Regulating Artificial Intelligence for Use in Criminal Justice Systems in the EU

Policy Paper

About Fair Trials

Fair Trials is a global criminal justice watchdog with offices in London, Brussels and Washington, D.C., focused on improving the right to a fair trial in accordance with international standards.

Fair Trials' work is premised on the belief that fair trials are one of the cornerstones of a just society: they prevent lives from being ruined by miscarriages of justice and make societies safer by contributing to transparent and reliable justice systems that maintain public trust. Although universally recognised in principle, in practice the basic human right to a fair trial is being routinely abused.

Its work combines: (a) helping suspects to understand and exercise their rights; (b) building an engaged and informed network of fair trial defenders (including NGOs, lawyers and academics); and (c) fighting the underlying causes of unfair trials through research, litigation, political advocacy and campaigns.

Contacts

Bruno Min
Legal Director (UK & International)
bruno.min@fairtrials.net

Griff Ferris
Legal & Policy Officer
griff.ferris@fairtrials.net

Executive Summary

'Artificial Intelligence' ('AI'), comprising machine-learning and other analytical algorithm-based automated systems, have become an important aspect of our lives. In recent years, this technology has been increasingly deployed in criminal justice systems across the world, playing an increasingly significant role in the administration of justice in criminal cases. This trend is often driven by perceptions about the reliability and impartiality of technological solutions, and pressures to make cost savings in policing and court services.

However, studies in various jurisdictions, including in Europe, provide substantial evidence that AI and machine-learning systems can have a significantly negative influence on criminal justice.

AI systems have been shown to directly generate and reinforce discriminatory and unjust outcomes; infringing fundamental rights, they have been found to have little to no positive influence on the quality of human decisions, and they have been criticised for poor design that does not comply with human rights standards.

Most AI used in criminal justice systems are statistical models, based on data which is representative of structural biases and inequalities in the societies which the data represents, and which is always comprehensively lacking in the kind of detail that is needed to make truly 'accurate' predictions or decisions. The data used to build and populate these systems is mostly or entirely from within criminal justice systems, such as law enforcement or crime records. This data does not represent an accurate record of criminality, but merely a record of law enforcement - the crimes, locations and groups that are policed within that society, rather than the actual occurrence of crime. The data reflects social inequalities or discriminatory policing patterns, and its use in these AI systems merely results in a reinforcing and re-entrenching of those inequalities and discrimination in criminal justice outcomes.

Given these extremely serious risks, strong regulatory frameworks are needed to govern the use of AI in criminal justice decision-making, and in some circumstances, restrict their use entirely.

Existing EU data protection laws restrict the use of automated decisions, but there are gaps and ambiguities that could result in the use of AI systems in ways that undermine human rights, if not accompanied by further guidance or legislation.

Firstly, EU laws currently only prohibit decisions that are solely based automated processes, but they do not regulate decision-making processes that are largely dependent on automated systems. Given that most AI systems in use today are designed and deployed to assist, rather than replace human decision-making in criminal justice systems, they largely fall outside the remit of EU data protection laws on automated decisions. Secondly, the prohibition on automated decisions are subject to broad exceptions. Individuals can be subject to decisions based solely on automated processes if authorised by EU or Member State law, and there are deemed to be appropriate human rights safeguards in place, including the right to obtain human intervention. However, there is not enough clarity on what safeguards are needed, and how 'human intervention' should be interpreted.

In order to regulate the use of AI in criminal justice proceedings, the EU must, at a minimum, set standards to address the following questions:

1) what standards are needed to govern the design and deployment of AI systems in criminal justice systems;
2) what safeguards needed in criminal justice proceedings to make sure that AI systems are used in accordance with human rights standards and prevent discrimination; and

3) how Member States should govern the deployment of AI systems and monitor their subsequent use.

The design of AI systems and their deployment in criminal justice proceedings should be regulated to generate human rights compliant, non-discriminatory outcomes. Minimum standards and safeguards should be set, which, if they cannot be adhered to, should preclude the use of the AI system in question. AI should also be regulated so that they are sufficiently transparent and explainable to enable effective independent scrutiny. AI systems should be designed and deployed to comply with and give effect to *inter alia* the right of access to court, the right to be presumed innocent, and the right to liberty. AI systems should not undermine the right to be tried by an impartial and independent tribunal, and in line with existing EU laws, no individual should be subject to an automated decision that results in a criminal record. AI systems should be designed so that they do not pre-designate an individual as a criminal before trial, nor should they allow the police to take unjustified, disproportionate measures against individuals without reasonable suspicion. AI systems that inform criminal justice outcomes should, as a general rule, favour outcomes that are favourable to the defendant. Where AI systems inform decisions on the deprivations of liberty, they should be calibrated to generate outcomes that favour release, and they should not facilitate detention other than as a measure of last resort. AI systems must be subject to rigorous testing to ensure they have the desired effect of reducing rates of pre-trial detention rates.

AI systems must be developed to guarantee that they do not generate discriminatory outcomes, ensuring that suspects and accused persons are not disadvantaged, either directly or indirectly, on account of their protected characteristics, including race, ethnicity, nationality or socioeconomic background. AI systems should be subject to mandatory testing before and after deployment so that any discriminatory impact can be identified and addressed. AI systems which cannot adhere to this minimum standard should have no place in the criminal justice system.

AI systems need to be transparent and explainable, so they can be understood and scrutinised by their primary users, suspects and accused persons, and the general public. Commercial or proprietary interests should never be a barrier to transparency. AI systems must be designed in a way that allows criminal defendants to understand and contest the decision made against them. It should be possible to carry out an independent audit of the AI system, and its processes should be reproducible for that purpose.

Member States should have laws that govern how AI systems are relied upon in criminal proceedings, and there must be adequate safeguards to prevent over-reliance on AI by decision-makers, to prevent discrimination and to ensure scrutiny and effective challenge by the defence.

Procedural safeguards should actively tackle automation-bias amongst criminal justice decision-makers. Examples include:

a) making it a legal requirement for decision-makers to be adequately alerted and informed about the risks associated with AI systems;
b) making AI systems' assessments intelligible to decision-makers;
c) requiring decision-makers to provide full, individualised reasoning for all decisions influenced by an AI system; and
d) making it easy for decision-makers to overrule AI assessments that produce unfavourable outcomes for defendants.

Criminal justice procedures should ensure that defendants are notified if an AI system has been used which has or may have influenced a decision taken about them at any point in the criminal justice system, from investigation to arrest, from charge to conviction, and sentence. Procedures should enable the full disclosure of all aspects of AI systems that are necessary for suspects and accused persons to contest their findings, and disclosure should be in a form which is clear, comprehensible and understandable to a layperson, without the need for technical or expert assistance, in order to ensure fairness, equality of arms, and to discharge the obligations to provide all relevant information and be given reasons for decisions under the right to a fair trial. Suspects and accused persons should also be given effective access to technical experts who can help to analyse and challenge otherwise incomprehensible aspects of AI systems. Training should be made available to all primary users of AI systems, and to criminal defence practitioners, so that there is greater awareness of AI technology, and of the risks of over-reliance on AI.

Effective regulation of AI systems should be facilitated by a governance and monitoring framework. AI systems should not be deployed unless they have undergone an independent public impact assessment with the involvement of appropriate experts, that is specific both to the purpose for which the AI system is deployed, and the locality where it is deployed. A requirement of the assessment should be a consideration whether it is necessary to use AI in the particular use case, or whether an alternative solution could achieve the same aims.

As far as it is possible to do so, AI systems should also be tested for impact pre-deployment, as part of which should be the minimum requirement to prove that the AI system has no discriminatory impact, either directly or indirectly, before it can be deployed. AI systems should be kept under regular review post-deployment. Effective monitoring of AI systems is not possible unless there is sufficient data that makes it possible to discern their real impact. In particular, Member States need to collect data that allow them to identify discriminatory impacts of AI systems, including discrimination on the basis of race and ethnicity.

Background

Rapid technological advancements in recent years have made artificial intelligence ('AI') an increasingly prominent aspect of our lives.

There are differences of opinion as to the definition of artificial intelligence and its true meaning, but for the purposes of this paper we are broadly referring to automated-decision making systems based on algorithms, including machine-learning, which are used in the criminal justice system.

There is little doubt that AI has great capacity to increase human potential and improve the lives of many, but the increasing role of AI in assisting important public functions has also highlighted serious risks and challenges. If not subject to proper regulations and oversight, AI can threaten fundamental human rights, and far from expanding human potential, it can amplify and worsen harmful aspects of our society, including inequality and injustice.

This challenge is particularly evident where AI has been used to assist the administration of justice in criminal cases. In recent years, more and more jurisdictions across the world have begun to use AI technology to inform and assist policing and judicial decisions, often driven by perceptions about the reliability and impartiality of technological solutions, and pressures to make cost-savings in policing and court services. In some countries, algorithmic processes can influence which geographic neighbourhoods should be subject to increased law enforcement and when, as well as which individuals should be specifically targeted by law enforcement. They can help to determine whether someone should be arrested, whether they should be charged with a criminal offence, whether they should be detained in prison before trial, and if convicted and sentenced, the length of their sentence. AI is being used more and more to influence highly sensitive, high impact decisions that have far-reaching, long-term implications for individuals' rights.

Research emerging from the United States, where the use of AI in criminal justice is particularly widespread, and from the United Kingdom and some EU Member States, however, seriously questions whether AI has a positive influence on criminal justice systems. AI tools and systems have been found to actively generate discriminatory criminal justice outcomes, they have been found to have little to no positive influence on the quality of human decisions, and they have been criticised for poor design, that does not reflect or give effect to human rights standards. These criticisms might not be justified for all AI systems, but these studies highlight the need for much stronger regulatory frameworks to govern the use of AI.

We believe that unless it is subject to robust regulation, it is unlikely that AI can be used in criminal justice systems without undermining the right to a fair trial. In some cases, it should be entirely restricted from doing so.

EU Member States should be encouraged to take a much more cautious approach to AI and subject automated processes to more stringent rules that are designed to ensure human rights compliance.

There is the potential for AI systems, if properly and robustly regulated, to have a positive impact on criminal justice system, advancing human rights, such as by analysing analyse law enforcement or judicial decisions to identify patterns of erroneous or poor decision-making, or discrimination.

The EU is already a world leader on AI regulation, having adopted ground-breaking data protection laws in recent years to shield individuals from automated decisions that have an adverse effect on their rights. We welcome the EU's commitment to build further on existing legal standards, and we emphasise that addressing the impact of AI on criminal justice has to be a primary consideration for EU policy makers when deciding on appropriate legal standards. Discussions around the impact of artificial intelligence on human rights have largely been centred on data protection, the right to privacy, and broader questions on ethics and human dignity. However, despite the increasing use of AI systems in criminal justice systems across the world, only limited discussion so far has focused on how these systems impact the right to a fair trial, and what regulations are needed to address those concerns.

About this Paper

Fair Trials has produced this policy paper to highlight the need for EU-wide standards on the regulation of AI in criminal justice, and to inform EU policy makers about the standards and safeguards needed to ensure effective protection of fair trial rights where criminal justice decisions are assisted by AI.

The EU Commission recognised that AI represents risks for fundamental rights, including the right to a fair trial in its 2020 White Paper, *'On Artificial Intelligence – A European approach to excellence and trust*'. It also recognised the need for improvements to the EU's legislative framework on AI, noting in particular, the challenges in the 'effective application and enforcement of existing EU and national legislation' and the 'limitations of scope of existing EU legislation'.

In this paper, we identify the most common fair trial rights issues raised by existing AI systems, based on examples and experiences from the EU, the United Kingdom, and the United States. We also offer examples of practical legal and policy solutions that could help to address these challenges, and to assist in the effective implementation of the EU's fundamental rights standards in this area. We recognise that the use of AI has a broader impact on human rights beyond the right to a fair trial, and that there are important social and ethical issues that also need to be addressed. However, we have narrowed the focus of this paper given Fair Trials' mission and field of expertise.

