statistical learning") and is currently the technique most widely used for complex applications, having only become possible in recent years thanks to increases in computer processing power and the availability of sufficient data. "Deep learning" is a particularly advanced form of machine learning, using multiple layers of "artificial neural networks" to process data. The algorithms developed by these systems may not be entirely susceptible to human analysis or comprehension, which is why they are sometimes described as "black boxes" (a term that is also, but for a different reason, sometimes used to describe proprietary AI systems protected by intellectual property rights).

All current forms of AI are "narrow", meaning they are dedicated to a single, defined task. "Narrow" AI is also sometimes described as "weak", even if modern facial recognition, natural language processing, autonomous driving and medical diagnostic systems, for example, are incredibly sophisticated and perform certain complex tasks with astonishing speed and accuracy. "Artificial general intelligence", sometimes known as "strong" AI, able to perform all functions of the human brain, still lies in the future. "Artificial super-intelligence" refers to a system whose capabilities exceed those of the human brain.

As the number of areas in which artificial intelligence systems are being applied grows, spreading into fields with significant potential impact on individual rights and freedoms and on systems of democracy and the rule of law, increasing and increasingly urgent attention has been paid to the ethical dimension.

Numerous proposals have been made by a wide range of actors for sets of ethical principles that should be applied to AI systems. These proposals are rarely identical, differing both in the principles that they include and the ways in which those principles are defined. Research has shown that there is nevertheless extensive agreement on the core content of ethical principles that should be applied to AI systems, notably the following:[47]

- *Transparency.* The principle of transparency can be interpreted widely to include accessibility, explainability and explicability of an AI system, in other words the possibilities for an individual to understand how the system works and how it produces its results.
- *Justice and fairness.* This principle includes non-discrimination, impartiality, consistency and respect for diversity and plurality. It further implies the possibility for the subject of an AI system's operation to challenge the results, with the possibility of remedy and redress.
- *Responsibility.* This principle encompasses the requirement that a human being should be responsible for any decision affecting individual rights and freedoms, with defined accountability and legal liability for those decisions. This principle is thus closely related to that of justice and fairness.
- *Safety and security.* This implies that AI systems should be robust, secure against outside interference and safe against performing unintended actions, in accordance with the precautionary principle.
- *Privacy.* Whilst respect for human rights generally might be considered inherent in the principles of justice and fairness and of safety and security, the right to privacy is particularly important wherever an AI system is processing personal or private data. AI systems must therefore respect the binding standards of the EU General Data Protection Regulation (GDPR) and the Council of Europe's data protection convention 108 (and the 'modernised' convention 108+), as applicable.

[46] For a wide-ranging overview of attempts to define 'artificial intelligence', see *AI Watch: Defining Artificial Intelligence – Towards an operational definition and taxonomy of artificial intelligence,* Samoili, S., López Cobo, M., Gómez, E., De Prato, G., Martínez-Plumed, F., and Delipetrev, B., European Commission Joint Research Centre, 2020.

[47] See *AI Ethics Guidelines: European and Global Perspectives,* Draft Report commissioned by the Council of Europe Ad Hoc Committee on Artificial Intelligence (CAHAI), Ienca & Vayena, March 2020.

The effective implementation of ethical principles in relation to AI systems requires an ‘ethics by design’ approach, including a human rights impact assessment so as to ensure compliance with established standards. It is not sufficient for systems to be designed on the basis of technical standards only and for elements to be added at later stages in an attempt to evince respect for ethical principles.

The extent to which respect for these principles should be built into particular AI systems depends on the intended and foreseeable uses to which those systems may be put: the greater the potential impact on public interests and individual rights and freedoms, the more stringent the safeguards that are needed. Ethical regulation can thus be implemented in various ways, from voluntary internal charters for the least sensitive areas to binding legal standards for the most sensitive. In all cases, it should include independent oversight mechanisms, as appropriate to the level of regulation.

These core principles focus on the AI system and its immediate context. They are not intended to be exhaustive or to exclude wider ethical concerns, such as democracy (pluralistic public involvement in the preparation of ethical and regulatory standards), solidarity (recognising the differing perspectives of diverse groups) or sustainability (preserving the planetary environment).