This paper should not be treated as an exhaustive list of fair trial rights standards that need to be introduced. AI is used in many ways in criminal justice systems cross the world, and as the technology continues to develop, it is likely that we will eventually see the deployment of AI technology in ways never imagined before. This paper focuses primarily on AI systems that carry out individualised risk assessments, given that these types of systems have had the most significant impact on individuals' rights so far, and we envisage that similar systems will become increasingly common in the near future.

Existing EU Legal Framework

Existing EU laws restrict the use of automated decisions in a wide variety of contexts. Article 22 of the General Data Protection Regulation ('GDPR') provides that data subjects have the right not to be subject to decisions '*solely*' based on automated processes, where they produce '*legal effects*' concerning them, or where they '*similarly significantly affect*' them. The Law Enforcement Directive ('LED') – the EU data legislation that governs the processing of data for criminal justice purposes – has a very similar provision at Article 11, which requires Member States to prohibit decisions based solely on automated processing, where they produce 'adverse legal effects' on the individual, or effects that are '*similarly significant*'.

However, there are two notable gaps in the existing legislative framework governing automated decision-making systems under both the GDPR and the LED. These ambiguities and potential loopholes could be exploited in ways that seriously undermine the general prohibition of automated decision-making processes, and adversely impact human rights. It is necessary, therefore, that the EU provides further guidance on how these provisions should be interpreted, including thorough legislation (if appropriate) to further clarify the circumstances in which Member States are allowed to deploy AI systems for criminal justice proceedings.

Firstly, the provisions in the GDPR and LED only prohibit *decisions* based '*solely*' on automated processes. In other words, the laws regulate the impact of decisions made through automated processing, but not the AI systems themselves. As discussed later in this paper, the main human rights challenges of AI systems can be attributed to how they are designed and trained, and the types of technology used such as machine-learning, so it is crucial that decisions about the design and deployment of AI systems are also regulated.

Secondly, neither the GDPR or LED provide regulatory standards to govern situations where automated processing is not the *sole* basis of a decision, but a primary influencer. In reality the difference between a fully automated decision and a decision made with a 'human-in-the-loop' is not always clear, and because of this strict classification, AI systems are able to be used and have significant legal effects without the corresponding safeguards. Stronger legal standards are needed to make sure that semi-automated decision-making processes do not become *de facto* automated processes.

Thirdly, the prohibition on automated decision-making is subject to two very broad exceptions. Automated decisions are prohibited under the GDPR and LED, '*unless authorised by Union or Member State law'* and there need to be '*appropriate safeguards for the rights and freedoms of the data subject, at least the right to obtain human intervention*'.[1] These provisions give extremely wide discretion on Member States to override the general prohibition. It is significant that EU laws emphasise the need for human rights safeguards, and the need to ensure the possibility of human interventions, but neither or these concepts have yet been adequately defined. Although influential actors like the EU and the Council of Europe have established principles on the ethical and responsible use of AI, there is currently no authoritative guidance on the practical safeguards that need to be in place.[2] Likewise, the meaning of 'human intervention' is open to interpretation. LED provides some

[1] Article 11(1), LED; Article 22(2)(c) and (3), GDPR

[2] On the other hand, Civil Society Organisations, such as the 'Partnership for AI' and 'AI Now' in the United States have attempted to address this gap through various recommendations and guidelines.

guidance on who should be carrying out the human intervention,[3] but there needs to be greater clarity on what meaningful human intervention entails in different contexts.

In order to regulate the use of AI in criminal justice proceedings, and close the gaps in existing data protection laws, the EU must, at a minimum, set standards to address the following questions:

1) what standards are needed to govern the design and deployment of AI systems in criminal justice systems;
2) what safeguards needed in criminal justice proceedings to make sure that AI systems are used in accordance with human rights standards and prevent discrimination; and
3) how Member States should govern the deployment of AI systems and monitor their subsequent use.

[3] Recital 38

Part 1: Regulating the Design and Deployment of AI Systems in Criminal Justice Systems

AI systems deployed to assist criminal justice decision-making have to be fit-for-purpose. The purposes of AI systems differ depending on the context in which they are deployed, but there are a few common considerations that need to be taken into account to determine whether it is appropriate for the AI system to be used.

Firstly, AI systems have to be designed to produce outcomes that are desirable from a human rights and non-discrimination perspective. This means that rather than being exclusively focused on delivering 'accurate' outcomes in criminal cases, AI systems have to be designed to facilitate fair, impartial and non-discriminatory criminal processes. Developers of AI systems and public entities that commission them should, in particular, make sure that AI systems are consciously designed to give effect to, and promote the right to fair trial. The fundamental issues with the way AI systems are designed and built, resulting in discriminatory outcomes, which must also be considered. Given the significant evidence of AI systems influencing discriminatory outcomes, special efforts must be made to ensure that AI systems do not produce discriminatory outcomes.

Secondly, AI systems need to be designed in a way that makes it possible for criminal defendants and the broader public to scrutinise them. This means that AI systems should not only be made open to scrutiny (rather than concealed to protect commercial interests), but their inner workings and processes should also be discernible and comprehensible.

AI Systems should be designed to protect and promote the right to a fair trial

Where AI systems are used to assist or inform criminal justice decisions, they support an important act of public administration that has a significant impact on the rights of suspects and accused persons. AI systems do more than just provide outputs that decision-makers can take into consideration as evidence. By attempting to mimic human analytical processes and reasoning, they can provide influential advisory input into human decision-making, or even replace it altogether. As such, it is right that human rights standards that govern criminal justice decision-making also apply to AI systems.

The Council of Europe and the EU Commission's High Level Expert Group on Artificial Intelligence ('AI HLEG') have both recognised that fundamental rights should be a key guiding principle for the design and deployment of AI systems.[4] The Council of Europe recommends that AI systems are built according to 'human rights by design' principles, and recognises that AI systems do not undermine the right to a fair trial under the European Convention on Human Rights ('ECHR'). The AI HLEG has similarly recognised that the respect for fundamental rights, as enshrined in the EU Charter of Fundamental Rights and international human rights instruments, form the foundations of trustworthy AI. AI HLEG's Ethics Guidelines for Trustworthy AI recognise the need for AI systems to also comply with other types of EU legislation. Although not mentioned explicitly in the Ethics Guidelines, Fair Trials would emphasise that the design of AI systems and the ways in which they are deployed in the EU should, in particular, be compatible with the standards set out in the procedural rights directives under the

[4] European Commission for the Efficiency of Justice, 'European ethical Charter on the use of Artificial Intelligence in judicial systems and their environment' (2018); Independent High-Level Expert Group on Artificial Intelligence, 'Ethics Guidelines for Trustworthy AI' (2019)

‘Roadmap for strengthening procedural rights of suspected or accused persons in criminal proceedings’.[5]

We would also like to note the potential for AI systems to have a positive impact on criminal justice systems. Public debate about the relationship between AI and human rights have predominantly been centred on the idea that AI is a threat to human rights. It is equally important, as technology takes an increasingly prominent role in public life, to consider what positive potential they may have. Policy-makers, developers, civil society activists, and other stakeholders should try to identify ways in which AI can also play an active role in advancing human rights, and improve the fairness of criminal justice systems. For example, AI systems could be used to analyse law enforcement or judicial decisions to identify patterns of erroneous or poor decision-making, or discrimination, for preventative purposes.

AI systems which are used as part of criminal justice decision-making should be designed not just to ensure that they do not undermine the right to a fair trial, but also to promote it. However, as explained below, given the embedded biases in the criminal data used to develop and train AI systems, there are serious doubts, based on recent studies, whether AI systems can promote fair criminal justice at all.

There are various aspects of the right to a fair trial, and without speculating what kind of AI systems will be developed in the future to support criminal justice decision-making, it is difficult to articulate how fair trial rights standards should inform the design of AI systems. However, examples of AI systems currently deployed in the EU and elsewhere suggest that there are certain aspects of the right to a fair trial that require special attention. These are:

a) the right of access to court
b) the presumption of innocence;
c) the principle of the equality of arms; and
d) the right to liberty.

Access to Court

The notion of AI systems replacing courts to determine the guilt or innocence of the accused may seem far-fetched at present, but there is a growing trend of automated administration of justice across the world that might threaten the right of access to court. For example, in several European countries, speeding and other minor traffic offences have been detected and enforced by means of automated processes for more than a decade.[6] Although nominally criminal processes, these types of proceedings are, in reality, normally administrative in nature, and they rarely have a ‘significant’ impact on the rights of individuals. However, as surveillance technology develops thanks to AI, there is a real likelihood that the scope of crimes punishable by way of automation will increase.[7]

In the United Kingdom, the government announced plans in 2017 that would enable defendants to enter guilty pleas via an online portal after viewing the charges and evidence against them, for a small

[5] Resolution of the Council of 30 November 2009 on a Roadmap for strengthening procedural rights of suspected or accused persons in criminal proceedings, 2009/C 295/01

[6] European Commission, ‘Speed Enforcement’ <https://ec.europa.eu/transport/road_safety/specialist/knowledge/speed/speed_limits/speed_enforcement_en>; Adam Snow, ‘Automated Road Traffic Enforcement: Regulation, Governance and Use – a Review’, RAC Foundation (2017)

[7] E.g. In China, AI systems are being used to enforce penalties for using a mobile phone whilst driving. BBC, ‘Chinese driver gets ticket for scratching his face’ (2019) https://www.bbc.co.uk/news/blogs-news-from-elsewhere-48401901

number of minor offences.[8] Under this procedure, known as 'automatic online conviction', defendants would be automatically convicted and fined without any judicial oversight if they accept the charges against them. Although it is debatable whether this system can truly be characterised as an AI system, it is an example of automated administration of criminal justice, that replaces a function usually played by courts.

It is worrying that the UK government has proposed expanding this scheme to other 'non-imprisonable' offences, if it is regarded as a success.[9] Fair Trials has outlined concerns about expanding the scope of cases where accused persons can be convicted without judicial oversight, even if such procedures are reserved solely for minor, non-imprisonable offences.[10] The impact of a criminal conviction, even for a minor offence, can be numerous, long-term, and hard to predict, affecting *inter alia* job prospects, educational opportunities, and immigration status. It is crucial that what amounts to 'legal effects' and 'similar significant effects' concerning the data subject for the purposes of automated decision-making are interpreted very broadly.[11] In particular, given that a criminal record always has a 'legal' or 'significant' effect, any automated decision-making process that directly results in a criminal record should be prohibited.

AI systems should not undermine the right to be tried by an impartial and independent tribunal, and in line with existing EU laws, no individual should be subject to an automated decision that results in their being held in custody or detention, gives them a criminal record or which determines a criminal sentence or sanction. No individual should be subject to an automated decision which engages their human rights without meaningful human input.

Presumption of Innocence

The right to be presumed innocent in criminal proceedings is a basic human right, and one that is expressly recognised and safeguarded by EU law under Directive 2016/343 (the 'Presumption of Innocence Directive').[12] The increasing use of AI in the sphere of criminal justice, however, raises questions about the scope of this right, and how AI systems should be built and used to protect it. Concerns about how AI systems undermine the presumption of innocence have been voiced in the context of certain types of predictive policing software.[13]

A variety of predictive policing tools that aim to facilitate preventative policing measures and to deter crimes before they have taken place have been developed and deployed across Europe.[14] Tools which predict the time and place where certain crimes are likely to take place have been used in many

[8] UK Ministry of Justice, 'Transforming our justice system: assisted digital strategy, automatic online conviction and statutory standard penalty, and panel composition in tribunals Government response' (2017)
[9] UK Ministry of Justice, 'Online convictions/statutory fixed fine Impact Assessment' (2016)
[10] Fair Trials, 'Written evidence from Fair Trials (CTS0079) (2019), http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/justice-committee/hmcts-court-and-tribunal-reforms/written/97940.pdf
[11] GDPR, Article 22(1)
[12] Directive (EU) 2016/343 of the European Parliament and of the Council of 9 March 2016 on the strengthening of certain aspects of the presumption of innocence and of the right to be present at the trial in criminal proceedings; Article 6(2), ECHR
[13] Alan Turing Institute, 'Using analytics in policing: Ethics Advisory Report for West Midlands police' (2018), https://www.turing.ac.uk/news/using-analytics-policing-ethics-advisory-report-west-midlands-police
[14] Fieke Jansen, 'Data Driven Policing in the Context of Europe' (2018) https://datajusticeproject.net/wp-content/uploads/sites/30/2019/05/Report-Data-Driven-Policing-EU.pdf

European countries. Similar tools have also been developed to identify potential suspects, which are used widely in the US, and now increasingly in Europe.[15]

An example is the 'Strategic Subject List' in Chicago, a police database of around 400,000 local residents who were assigned threat scores that determine the likelihood that they will commit crimes.[16] The algorithm used to generate these scores were not open to the public, so the exact process by which individual risk levels are assessed were not known. Despite this lack of transparency, it is clear that threat scores generated by the software had significant impacts on individuals' rights – in particular, their right to privacy. Individuals with higher threat scores were, for example, more likely to be subject to targeted police surveillance, or home visits – as though they were officially recognised as predisposed to commit crimes, irrespective of any credible suspicion of wrongdoing.[17] The Strategic Subject List was decommissioned in January 2020 by the Chicago police who cited ineffectiveness as the primary reason for the decision.[18]

These types of predictive policing tools are now being used in Europe. In the United Kingdom, a coalition of police forces have been developing a system not dissimilar to the Strategic Subject List, that aims to identify individuals who are likely to commit crimes.[19] Known as the National Data Analytics Solution ('NDAS'), this risk assessment tool uses statistical analysis and machine-learning to inform policing decisions, and to facilitate 'early interventions' where appropriate.[20] The sources of data that the system uses to conduct its risk assessments raises concerns that the system will be built to profile individuals on the basis of very sensitive, personal information, including stop and search data, data from social services, and the National Health Service.[21] Where this data is used to indicate the likelihood of people's criminality, it will inevitably flag up people whose profiles fit those who are over-represented in that data as being higher risk. It is particularly worrying that an individual might be profiled for policing purposes on the basis of their health conditions or their access to essential services, such as welfare or benefits. These factors should not be regarded as relevant factors for determining whether someone may commit criminal offences.

Also in the UK, the Metropolitan Police in London operates a database called the Gangs Matrix, which contains information and risk-assessments on individuals who are alleged 'gang' members.[22] This database was created using criminal justice data, including police and crime records. The Gangs Matrix and the assessments it produces assists policing decisions, including the deployment of stop and search, and further enforcement action, such as imprisonment and deportation. A further tactic resulting from the risk assessments made by the Gangs Matrix is the threat of eviction or exclusion

[15] Don Casey et al. 'Decision Support Systems in Policing', *European Law Enforcement Research Bulletin*, (2019) https://bulletin.cepol.europa.eu/index.php/bulletin/article/view/345

[16] Andrew Guthrie Ferguson, 'The Police Are Using Computer Algorithms to Tell If You're a Threat', Time (3 October 2017), https://time.com/4966125/police-departments-algorithms-chicago/

[17] Ibid.

[18] Sam Charles, 'CPD decommissions 'Strategic Subject List' (27 January 2020) https://chicago.suntimes.com/city-hall/2020/1/27/21084030/chicago-police-strategic-subject-list-party-to-violence-inspector-general-joe-ferguson

[19] Hettie O'Brien, 'The police know what you'll do next summer', *New Statesman* (15 August 2019) https://www.newstatesman.com/politics/uk/2019/08/police-know-what-you-ll-do-next-summer

[20] Police Transformation Fund, 'National Analytics Solution, Final Business Case v6.0' http://foi.westmidlands.police.uk/wp-content/uploads/2019/01/report1_.pdf

[21] Sarah Marsh, 'Ethics committee raises alarm over 'predictive policing' tool', *The Guardian* (20 April 2019)

[22] Metropolitan Police, 'Gangs violence matrix', https://www.met.police.uk/police-forces/metropolitan-police/areas/about-us/about-the-met/gangs-violence-matrix/

from education, as names and details of these alleged gang members have been shared with education, healthcare and housing providers. [23]

In the Netherlands, the government has been running an algorithmic risk assessment tool, ProKid 12-SI, which purports to assess the risk of criminality of 12-year-old children since 2009.[24] ProKid uses existing police data on these children, such as reports of where children have come into contact with the police, their addresses, information about their 'living environment', even including whether they are victims of violence, to identify them as being in one of four categories of 'risk' of committing crime in future.[25] The system assesses children based on their relationships with other people and their supposed risk levels, meaning that individuals can be deemed higher risk by being linked to another individual with a high risk assessment, such as a sibling or a friend.[26] Parents' assessed risk can also impact a child's risk level. ProKid's algorithms assess risks in relation to future actions that the children have not yet carried out, and judges them on the basis of the actions of others close to them.[27] These risk assessments result in police 'registering' these children on their systems and monitoring them, and then referring them to youth 'care' services.[28] ProKid frames children as potential perpetrators even when they are registered as victims of violence; having potentially serious implications on the presumption of innocence.[29]

Several similar tools are also used in the Netherlands, including the Reference Index for High Risk Youth, a large scale risk assessment system that focuses on assessing under-23 year olds.[30]

Predictive policing tools like NDAS, ProKid and the Gangs Matrix can be regarded as part of a broader trend in law enforcement that moves away from 'reactive' policing, and towards 'preventative' or 'proactive' policing.[31] NDAS and other similar predictive policing tools intend to pursue legitimate objectives of preventing, or reducing harm,[32] but there are serious concerns that these systems single-out individuals as 'pre-criminals', who are subject to police interventions even though they are not formally suspected of any crime, and there is no evidence that they have done anything wrong.[33] It is of further concern that these types of predictive policing tools do not necessarily designate individuals' risk levels on the basis of their past actions, or behaviour that can be regarded as 'suspicious' in any

[23] Amnesty International, 'Trapped in the Matrix', (2018), https://www.amnesty.org.uk/files/reports/Trapped%20in%20the%20Matrix%20Amnesty%20report.pdf

[24] Abraham et al, "ProKid 2- identification tool evaluated", WODC DSP-groep (2011) https://english.wodc.nl/onderzoeksdatabase/evaluatie-signaleringsinstrumenten-prokid.aspx

[25] K La Fors-Owyczynik, 'Prevention strategies, vulnerable positions and risking the 'identity trap': digitalized risk assessments and their legal and socio-technical implications on children and migrants', (2016) https://www.tandfonline.com/doi/full/10.1080/13600834.2016.1183307

[26] Ibid.

[27] K La Fors-Owyczynik, 'Profiling 'Anomalies' and the Anomalies of Profiling: Digitalized Risk Assessments of Dutch Youth and the New European Data Protection Regime' (2016), https://link.springer.com/chapter/10.1007/978-3-319-48342-9_7)

[28] Abraham et al, "ProKid 2- identification tool evaluated", WODC DSP-groep (2011) https://english.wodc.nl/onderzoeksdatabase/evaluatie-signaleringsinstrumenten-prokid.aspx

[29] K La Fors , 'Minor protection or major injustice? – Children's rights and digital preventions directed at youth in the Dutch justice system', Computer Law and Security Review, (2015) https://research.tilburguniversity.edu/en/publications/minor-protection-or-major-injustice-childrens-rights-and-digital-

[30] Netherlands Youth Institute, 'Reference Index for youth at risk: factsheet', https://www.nji.nl/nl/Download-NJi/Publicatie-NJi/Reference_Index_Youth_at_Risk.pdf

[31] Alan Turing Institute (n 13)

[32] Ibid.

[33] Hettie O'Brien, 'The police know what you'll do next summer', *New Statesman* (15 August 2019) https://www.newstatesman.com/politics/uk/2019/08/police-know-what-you-ll-do-next-summer

way, but on account of factors far beyond their control and immutable characteristics. In particular, there is strong evidence to suggest that AI systems have a tendency to overestimate risks and criminality or certain ethnic and racial groups. For example, out of 3,800 people on the Gangs Matrix, 80% are 12-24 years old, and 78% of them are black – clearly disproportionate and discriminatory. The discriminatory impact of AI in criminal justice systems is discussed in further detail in the following section.

Although predictive policing tools do not directly 'convict' people, they not only allow the police to treat legally innocent individuals as pseudo-criminals, but they can also result individuals being deprived of their basic rights with regard to education, housing, and other public services – effectively 'punishing' them on account of their profiles. This seriously damages the fundamental human rights principle that the matter of guilt or innocence can only be determined by means of a fair and lawful criminal justice process.[34]

While it is clear that certain types of predictive policing can infringe the presumption of innocence from a moral and ethical viewpoint, it is debatable whether these systems also violate the *legal* presumption of innocence under EU law and international human rights law. The Presumption of Innocence Directive applies to natural persons who are 'suspects' and 'accused persons', from the moment they are suspected or accused of a crime.[35] However, there is some ambiguity about the exact stage at which an individual attains the status of a 'suspect' under the Directive,[36] and about whether the scope of the Directive extends to decisions to designate an individual as a suspect (or a 'pre-criminal'). On the other hand, the European Court of Human Rights appears to have taken a clearer position that measures undertaken pre-charge, as a general rule, fall outside the scope of the presumption of innocence.[37] It has also held that preventative measures, such as surveillance do not amount to a criminal sanctions for the purposes of Article 6 ECHR.[38]

Even if the current language on the presumption of innocence is such that it is not directly applicable to the predictive policing context, it must be recognised that these tools nevertheless interfere with human rights. In particular, the targeted surveillance that results from predictive policing has clear implications on the right to privacy. The acceptable degree to which criminal justice processes can interfere with this right is a matter that might require better articulation, as is the question of the impact of Article 8 ECHR violations on criminal proceedings.

AI systems that inform charging decisions have also been developed and deployed. An example of this is the Harm Assessment Risk Tool ('HART') currently being used by Durham Constabulary in the United Kingdom. HART uses a machine-learning algorithm to assess a suspect's risk of reoffending, using over thirty variables that characterise an individual's criminal history and socio-demographic background. The risk assessments conducted by HART are used by the local police to determine whether an individual should be charged, or diverted into a rehabilitation programme. HART does not determine whether an individual is guilty or innocent, but its assessment can trigger a chain of events that can result in the deprivation of liberty, and/or a criminal conviction. Charging decisions should surely be

[34] ECHR, Article 6(2)

[35] Article 2

[36] cf. ECtHR's definition of 'suspect' and 'charge' in *Mikolajova v. Slovakia*, App No. 4479/02 (Judgment of 18 January 2011), paras 40-41 and *Bandeltov v. Ukraine*, App No. 23180/06 (Judgment of 31 October 2013), para. 56

[37] ECtHR, *Gogitizde and Others v. Georgia*, App. No. 36862/05 (Judgment of 12 May 2015)

[38] ECtHR, *Raimondo v. Italy*, App. No. 12954/87 (Judgment of 22 February 1994)

based on the merits of individual cases, and it is difficult to imagine how decisions on entry into diversion programmes can be made by means other than a careful consideration of individual circumstances. These types of high impact, fact-sensitive decisions should never be delegated to automated processes, particularly those which operate by identifying correlations rather than causal links between an individual's characteristics and their likely behaviour.

An examination of HART also reveals flaws in how the tool is designed. HART is calibrated to err on the side of caution,[39] because it regards under-estimation of risk levels as a more serious error than over-estimations, so that under-estimations occur less frequently. In other words, HART is deliberately designed to underestimate who is eligible for entry into the diversion programme, so it is predisposed to over-criminalise. This approach conflicts with the notion that any doubt in a criminal case should be interpreted in favour of the defendant ('*in dubio reo*').[40] A human rights compliant approach to criminal justice decision-making would do the opposite of what HART does – it would need to err on the side of the defendant.

AI systems should respect the presumption of innocence and they must be designed so that they do not pre-designate an individual as a criminal before trial, nor should they allow or assist the police to take unjustified, disproportionate measures against individuals without reasonable suspicion. AI systems that inform criminal justice outcomes should, as a general rule, favour outcomes that are favourable to the defendant.

Equality of Arms

A major concern raised in the studies of certain AI systems is that they are inaccessible for adequate scrutiny by defendants and their lawyers. This has serious implications for the principle of equality of arms and the right to an adversarial process, because without information about how a decision is made, it is difficult to envisage how defendants can question the accuracy and legality of the decision. The need for AI systems used in criminal justice to be transparent, explainable and understandable to all is addressed in more detail below.

The Right to Liberty

In the United States, 'risk-assessment' tools that use AI technology have been used to assist pre-trial assessments that determine whether a defendant should be released on bail, or held on remand pending their trial. Examples of risk-assessment tools currently being used in the United States include COMPAS, the Public Safety Assessment ('PSA'), and the Federal Pre-Trial Risk Assessment Instrument ('PTRA'). Many of these tools are also used to inform decisions on parole and sentencing.

These tools have, however, been subject to intense criticism for several reasons. Studies have shown *inter alia* that risk assessments make inaccurate predictions that are no better than those made by non-expert humans. They do not result in a significant reduction in pre-trial detention rates, and that they produce disparate outcomes for different racial groups. US-based NGO Partnership on AI has found that AI risk assessment tools currently being used in the United States are unfit for use in pre-trial assessments, and it has recommended that policymakers cease the deployment of risk assessment tools until such time that the challenges affecting such tools have been adequately addressed.[41]

[39] Marion Oswald et al., 'Algorithmic risk assessment models: lessons from the Durham HART model and 'Experimental proportionality' Information & Communications Technology Law, Vol 27, Issue 2 (2018)
[40] ECtHR, *Barbera, Messegue, and Jabardo v Spain*, App. No. 10590/83 (Judgment of 6 December 1988)
[41] Partnership on AI, 'Report on Algorithmic Risk Assessment Tools in the US Criminal Justice System' (2018)

The adoption of pre-trial risk-assessments tools in the United States has largely been driven by the desire to address high imprisonment rates in the country by making pre-trial decision-making fairer. In particular, these tools have been promoted as an alternative to cash bail - a system often criticised for disadvantaging poorer defendants and worsening social injustices.[42] Cash bail is a relatively rare concept in the EU, but there are concerns about the quality of pre-trial detention decisions in many Member States, which have been criticised for failing to carry out case-specific reviews and fully consider alternatives to detention.[43]

We are currently unaware of any attempts in EU Member States to introduce algorithmic risk assessments to supplement or replace existing pre-trial decision-making processes. However, it is possible that risk-assessment tools will also be recommended as a solution to address the pre-trial detention challenge in Europe, especially given that many of these tools are developed by private companies that actively market their products to governments and local police forces.

Risk-assessment tools are usually designed to assess the likelihood of re-arrest, and/or of failure to turn up to court after being released based on the profiles of the defendant. Based on these assessments, risk assessment tools either assign risk levels to defendants, or they provide direct advice to decision-makers on whether or not the defendant should be released. There is only limited research about the extent to which pre-trial risk-assessment tools influence judges' decisions in practice,[44] but concerns have been raised about the ability of AI systems to recommend detention at all.[45] There is a risk that recommendations made by AI systems to detain individuals compromise the presumption of release. This is a particularly valid concern, in light of research suggesting that decision-makers have a tendency to err on the side of caution when they are 'advised' by AI systems, and that they have a greater propensity to override risk assessment tools to detain, rather than release defendants.[46] Pre-trial detention should always be a measure of last resort, and no risk-assessment can be regarded as human rights compliant, unless it recommends its users to consider detention as a measure of last resort, after all other alternatives have been fully considered.

Pre-trial risk assessment tools in the United States and elsewhere have also been criticised for (unintentionally) over-estimating risks, because of the nature of the data used to train its algorithms.

Pre-trial risk assessment tools typically rely only on data regarding individuals who have been released, and they ignore those who were detained, but would have otherwise 'succeeded' by not being arrested, and by appearing in court.[47] In other words, algorithms are based on the assumption that individuals who have been detained by courts in the past have been rightfully deprived of their liberty. Any AI system developed to assist pre-trial detention decision-making must be designed to give effect to the presumption in favour of release. This means that risk-assessment tools need to be deliberately calibrated to generate outcomes that favourable to the defendant. Data used to train the AI system

[42] National Association of Criminal Defence Lawyers (NACDL), 'Making Sense of Pretrial Risk Assessments' (2018) https://www.nacdl.org/Article/June2018-MakingSenseofPretrialRiskAsses
; Pretrial Justice Institute, 'Pretrial risk assessments can produce race-neutral results. Report' (2017)

[43] Fair Trials, 'Measure of Last Resort' (2016) https://www.fairtrials.org/sites/default/files/publication_pdf/A-Measure-of-Last-Resort-Full-Version.pdf

[44] E.g. Alex Albright. 'If You Give a Judge a Risk Score: Evidence from Kentucky Bail Decisions' (2019)

[45] The Leadership Conference on Civil and Human Rights, 'The Use of Pretrial "Risk Assessment" Instruments – A Shared Statement of Civil Rights Concerns' http://civilrightsdocs.info/pdf/criminal-justice/Pretrial-Risk-Assessment-Full.pdf

[46] NACDL (n 42)

[47] NACDL (n 42)

should be carefully scrutinised so that it reflects the inevitable fact that a significant proportion of individuals in pre-trial detention have been deprived of their liberty in violation of their human rights.

Studies of pre-trial risk-assessment tools used in the United States cast doubt on their effectiveness at reducing pre-trial detention rates, and their ability to make accurate predictions of risks. A study in Kentucky, for example, found that the likelihood of defendants being released within the first three days of their arrest went down after the risk-assessment tool was deployed, and that there were no significant changes in the number of re-arrests and failure-to-appear rates amongst defendants released on bail during the same period.[48] This was the case even after the risk-assessment tool was modified post-deployment to improve the accuracy of predictions. Another study has found that the COMPAS risk-assessment tool is no better at predicting the likelihood of defendants reoffending than non-expert human volunteers.[49] These studies do not necessarily prove that AI systems are incapable of reducing pre-trial detention rates at all, but they do raise questions about their usefulness, and they strongly challenge claims that algorithmic risk-assessment tools help to improve the quality of pre-trial detention decisions. They also highlight the need for post-deployment testing and monitoring of AI systems, to ensure that they have the desired effect of ensuring that individuals are detained only as a measure of last resort.

Post-trial assessment systems are also being increasingly used, such as to assist sentencing decisions or prisoner release.

In England and Wales, the Prison and Probation Services has developed and operates the Offender Assessment System (OASys), an automated risk-assessment tool.[50] It assesses the risk of harm offenders pose to others and how likely an offender is to reoffend, as well as assessing offender needs. These risk assessments are used to decide “interventions” and to influence the sentence plans given to offenders.[51] Millions of these assessments have been carried out.[52] The system collates information on offenders’ previous offences, education, training, employment, alcohol and drug misuse; as well as their “attitudes”, “thinking and behaviour”, “relationships”, and “lifestyle”.[53] This data is used alongside the individual’s offending record and “offender demographic information” to inform two predictive algorithms: OASys General Reoffending Predictor (OGP1) and OASys Violence Predictor (OVP1).[54] A 2014 National Offender Management Service analysis found that the OGP1 and OVP1 generated different predictions based on race and gender. They found that relative predictive validity was better for white offenders than for Asian, black, or mixed ethnicity offenders. The Offender Group Reconviction Scale (OGRS) is another algorithmic risk assessment tool, which is used in England and

[48] Megan Stevenson, ‘Assessing Risk Assessment in Action’, 103 Minnesota Law Review 303 (2018)

[49] Julia Dressel and Hany Farid, ‘The accuracy, fairness, and limits of predicting recidivism’, Science Advances Vol 4, no. 1 (2018)

[50] Prison Service Order, Offender Assessment and Sentence Management – OASys (2005), https://www.justice.gov.uk/downloads/offenders/psipso/pso/PSO_2205_offender_assessment_and_sentence_management.doc; National Offender Management Service, ‘A compendium of research and analysis on the Offender Assessment System (OASys) 2009–2013’ (2014), https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/449357/research-analysis-offender-assessment-system.pdf

[51] Ibid

[52] National Offender Management Service, ‘A compendium of research and analysis on the Offender Assessment System (OASys) 2009–2013’ (2014), https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/449357/research-analysis-offender-assessment-system.pdf

[53] Non-scored categories: Health and other, emotional wellbeing, financial management

[54] Ibid

Wales to assess and predict an offender's likelihood of reoffending.[55] The OGRS algorithm uses data on the individual's official criminal history, as well as their age and gender, to produce a risk score between 0 and 1 of how likely an offender is to reoffend within one or two years.

The use of these AI systems in a post-trial setting, and the documented differences in predictive outcomes based on, among others, race, highlight the clear need for strict testing and monitoring of such systems. These systems used in a post-trial setting could very easily be transferred to a pre-trial risk assessment setting; the principles and aims of these systems and the data used are very similar. For example, the COMPAS system, mentioned above and considered in more detail below, was originally designed as a recidivism risk assessment tool, and it is also used as a pre-trial risk assessment tool. [56]

Where AI systems inform decisions on the deprivations of liberty, they should be calibrated to generate outcomes that favour release, and they should not facilitate detention other than as a measure of last resort. AI systems must be subject to rigorous testing to ensure they have the desired effect of reducing rates of pre-trial detention rates.

AI systems should be designed to be non-discriminatory

One of the most frequent criticisms of AI systems and their use in criminal justice systems is that they can lead to discriminatory outcomes, especially along racial and ethnic lines.

The best-known example of this is the study by the US media outlet ProPublica into COMPAS, a risk-assessment tool designed to predict the likelihood of reoffending in Broward County in Florida. ProPublica found that COMPAS was 77% more likely to rate black defendants as 'high-risk' than white defendants, and it was almost twice as likely to mislabel white defendants as lower risk than black defendants.[57]

The dangers of the failure to adequately regulate the use of AI to prevent discrimination have also been witnessed in Europe. The 'Crime Anticipation System' ('CAS'), a predictive policing software being used across the Netherlands, was initially designed to consider ethnicity as a relevant factor for determining the likelihood of a crime being committed. Amongst the indicators used by CAS to predict crimes in a particular area was the number of '*non-Western allochtones*' in the area – in other words, 'non-Western' individuals with at least one foreign-born parent.[58] The software not only presupposed the existence of a correlation between ethnicity and crime, but also singled out a category of ethnicities to be of particular concern, given that the presence of '*Western*', '*autochtone*' individuals were not used as indicators. Furthermore, given that '*Western*' was defined somewhat subjectively (for example, including individuals of Japanese or Indonesian origin, and including all European nationalities, apart from Turkish), CAS incorporated highly questionable societal categorisations and biases.

[55] Howard et al, 'Offender Group Reconviction Scale' (2017), https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119184256.ch11

[56] Northpointe, 'PractitionersGuide to COMPAS Core', (2015) https://assets.documentcloud.org/documents/2840784/Practitioner-s-Guide-to-COMPAS-Core.pdf

[57] Julia Angwin et al., 'Machine Bias', ProPublica (2016)

[58] Serena Oosterloo and Gerwin van Shie, 'The Politics and Biases of the "Crime Anticipation System" of the Dutch Police' (2018), http://ceur-ws.org/Vol-2103/paper_6.pdf

In the United Kingdom, a major criticism against HART has been that it included data collated and classified by a private company for marketing purposes that could very easily to biased outcomes. HART relied on the 'Mosaic' code developed by a consumer credit reporting company, that categorised individuals into various groups according to *inter alia* their ethnic origin, income, and education levels. It was of particular concern that some socio-demographic categories used by Mosaic were blatantly racialised, including, for example, 'Asian Heritage', which stereotyped individuals of 'Asian' origin as being unemployed or having low-paid jobs, and living with extended families.[59]

In Denmark, an automated algorithmic assessment has been used to classify different neighbourhoods, based on criteria such as unemployment, crime rates, educational attainment, and other 'risk indicators', as well as whether the levels of first and second-generation migrants in the population is more than 50%. Neighbourhoods which meet these criteria are classified as 'ghettos'. These neighbourhoods are then subject to special measures, including higher punishments for crimes.[60] It is clearly discriminatory, as well as entirely unfair, for people living in certain areas to be punished more severely than others in different areas for the same crimes.

Further examples of criminal justice AI which have been identified as producing discriminatory outcomes include the previously mentioned OASys, NDAS and the Gangs Matrix in the UK, and the Netherland's ProKid 12.

These examples illustrate the need for regulations to ensure that AI systems are designed to be non-discriminatory, and that they exclude categorisations and classifications that deepen and legitimise social biases and stereotypes. However, policy makers should not assume that making AI systems blind to all protected characteristics will always help to produce non-discriminatory outcomes. In certain scenarios, the removal of protected characteristics from the data could worsen discrimination. For example, it has been suggested on the basis of research into COMPAS in the United States, that excluding gender as a variable for risk assessments would fail to reflect a well-established statistical fact that in most countries, women are less likely to reoffend than men.[61] Making COMPAS gender-blind would unfairly and inaccurately assume women to be as equally likely to reoffend as men, and discriminate against them by overestimating their risk scores.

Removing visible biases from AI systems cannot be the sole or primary solution to their discriminatory impact also because AI systems can be biased even if they have not been deliberately designed in that way. Bias is often unintentional, and even if the AI system appears on the surface to be neutral, their algorithms can lead to discriminatory assessments and outcomes. COMPAS, for example, does not include race or ethnicity as a variable, yet research has found that it consistently gives black defendants higher risk scores than their white counterparts, making them less likely to be released from detention.[62]

[59] Big Brother Watch, 'Written evidence on algorithms in the justice system for the Law Society's Technology and the Law Policy Commission' (2019), https://bigbrotherwatch.org.uk/wp-content/uploads/2019/02/Big-Brother-Watch-written-evidence-on-algorithms-in-the-justice-system-for-the-Law-Societys-Technology-and-the-Law-Policy-Commission-Feb-2019.pdf

[60] Algorithm Watch, 'Automating Society' (2019), https://algorithmwatch.org/en/automating-society-denmark/

[61] Nicol Turner Lee, Paul Resnick, and Genie Barton, (2019) 'Algorithmic bias and mitigation: Best practices and policies to reduce consumer harms' (https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms)

[62] Sam Corbett-Davies et al. 'Algorithmic decision making and the cost of fairness' (2017), arXiv:1701.08230

Hidden biases can arise in AI systems in numerous ways. Although a comprehensive analysis of how they can cause unintentional biases are beyond the scope of this paper,[63] the way in which AI systems are themselves created and built illustrate the difficulty, complexity, and sometimes impossibility, in preventing discriminatory outputs and effects of AI systems.

There are fundamental issues with the way AI systems are designed and created which can lead to bias. Where the AI system is based on machine-learning, biases can result from faults in the data that is used to train its algorithms. Machine learning systems 'learn' how to make assessments or decisions on the basis of its analysis of data to which it has previously been exposed. However, the data used to train the machine learning system might be incomplete, inaccurate, or selected for improper reasons, and this could lead to AI systems producing unwanted outcomes. What amounts to appropriate, good quality data for the purpose of training algorithms depends on what the machine learning system is being designed to do,[64] so it might not always be obvious what dataset is needed to train algorithms to be non-discriminatory.

AI designed or created for use in the criminal justice system will almost inevitably use data which is heavily reliant on, or entirely from within the criminal justice system itself, such as policing or crime records, which do not represent an accurate record of criminality, but merely a record of policing - the crimes, locations and groups that are policed within that society, rather than the actual occurrence of crime. The data might not be categorised or deliberately manipulated to yield discriminatory results, but it may reflect the structural biases and inequalities in the society which the data represents.

Where there are discriminatory policing patterns targeting certain demographics, or the systematic under-reporting and systematic over-reporting of certain types of crime and in certain locations,[65] the use of such data merely results in a reinforcing and re-entrenching of those inequalities and discrimination in criminal justice outcomes. For example, according to UK crime data, black people are over 9 times more likely to be stopped and searched than white people,[66] and black men are more than 3 times more likely to be arrested than white men.[67] Despite these statistics, NDAS (mentioned above) in the United Kingdom explicitly relies on stop and search data to determine an individual's propensity to commit a criminal offence. The fact that stop and search is disproportionately used against black people means that there will inevitably be an overrepresentation of black people in NDAS and that their risk levels will be inflated in comparison to white people.

[63] A full analysis can be found in Frederik Zuiderveen Borgerius, 'Discrimination, artificial intelligence, and algorithmic decision-making', Council of Europe (2018)

[64] Fundamental Rights Agency, 'Data quality and artificial intelligence – mitigating bias and error to protect fundamental rights' (2019)

[65] Lum, Kristian, and William Isaac. 2016. 'To Predict and Serve?', Significance 13 (5): 14–19, https://rss.onlinelibrary.wiley.com/doi/full/10.1111/j.1740-9713.2016.00960.x; Bennett Moses, L., & Chan, J. (2016). 'Algorithmic prediction in policing: Assumptions, evaluation, and accountability'. *Policing and Society*. https://www.tandfonline.com/doi/10.1080/10439463.2016.1253695; Barocas, S. and Selbst, A.D., 2016. 'Big Data's disparate impact'. *California Law Review*, 104, 671. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2477899

[66] UK Government Stop & Search facts and figures, February 2019: https://www.ethnicity-facts-figures.service.gov.uk/crime-justice-and-the-law/policing/stop-and-search/latest

[67] Ministry of Justice, 'Black, Asian and Minority Ethnic disproportionality in the Criminal Justice System in England and Wales', 2016, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/639261/bame-disproportionality-in-the-cjs.pdf

Comparable statistics on stop and search are not available in most EU Member States, where the official collection of racially disaggregated criminal justice data is either forbidden by law, or not standard practice. However, recent studies show that racially biased policing practices are prevalent throughout the EU. Data collected from a survey by the Fundamental Rights Agency, for example, has shown that during a 5 year period, 66% of individuals of Sub-Saharan African origin in Austria, and over half of respondents of South Asian origin in Greece were stopped and searched.[68]

AI built on data embedded with such biases to assist, inform, or make decisions in the criminal justice system can expand and entrench the biases represented in the data.[69] When AI systems result in criminal justice outcomes which repeat the discrimination inherent in the historic data, such as targeting individuals from a particular demographic, that decision will itself be preserved in the data. This leads to self-perpetuating 'feedback loops' which reinforce patterns of inequality.[70]

Another way in which AI systems can produce unintentional biases is by way of proxies. Data used by AI systems might be classified in seemingly legitimate ways, but those classifications can sometimes act as proxies for protected characteristics. A common example used to illustrate this point is how home addresses or postcodes can be proxies for race or ethnicity.[71] Certain AI systems, such as HART, were initially trained to find correlations between home addresses and the risk of reoffending – in other words, to identify which postcode areas have 'higher-risk' residents than others.[72] This approach overlooks the fact that there is very pronounced ethnic residential segregation in many countries,[73] making it highly probable in practice, for AI systems to inadvertently establish a link between ethnic origin and risk.

Roma are especially vulnerable to this form of proxy discrimination, given that in many EU Member States, Roma are reported to live primarily in segregated areas inhabited mostly or exclusively by Roma.[74]

There are several ways in which AI systems can be designed to mitigate the risks of discrimination, including by identifying and excluding data classifications that act as proxies for protected characteristics.[75] However, it can be difficult in practice to identify which variables are proxies for protected characteristics (and how they do so), and removing too many 'offending' variables might result in the AI system losing much of its functional utility.[76] There is no one-size-fits-all method of

[68] Fundamental Rights Agency, 'EU-MIDIS II Second European Union Minorities and Discrimination Survey' (2018), https://ec.europa.eu/knowledge4policy/dataset/ds00141_en

[69] Lum, Kristian, and William Isaac. 2016. 'To Predict and Serve?' Significance 13 (5): 14–19, https://rss.onlinelibrary.wiley.com/doi/full/10.1111/j.1740-9713.2016.00960.x

[70] Ibid; Ensign et al, (2017) 'Runaway Feedback Loops in Predictive Policing', Cornell University Library, 29 June 2019 (https://arxiv.org/abs/1706.0984); Lyria Bennett Moses & Janet Chan (2018) Algorithmic prediction in policing: assumptions, evaluation, and accountability, Policing and Society, 28:7, 806-822, https://www.tandfonline.com/doi/10.1080/10439463.2016.1253695

[71] Fredreik Zuiderveen Borgesius, 'Discrimination, artificial intelligence, and algorithmic decision-making' (2018), Directorate General of Democracy, Council of Europe

[72] Marion Oswald et al., (2018) 'Algorithmic risk assessment policing models: lessons from the Durham HART model and 'Experimental' proportionality', Information & Communications Technology Law. 27:2, 223-250

[73] E.g. Sweden. See Bo Malberg (2018) 'Residential Segregation of European and Non-European Migrants in Sweden', Eur J Popul 34(2): 169-193

[74] FRA, 'Summary Report – The State of Roma and Traveller Housing in the European Union – Steps towards Equality' (2010)

[75] Information Commissioner's Office ('ICO'), (2019) 'Human bias and discrimination in AI systems', https://ico.org.uk/about-the-ico/news-and-events/ai-blog-human-bias-and-discrimination-in-ai-systems/

[76] Ibid.

ensuring that AI systems do not produce discriminatory outcomes. Different approaches to de-biasing AI systems can conflict with one another, and the suitability of a particular de-biasing method might depend on the AI tool itself, and the legal and policy context in which it is designed to operate.[77] Biases in AI systems are often not easy to detect, and in many cases, it might also be difficult to pinpoint flaws either in the system itself, or in the training data that has been caused the bias. The structural bias within data that AI systems are built and operated on, something which is particularly deep-rooted in criminal justice data, is a fundamental issue, and which is likely to result in AI systems being fundamentally inoperable – both because the bias makes them morally and ethically inoperable, if not yet legally, and any attempts to remove the bias will make the data to operate these systems unusable.

Fair Trials' view is that the only effective way in which AI systems can be regarded as non-discriminatory is if they have been subject to rigorous independent testing for biases. These tests must be mandated by law, must be independently run, must have clearly stated aims or objectives and must be carried out pre-deployment to reduce the likelihood of individuals being affected by discriminatory profiling and decisions. AI can be tested using test data; datasets which are either synthetic,[78] or using historic data with permissions, and running them through an AI system, analysing the outputs.[79] For example, a trial of retrospective facial recognition video analysis is being run by a police oversight Ethics Committee in the UK. The trial is using historic data - CCTV footage – as the basis for simulated investigations in a controlled environment, monitored by researchers. The trial has clearly stated aims and signifiers of success, and all outcomes will be examined. There are significant human rights, data protection and ethical concerns involved with this particular technology, including the right to privacy, and the testing is not being independently done as it should be, but as above, there are positive aspects of the testing methodology.[80]

An alternative could be to 'test' a system in a strictly academic sense by running it alongside actual criminal justice processes, but with the system not having any effect on decision-making, and analysing the system's proposed decisions or outcomes for bias.

AI should never be used or even 'tested' in real-world situations where they have actual effects on individuals or criminal justice outcomes, before they have been tested. These types of tests also need to be carried out in the broader context of an AI governance framework that not only analyses the potential impact of the AI system pre-deployment, but also continues to monitor its impact afterwards.

If these tests are not carried out, and/or if an AI system cannot be proven to be non-discriminatory, it should be legally precluded from deployment. However, as explained in the final section of this paper, it is questionable whether such tests are feasible in many Member States, where local laws prohibit the collection of racially-disaggregated data.

AI systems should be developed to generate non-discriminatory outcomes, ensuring that suspects and accused persons are not disadvantaged, either directly or indirectly, on account of their

[77] Partnership on AI (n 41), ICO (n 75)

[78] https://towardsdatascience.com/reducing-ai-bias-with-synthetic-data-7bddc39f290d

[79] https://www.theguardian.com/technology/2016/dec/19/discrimination-by-algorithm-scientists-devise-test-to-detect-ai-bias; https://arxiv.org/abs/1610.02413

[80] West Midlands Police Office of the Police and Crime Commissioner, 'Research Protocol: Using CCTV in Police Investigations: A comparison of facial recognition technology-assisted reviews and manual reviews' (2020), https://www.westmidlands-pcc.gov.uk/wp-content/uploads/2020/05/26032020-EC-Item-3-Trial-Protocol.pdf?x56534; and West Midlands Police Office of the Police and Crime Commissioner, 'Retrospective Assisted Facial Recognition Trial on Historic Criminal Case Data', (2020) https://www.westmidlands-pcc.gov.uk/wp-content/uploads/2020/05/26032020-EC-Item-3-AFR-Summary.pdf?x56534

protected characteristics, including race or ethnicity. AI systems should be subject to mandatory testing before and after deployment so that any discriminatory impact can be identified and addressed. If an AI system cannot be proven not to generate discriminatory outcomes, it should not be used.

AI Systems need to be transparent and explainable

AI systems can have a significant influence over criminal justice decisions, and they should be open to public scrutiny in the same way that all decision-making processes by public entities should be. However, a common criticism of many AI systems is that they lack transparency, which often makes it difficult, if not outright impossible, to subject them to meaningful impartial analysis and criticism. This lack of transparency is both as a result of deliberate efforts to conceal the inner workings of AI systems for legal or profit-driven reasons, and of the nature of the technology used to build AI systems that are uninterpretable for most, if not all humans.

There are several reasons why it is necessary for AI systems to be transparent. Firstly, transparency is essential for strengthening confidence of both primary users of the system, as well as the general public in AI systems. Democratic values demand that the public needs to be aware of how powerful public institutions, such as the police and the judiciary, operate so that they can be held accountable for their actions. It is also crucial for primary users of AI systems to understand how they work, so that they can make informed decisions about how much influence they should have on criminal justice decisions.

Secondly, decisions made by AI systems need to be contestable at an individual level. Standards on the right to a fair trial and the right to liberty demand that defendants should have access to materials that inform decisions regarding them, so that they can challenge the accuracy and lawfulness of those decisions.

Transparency also acts as a safeguard against bias and inaccuracies. It is difficult to imagine how issues that undermine the fairness and accuracies of AI systems (such as racial biases) can be detected and ultimately, fixed, if they cannot be properly accessed and analysed. As explained above, certain AI systems, such as CAS, have been found to have serious, but very obvious flaws. In CAS's case, however, the fault in the software could be detected easily, which meant that the discriminatory impact of the tool could be mitigated. The indicator for '*non-Western allochtones*' in CAS was removed in 2017,[81] ostensibly because it served no useful purpose, but presumably also because of the very obvious bias. This mitigation was possible because CAS is a transparent software, that was developed in-house by the Dutch police. The types of indicators used to predict crime were made openly available, and information about the method by which the software made predictions could easily be accessed and understood.[82]

This, however, is not the case for all AI systems, because AI systems are often developed by for-profit companies with little to no meaningful input from the public. As such, details of how they are designed, and how they make decisions and assessments are, in many cases, closely guarded as trade secrets that are protected by law.[83] Often, AI systems are 'black boxes' because they are deliberately

[81] Ibid.

[82] Ibid.

[83] Taylor R Moore, 'Trade Secrets and Algorithms as Barriers to Social Justice', Center for Democracy & Technology (2017), https://cdt.org/files/2017/08/2017-07-31-Trade-Secret-Algorithms-as-Barriers-to-Social-Justice.pdf

kept that way. While it is accepted that strong, enforceable intellectual property laws are needed to promote advancements in what is a very dynamic field of scientific research and innovation, it is not acceptable that these concerns trump the rights of individuals suspected or accused of crimes. In light of this, it is concerning that the Commission's White Paper focuses on, and strongly promotes, the concept of a "partnership between the private and the public sector" in relation to AI.[84] Fair Trials appreciates that effective public-private collaboration could help to fill in gaps in public sector expertise and capacity for the development of AI systems, but given the transparency challenges, it is essential that such partnerships are accompanied by robust regulations and rules that ensure effective and open scrutiny.

However, even if AI systems are completely exposed to public scrutiny, and their source code[85] and input data, for example, are openly disclosed, there is still no guarantee that they will be sufficiently transparent to enable adequate independent scrutiny. AI systems can be black boxes by nature of the technology that makes their decision-making processes complicated beyond comprehension for most (in some cases, too complicated even for computer scientists to understand).[86] This is especially the case where AI systems are based on machine-learning algorithms.

One possible reason for the unintelligibility of AI systems is that they sometimes use machine-learning algorithms that are simply too complex to be understood to a reasonable degree of precision.[87] This is especially the case where AI systems incorporate 'Deep Neural Networks' – a machine-learning algorithmic architecture inspired by the structure and mechanics of human brains. Rather than relying on a set of man-made instructions, these types of AI systems make decisions based on experience and learning. Decision-making processes of this kind have been described to be 'intuitive', because they do not follow a defined logical method, making it impossible to analyse the exact process by which a particular decision is reached.[88] It has also been suggested that some AI systems are uninterpretable to humans because the machine-learning algorithms that support them are able to identify and rely on geometric relationships that humans cannot visualise. Certain machine-learning algorithms are able to make decisions by analysing many variables at once, and by finding correlations and geometric patterns between them in ways that are beyond the capabilities of human brains.[89]

Given these challenges, there is widespread recognition that states should require AI systems to not only be 'transparent', but also explainable and intelligible.[90] GDPR already recognises that individuals should have the right to an explanation of how a decision was reached, if they have been subject to an automated decision.[91] In principle, this is an essential and very useful requirement, but it is also one that seems difficult to implement in practice, given that both 'explainability' and intelligibility are highly subjective concepts. Arguably, AI systems' computing processes are inherently difficult to explain and understand for most people, including for most criminal justice decision-makers, but this

[84] European Commission, 'On Artificial Intelligence – A European approach to excellence and trust', Brussels, 19.2.2020 COM(2020) 65 final

[85] Computer programming codes that are readable to humans

[86] Royal Society, 'Explainable AI: the basics', https://royalsociety.org/-/media/policy/projects/explainable-ai/AI-and-interpretability-policy-briefing.pdf

[87] Yavar Bathaee, 'The Artificial Intelligence Black Box and the Failure of Intent and Causation', Harvard Journal of Law & Technology, vol 31, no. 2, 890 (2018)

[88] Ibid.

[89] Ibid.

[90] Toronto Declaration, Art 32

[91] GDPR, Recital 71

surely should not be the sole basis for oversimplifying the technology, or for banning the use of AI outright.

Computer scientists have been theorising different ways of ensuring that decisions made through complex algorithms can be explained and understood. An example is the 'explainable AI' movement ('xAI') that aims to build AI systems that can show more discernible links between inputted data and decisions. xAI systems measure how each input influences the final decision, so it is possible figure out how much weight is given to each input.[92] This seems to be an innovative response to the 'black box' challenge, establishing clearer, more helpful relationships between inputs and final decisions. However, it appears to fall short of explaining what happens between data being inputted into the system and the final decision, and it does not enable users to impute any logic to the decision-making process.[93]

As explained above, there are various reasons why AI systems need to be transparent and intelligible, but the effective of exercise of the rights of the defence must be recognised as a crucial test for determining whether an AI system is sufficiently explainable and intelligible. AI systems have to be designed in a way that allows criminal defendants to understand and contest the decision made against them. Partnership for AI has suggested that a central factor that determines contestability of AI systems is the possibility of carrying out an audit trail of the AI decision.[94] In particular, it has to be possible for an auditor to follow and reproduce the process and come to the same conclusion reached by the AI system at the end.

Furthermore, as explained in further detail below, criminal justice procedures should require the full disclosure of all aspects of AI systems that are necessary for suspects and accused persons to contest their findings, and this disclosure should be in a form which is understandable to a layperson, without the need for technical or expert assistance.

AI systems need to be transparent and explainable, so they can be understood and scrutinised by their primary users, suspects and accused persons, as well as the general public. Commercial or proprietary interests, or technical concerns, should never be a barrier to transparency. AI systems must be designed in a way that allows criminal defendants to understand and contest the decision made against them. It should be possible to carry out an independent audit, and processes should be reproducible.

[92] Kartik Hosanagar and Vivian Jair, 'We Need Transparency in Algorithms, But Too Much Can Back Fire', Harvard Business Review, (2018) https://hbr.org/2018/07/we-need-transparency-in-algorithms-but-too-much-can-backfire

[93] Brent Mittelstadt et al, 'Explaining the Explanations of AI' (2018), arXiv:1811.01439v1 [cs.AI]

[94] Partnership on AI (n 41)

Part 2: Safeguards for the use of AI Systems in Criminal Proceedings

AI systems have to be built in accordance with human rights principles, and to give effect to human rights in practice, but it is unlikely that their design alone will guarantee that they are used in ways that comply with human rights. Regulatory frameworks for the design and deployment of AI systems have to be accompanied by appropriate legal safeguards that ensure they are used responsibly and lawfully. There are two primary questions that need to be addressed:

1) how procedural rules ensure that decision-makers do not over-rely on AI systems; and
2) how decisions and assessments made by AI systems can be analysed independently and challenged.

Combatting 'Automation Bias' and Reinforcing Meaningful Human Input

One of the main challenges of automated, or semi-automated decision-making systems is that of 'automation bias' – the tendency to over-rely on automation in ways that can cause errors in decision-making. Automation bias occurs primarily due to the perception that automated decision-making processes are generally trustworthy and reliable. Automated cues have been found to be particularly salient to decision-makers, and research has shown that users of automated decision-making systems have a tendency to place greater weight on automated assessments over other sources of advice.[95] The disproportionate influence of automated systems can undermine the quality of decision-making, by discouraging its users from consulting a wider range of factors that could inform more accurate decisions.

Most AI systems currently being used to assist criminal justice decision-making do not completely replace human decision-making. They are instead designed and deployed to be used as decision aids, whose outputs are factored into consideration for the purposes of human decision-making. The phenomenon of automation bias however, raises questions about whether AI systems are being used in reality in accordance with their intended purpose as decision aids, and not as *de facto* replacements for human decision-making processes.

There is strong evidentiary basis for automation bias amongst pilots who, like judges and other decision-makers in criminal justice proceedings, have typically been through a high level of training to make appropriate decisions in highly complex settings.[96] However, limited research into automation bias amongst judges suggests that AI systems might have a more complex impact on judges' behaviour. For example, a study conducted in 2019 in Kentucky seems to suggest that the degree to which judges rely on predictive tools for pre-trial detention decision-making could be influenced by the ethnicity of the defendant.[97] The research indicates that judges had a greater tendency to rely on algorithmic risk assessments where the defendant was white, whereas in cases where the defendant was black, judges were more likely to overrule the risk-assessment in favour of detaining them. This study appears to show that AI systems can influence judges' behaviour in unpredictable ways, especially where there are interactions or conflicts between automation and human biases, and that AI systems might be an ineffective tool for challenging human prejudices.

It is crucial that rules governing the use of AI systems in criminal proceedings actively try to counter automation bias, and encourage decision-makers to make independent determinations. A simple

[95] Raja Parasuraman and Dietrich Manzey, 'Complacency and Bias in Human Use of Automation: An Attentional Integration', Human Factors, The Journal of Human Factors and Ergonomics Society (2010)
[96] Ibid.
[97] Alex Albright, 'If You Give a Judge a Risk Score: Evidence from Kentucky Bail Decisions', Harvard University (2019) https://thelittledataset.com/about_files/albright_judge_score.pdf

requirement to have a human decision-maker 'in the loop' or to have a human decision-maker review or check the automated decision is insufficient, because this risks overestimating the capacity or willingness of human decision-makers to question and overrule automated decisions. A mere requirement to have an automated decision reviewed by a human, on its own, could reduce the human review into a rubber-stamping exercise which, in practice, is no oversight at all.

In recognition of this challenge, the European Data Protection Board has recommended that in order for decisions to be regarded as *not* 'based solely' on automated processing for the purposes of Article 22 GDPR, there has to be 'meaningful' human oversight, rather than just a token gesture.[98] What qualifies as 'meaningful' intervention is open to interpretation, and it is likely to differ depending on the circumstances and the type of decision being made. In the context of criminal justice procedures, where decisions often have particularly severe and far-reaching implications for individuals' rights, safeguards for ensuring meaningful human intervention have to be especially robust.

Procedural safeguards that ensure 'meaningful' human oversight

Rules governing the use of AI systems in criminal justice proceedings have to counter automation bias by encouraging human decision-makers to treat their processes with scepticism, and to force them to challenge and scrutinise the outcomes of algorithmic assessments.

Procedural safeguards that can be put in place to tackle automation bias include:

a) making it a legal requirement for decision-makers to be adequately alerted and informed about the risks associated with AI systems;
b) making AI systems' assessments intelligible to decision-makers;
c) requiring decision-makers to provide full, individualised reasoning for all decisions influenced by an AI system; and
d) making it easier for decision-makers to overrule AI assessments that produce unfavourable outcomes for defendants.

One way of ensuring that automated assessments and decisions might be to ensure that decision-makers are sufficiently informed and alerted about the risks of relying on AI systems. This seems to be the approach taken by the Wisconsin Supreme Court in the United States in the case of *Loomis*,[99] in which the Court considered whether or not the use of the COMPAS risk assessment tool for sentencing purposes violated due process rights. The judgment in *Loomis* recognises the importance of procedural safeguards as a way of safeguarding fairness of decisions, by requiring the use of 'written advisements' to alert decision-makers about the potential risks of AI risk assessments. Specifically, the court mandated that these advisements had to include warnings that a) the process by which the COMPAS produces risk scores were not disclosed due to its 'proprietary nature'; b) the accuracy of risk scores are undermined by the fact that COMPAS relied on group data; c) the risk-assessment tool had never been tested locally for accuracy; d) 'questions' have been raised about the discriminatory effect of COMPAS risk-assessments; and e) COMPAS was developed to inform *post*-sentencing decisions, but not sentencing decisions themselves.

These warnings are clearly very specific to COMPAS and the context in which it is used in Wisconsin. If similar safeguards were adopted in different contexts and with regard to different AI systems, advisements will no doubt need to be adapted. The warnings used in *Loomis* have, however, been

[98] Article 29 Data Protection Working Party, 'Guidelines on Automated individual decision-making and profiling for the purposes of Regulation 2016/679' (3 October 2017)
[99] *State v. Loomis*, 881 N.W.2d 749 (Wis. 2016)

criticised because they do not give enough information to decision-makers to enable them to appreciate the degree to which these risk-assessments should be discounted.[100] In particular, the advisements are silent on the strength of the criticisms against COMPAS, and they say nothing about the basis on which questions about their discriminatory effect have been raised.[101] These warnings also give no indication about likely margin of error of the assessment, so although judges are informed that some assessments might be inaccurate, they are not in a position to appreciate how serious or frequent these errors might be.

'Advisements', or warnings that encourage decision-makers to be sceptical of AI systems cannot be considered as effective safeguards, unless they contain sufficiently helpful information for decision-makers. However, even if judges are given stronger warnings than those in the *Loomis* advisements, it is still doubtful whether they alone will adequately mitigate automation bias. One reason for this is that many criminal justice decisions (such as pre-trial detention decisions) are, in practice, made very routinely by judges. Although written advisements might initially help judges think more critically about automated risk assessments, over time, these advisements could become repetitive and routine, and lose much of the intended meaning and effect.[102]

An effective safeguard that could work in conjunction with mandatory warnings could be for decision-makers to be given a better insight into how AI systems produce a particular assessment or calculation. As mentioned above, the lack of information about how assessments are made by AI systems makes it harder for criminal defendants to scrutinise and challenge them. Surely, this has to be true also for decision-makers. It is much harder, if not impossible, to analyse and criticise decisions if there is no reasoning behind them. While AI systems do not rely on 'reasoning' *per se*, information given to decisions about how a specific assessment was made, including what factors were relevant, and how much weight was given to each factor could give decision-makers more confidence to decide whether to agree or disagree with an AI-generated decision.

Decisions or assessments made by AI systems cannot be the sole basis of criminal justice decisions – they should be no more than a factor that can influence human-decision making. As such, decision-makers should be required to show that decisions were influenced by a broader range of factors other than the AI system, by way of fully reasoned, case-specific, written decisions. Research has shown that the lack of case-specific reasoning in pre-trial detention decisions is already a serious challenge in many EU Member States,[103] and AI systems risk worsening the standardisation of such decision-making processes. Where AI systems are used to inform pre-trial detention decisions, or any other criminal justice decision that has a significant impact on the rights of the defendant, reasoned decisions must be specific to the defendant's case, and in particular, they must reveal what factors influenced the decision, and to what degree. In particular, decisions have to make it clear how much weight was given to assessments by AI systems.

It is also crucial that decision-makers are able to override decisions made by AI systems, and that they are confident about doing so where the tool produces assessments or recommendations that are unfavourable to the defendant (e.g. where the AI system advises against releasing the defendant). It has been reported that members of the police force in Avon and Somerset Police in the United Kingdom are expected to record incidences where they have disagreed with assessments made by a

[100] Harvard Law Review, '*State v. Loomis* – Wisconsin Supreme Court Requires Warning Before Use of Risk Assessments in Sentencing' 130 Harv L Rev 1530 (2017)
[101] Ibid.
[102] Partnership for AI, (n 41)
[103] Fair Trials (n 34)

predictive policing tool, and to explain their reasons for the disagreement.[104] This is likely to act as a strong disincentive for overriding decisions made by the AI system, and as such, it actively facilitates automation bias. Furthermore, it seems to interfere with the presumption of innocence by making it difficult for decision-makers to override AI systems to make decisions that favour the defendant. If an AI system recommends the arrest or the detention of an individual, decision-makers should feel that they have a genuine choice of overruling the AI system, and not be pressured into compliance. Criminal justice decision-making processes should, as a general rule, be skewed in favour of the defence to give effect to the presumption of innocence, and rules governing the use of AI systems should favour favourable outcomes for defendants.

On the other hand, in cases where a decision-maker acts against the advice of an AI system that recommends a favourable outcome for the defendant, there should be a requirement for reasons to be given for their decision. This is to prevent unfavourable outcomes for defendants that are motivated by improper reasons, and to mitigate the risk of unconscious bias.

Challenging AI in criminal proceedings

AI systems need to be contestable by criminal defendants. This is so that they can not only challenge the outcomes of the AI systems' calculations and analyses, but also scrutinise the legality of their use. In other words, being able to challenge AI systems in criminal proceedings is not only a procedural fairness requirement for defendants, it is also a means by which legal standards governing AI systems and their use can be enforced.

One of the major issues preventing the sufficient contestability of AI systems in criminal proceedings is the lack of notification. If an individual is not notified that they have been subject to an AI system automated decision, they will not have the ability to challenge that decision, or the information that the decision was based on.

For example, in the United Kingdom the Data Protection Act 2018 sets out the applicability of the GDPR and sets out the UK's interpretations on the GDPR's requirements and safeguards. Section 14 of the Data Protection Act significantly dilutes the requirements of Article 22 of the GDPR, permitting purely automated decisions which have legal or similar significant effects on a data subject, in absence of the subject's consent, as long as the data subject is *notified* that a purely automated decision has been taken about them, after the decision has been made. The data subject then has the opportunity to request a new decision.

However, despite the operation of such AI systems in the UK, for example, the HART system described above, which influences whether an individual is held in custody or released onto a rehabilitation programme, it has been reported that individuals are not notified that they have been subject to such an automated decision, assistive or otherwise.[105] This is likely because under the Data Protection Act 2018, automated decisions which have legal or similar significant effects on a subject are not necessarily classified as "purely automated" if a human has administrative input. In order to meet this requirement, the human input can be as minimal as checking a box to accept the automated-decision, which may have significant impact on an individual, such as holding them in custody - and the

[104] Lina Dencik et al., 'Data Scores as Governance: Investigating uses of citizen scoring in public services', Data Justice Lab, Cardiff University (2018)

[105] Big Brother Watch, 'Big Brother Watch submission to the Centre for Data Ethics and Innovation: Bias in Algorithmic Decision-Making (Crime and Justice)' (2019), https://bigbrotherwatch.org.uk/wp-content/uploads/2019/06/Big-Brother-Watch-submission-to-the-Centre-for-Data-Ethics-and-Innovation-Bias-in-Algorithmic-Decision-Making-Crime-and-Justice-June-2019.pdf

requirement to notify and ability to request a new decision are nullified. In this way, systems such as HART continue to be used, with people subject to its decisions completely uninformed.

Therefore, defendants and/or their representatives must always be notified where there has been an automated decision-making system involved, assistive or otherwise, which has or may have impacted the defendant pre-arrest, pre-custody, and pre- or post-trial.

This notification should include clear and comprehensible information about the decision that has been taken, how that decision was reached, including details of the information or data involved in reaching that decision, what the result or outcomes of the decision are and what effects, legal or otherwise they have, and information on how to challenge that decision.

As discussed in the previous section, a further major barrier to the contestability of AI systems is a technical one. The 'black box' nature of certain AI systems can be largely attributed to their design, so it is important that there are rules governing the interpretability of these systems so that when they are in use, their processes can be understood at all. However, there are also legal barriers to the full disclosure of AI systems, which are often put in place to protect commercial interests. Procedural safeguards play a particular important and effective role in addressing these types of opacity challenges.

Transparency is a fundamental aspect of an adversarial process that underpins the right to a fair trial, and human rights standards require that as a general rule defendants should be given unrestricted access to their case-file,[106] and to be given the opportunity to comment on the evidence used against them.[107] These standards are further reinforced by Directive 2012/13/EU,[108] which requires Member States to grant access to all material evidence in possession of the competent authorities to the defence to safeguard the fairness of the proceedings and to enable defendants to prepare their defence.[109] The procedural requirement of an adversarial process is not one that is limited to substantive criminal proceedings – it also applies in the context of pre-trial decision-making processes, especially for decisions on the deprivation of liberty.[110] While EU law and international human rights law also recognise that there might be certain justifications for non-disclosure of materials used against the defendant in criminal proceedings, but these are narrow restrictions, and commercial interests are not regarded as a valid justification.[111] Furthermore, EU law does not explicitly recognise any derogations from the right of access to materials that are essential to challenging the lawfulness of an arrest or detention.[112] In order for Member States to comply with these standards, any exceptions to the disclosure of information regarding AI systems have to be applied very narrowly.

It must also be recognised that barriers to scrutiny and accountability of AI systems are not only legal, but also technical. As explained in previous sections, many AI systems suffer from interpretability issues because of their design and by the nature of the machine-learning technology upon which they rely. In the absence of specific expertise on AI, it is difficult to imagine how, in practice, defendants and their lawyers will be able to challenge AI systems.

[106] ECtHR, *Beraru v Romania* App. No. 40107/04 (Judgment of 18 March 2014)

[107] ECtHR, *Kuopila v Finland*, App. No. 27752/95 (Judgment of 27 April 2000)

[108] Directive 2012/13/EU of the European Parliament and of the Council of 22 May 2012 on the right to information in criminal proceedings ('Access to Information Directive')

[109] Access to Information Directive, Article 7(2)

[110] Access to Information Directive, Article 7(1); ECtHR, Wloch v Poland, App. No. 27785/95 (Judgment of 19 October 2000)

[111] Access to Information Directive, Article 7(1),(2), and (4)

[112] Ibid., Article 7(1)

Another possible solution to this challenge, as explained below, is training for defence lawyers – but it is unreasonable to expect lawyers to develop expertise that would enable them to analyse and scrutinise AI systems at a technical level. A further solution could be that defence lawyers have access to the relevant expertise from suitably qualified professionals.

Criminal justice procedures should require the notification to suspects and accused persons where an AI system has been used which has or may have impacted a decision made about that individual. Procedures should enable the full disclosure of all aspects of AI systems that are necessary for suspects and accused persons to contest their findings and this disclosure should be in a form which is comprehensible to a layperson, without the need for technical or expert assistance. Suspects and accused persons should also be given effective access to technical experts who can help to analyse and challenge otherwise incomprehensible aspects of AI systems.

Training

AI systems use technology not well understood by many people. Without proper training, outputs of AI systems might not be easy to interpret, and it might be difficult to appreciate what factors undermine the reliability of AI systems, so that appropriate weight can be attached to their findings. As mentioned above, decision-makers can be warned about the weaknesses of AI systems as part of their decision-making process, but the effectiveness of this safeguard can be questioned, because it is unlikely to provide decision-makers with all the information they need, and there is no guarantee that the warnings will be taken seriously in all cases.

Training is not just needed for the primary users of AI systems, such as judges and police officers who use them to inform their own decisions. The training must also be available criminal defence lawyers, so that they are in a better position to challenge AI systems, where necessary. If AI systems are used routinely to aid criminal justice decisions or even made mandatory (as is the case in certain states in the United States), there would be strong justification for governing bodies to make training on AI mandatory for criminal justice practitioners.

Part 3: Governance and Monitoring

Criminal justice processes are an important enforcement mechanism for ensuring that AI systems are designed and used lawfully, but they cannot be the sole, or even the primary means of implementing legal and ethical standards. Of equal, if not greater importance is a framework that ensures that policy decisions on the design and deployment of AI systems are made in systematised way, and that unlawful or harmful AI systems never enter into public service. Member States that deploy AI systems for criminal justice purposes should have regulatory mechanisms that are fit for purpose. At a minimum, these should include frameworks for a) pre-deployment impact assessments; b) post-deployment monitoring and evaluations; and c) collection of data needed for effective comparative analysis.

Pre-Deployment

Both the GDPR and LED recognise the need for AI systems to be analysed before they are deployed, so that they comply with existing regulatory and human rights standards. Under Article 35 GDPR, Member States are required to carry out a 'Data Protection Impact Assessment' ('DPIA') for data processing systems that carry out 'a systematic and extensive evaluation of personal aspects relating to natural persons which is based on automated processing, including profiling and on which decision are based that produce legal effects concerning the natural person or similarly significantly affect the natural person'. The corresponding provision in the LED is Article 27, which similarly calls for DPIAs to be carried out where processing of data is likely to result in a 'high risk to the rights and freedoms of natural persons'. DPIAs under both laws have to carry out *inter alia* an assessment of the possible impact of the data processing system on the rights or individuals, and they need to mention what measures will be in place to ensure that their rights are properly protected.

DPIAs help to address a serious accountability challenge, but EU laws do not provide sufficiently helpful standards on how they should be conducted. Article 27 LED does not lay down minimum requirements for how DPIAs should be carried out. On the other hand, there are aspects of Article 35 GDPR, which if used to guide how DPIAs should be conducted for AI systems used in criminal justice, would raise concerns. The foremost challenge is the level of transparency mandated by the GDPR. DPIAs are envisaged largely as internal processes led by the data controller, who may seek the opinions of data subjects (such as members of the public or their representatives), where it is 'appropriate' to do so. The GDPR also explicitly recognises that the requirement to seek the views of data subject is 'without prejudice to the protection of commercial interests'.[113]

As outlined above, transparency is a key aspect of a fair criminal justice system, and as a general rule, all criminal justice decision-making processes need to be open to public scrutiny. There is no reason why AI systems should be exempt from this requirement, and given that administration of criminal justice is a matter of strong public interest, the public should have the right to voice their opinions and raise objections whenever AI systems impact criminal justice processes. Also, given the highly technical nature of AI systems, and their (as yet) poorly understood impact on society, impact assessments must have multi-disciplinary expert engagement.[114] In particular, DPIAs should always involve independent experts (computer scientists, in particular) who can audit, analyse, and if possible, 'explain' AI systems, so that they can help legal, policy and social science experts to determine the likely implications for the individuals' rights.

[113] Art 35(9), GDPR

[114] EFF

For public and expert consultations to be meaningful and effective, sufficient information should be made available to interested parties so that the AI system can be thoroughly understood and researched. Partnership on AI has recommended that for criminal justice risk-assessment tools, training datasets,[115] architectures and algorithms of AI systems should be made available to ensure meaningful scrutiny.[116] Commercial interests should not be regarded as a legitimate ground for limiting the disclosure of this information.

Secondly, Article 35 GDPR allows data controllers to carry out a single DPIA 'for a set of similar processing operations that present similar high risks'. There is a danger that this provision could be interpreted too broadly if Member States are given free rein to determine what two systems can be regarded as sufficiently 'similar'. There are risks in assuming that an AI system well-suited for use in a particular context or within a particular geographic area will be equally useful in another. AI systems built using data from one jurisdiction might not be able to reflect differences in, for example, law enforcement culture and patterns of behaviour, laws and policies, and socio-demographic characteristics of another jurisdiction.[117] Sometimes, these differences can be seen in the same country or even within the same region. For example, a study of 'PRECOBS' a predictive policing tool used in Baden-Wurttemberg in Germany, found significant differences in predictive utility between rural and urban areas.[118]

Finally, DPIAs seem to require data controllers to theorise the possible impact of AI systems, but there is no strict requirement for AI systems to be subject to testing or auditing before, or immediately after deployment. This overlooks the fact that flaws in AI systems, including unintentional biases, are not always easily detectable, and that they might only surface once the system is put into operation. As discussed earlier, the causes of biases in AI systems can be difficult to identify, and it is difficult to appreciate how, short of thorough testing, the true impact of AI decisions can be known.

In New York, the AI Now Institute has proposed an alternative model for impact assessments, known as 'Algorithmic Impact Assessments' ('AIAs').[119] The AIA framework sets out in detail how public authorities should conduct impact assessments of AI systems, and it can be contrasted with the provisions of the GDPR in that AIAs place much greater emphasis on the need for community engagement and consultations with external experts. This framework could serve as a useful guide for Member States seeking to establish pre-deployment procedures for approving AI systems.

AI systems should not be deployed unless they have undergone an independent public impact assessment with the involvement of appropriate experts, that is specific both to the purpose for which the AI system is deployed, and the locality where it is deployed. AI systems must be tested for impact pre-deployment, and systems should be precluded from deployment until they have undergone this testing and achieved minimum standards, such as non-discrimination.

Post-Deployment

Impact assessments of AI systems should not be regarded as 'one-off' processes. They have to be followed up with ongoing post-deployment monitoring and evaluation, so that the longer-term

[115] Ensuring that training datasets have been sufficiently anonymised, in accordance with data protection laws
[116] Partnership on AI, (n 41)
[117] John Logan Koepke and David G. Robinson, 'Danger Ahead: Risk Assessment and the Future of Bail Reform' (2018), Washington Law Review, Vol 93, Issue 4, 1725
[118] Dominik Gerstner, 'Predictive Policing in the Context of Residential Burglary: An Empirical Illustration on the Basis Pilot Project in Baden-Wurttemberg, Germany' European Journal for Security Research 3, 115 (2018)
[119] AI Now, 'Algorithmic Impact Assessments: A practical framework for public agency accountability' (2018)

impact of AI systems can be understood, and shortcomings and biases that affect the rights of individuals can be identified and fixed.

The ability of AI systems to deliver fair and just outcomes, and to meet policy objectives can be difficult to predict from the outset. Although AI systems can be validated and tested prior to deployment to check if they are likely to produce desired outcomes, their impact in the real world might be different. Furthermore, even if the likely outputs of AI systems can be predicted, it is much harder to estimate the likely impact they will have on human decision-making.[120]

Further reviews of AI systems are also necessary because criminal justice systems and the societies in which they operate change over time. A study in the United States, for example, theorises that many pre-trial risk assessment tools might be making predictions based on historic data that is no longer fit for purpose. It has been suggested that because data used to train risk assessment algorithms pre-date bail reforms in many US jurisdictions, the impact of recent measures introduced to reduce the risk of failure-to-appear, such as transportation assistance and text message alerts are not taken into consideration – potentially leading to over-incarceration.[121] Socio-demographic changes might also require AI systems to be altered so that they continue to be fit for purpose. If, for example, an area experiences high levels of net migration which results in rapid changes to policing patterns and judicial behaviour, AI systems might need to be reviewed to make sure they are not unintentionally worsening racial discrimination.

Data Collection

It is difficult to imagine how the impact of AI systems can be assessed, if there is inadequate data to support effective monitoring. The deficiency of criminal justice data across the EU has been subject to criticism. In particular, Fair Trials has found that most EU Member States do not systemically collect statistics on the duration of pre-trial detention, outcomes of criminal cases of pre-trial detainees, and the likelihood of a suspect or accused person being released by the court.[122] The data needed for effective monitoring and evaluation depends on the function of the AI system and its intended objectives, but the lack of criminal justice data more generally questions whether Member States currently have adequate legal and policy foundations for introducing AI systems responsibly into criminal justice processes. Data needed for monitoring and evaluation purposes will, of course, need to have been collected from well before the introduction of the AI system, so that a proper pre- and post- analysis comparison can be made.

Of particular concern is that in most EU Member States, race or ethnic data on criminal justice is not available, either because there is no systemised process for collecting it, or because local laws ban this practice altogether.[123] This is a serious challenge because the most predominant criticism against the use of AI systems in the United States and elsewhere is that it worsens racial and ethnic bias in criminal justice decisions. Even without official statistics, there is strong evidence in many EU Member States that certain ethnic minorities, and in particular, Roma and people of colour are unfairly overrepresented in criminal justice systems.[124] It is worrying that AI systems might worsen this discrimination, but that there will be no way of detecting this trend, because of the lack of data.

[120] See for example, Albright (n 97)

[121] Koepke and Robinson (n 117)

[122] Fair Trials (n 43)

[123] Justicia, 'Comparative Report – Ethnic, Racial Disparities in Criminal Justice' (2018), http://www.eujusticia.net/images/uploads/pdf/Justicia_Network_Disparities_in_Criminal_Justice_Comparative_Report_2018-1.pdf

[124] Ibid.

Furthermore, the absence of racial and ethnic data could also prevent pre-emptive measures to combat racial bias. It is doubtful that developers will be able to design systems free from racial bias, if they have no data against which to measure their performance.

On data collection, Fair Trials believe that EU and its Member States will need to make a strict choice. Either they should ensure that racially disaggregated criminal justice data is collected, or AI systems should be banned where they make individualised assessments for criminal justice purposes.

Effective monitoring of AI systems is not possible unless there is sufficient data that makes it possible to discern their real impact. In particular, Member States need to collect data that allow them to identify discriminatory impacts of AI systems, including discrimination on the basis of race and ethnicity.