EU Commission

Roadmap Inception impact assessment: Proposal for a legal act of the European Parliament and the Council laying down requirements for Artificial Intelligence

Fair Trials response

(10 September)

Introduction

Fair Trials is aware that the Inception Impact Assessment considers the entire and wide array of economic and societal issues in relation to AI across a spectrum of industries and social activities. However, this response is concerned specifically with AI in the context of criminal justice and will consider the legal and policy measures needed for the regulation of AI in criminal justice.

Rapid technological advancements in recent years have made machine learning and algorithmic automated decision-making systems, often referred to as artificial intelligence ('AI') a prominent aspect of our lives. There is little doubt that these systems have great capacity to increase human potential and improve the lives of many, but their increasing role in assisting important public functions has also highlighted serious risks and challenges, particularly in the context of criminal justice. If not subject to proper regulation 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.

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

Fair Trials is grateful for the opportunity to submit feedback on the EU Commission's AI Inception Impact Assessment. We submitted a response to the EU Commission's consultation on the White Paper, *'Regulating Artificial Intelligence for Use in Criminal Justice Systems in the EU'*, which provided evidence and analysis about the negative impact that 'AI' can have on criminal justice.

We are pleased that the EU Commission recognises that AI represents risks for fundamental rights, including the right to a fair trial, as well as the need for improvements to the EU's legislative framework

on AI, particularly the 'effective application and enforcement of existing EU and national legislation', as well as the 'limitations of scope of existing EU legislation'.[1]

We are also pleased to see that the objectives identified in the Roadmap include preventing or minimising significant risks for fundamental rights, as well as ensuring the effective enforcement of rules of existing EU law to protect fundamental rights and avoid illegal discrimination.[2]

However, with regards to the policy options stated in the Roadmap, in order to achieve the above stated objectives to protect and minimise risks to fundamental rights and prevent discrimination, we believe that there is a clear need for a legislative solution under Option 3. For the purposes of protecting rights in criminal justice, any legislative proposal would at the very least need to cover 'high-risk' applications (Option 3, second sub-option b), but more comprehensive protection may be offered by a legislative act which covered all AI applications (Option 3, third sub-option c).

Artificial intelligence and criminal justice: EU legislative proposals

In recent years, AI, comprising machine-learning and other analytical algorithm-based automated decision-making systems, 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 systems 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 and discriminatory policing patterns, and its use in these AI systems merely results in a reinforcement and re-entrenchment of those inequalities and discrimination in criminal justice outcomes.

[1] https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf
[2] https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=PI_COM:Ares(2020)3896535&rid=3

Given these extremely serious risks, strong regulatory frameworks are needed to govern the use of AI in criminal justice decision-making and, which in some circumstances, may restrict its use entirely. 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. There is a need for clear and enforceable standards and safeguards to protect the right to a fair trial, as well as preventing discrimination in the criminal justice system.

Issues with the existing EU legal framework

In theory, 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. The current legal framework governing the use of automated decision-making systems do not sufficiently protect individuals, such as those engaged with the criminal justice system, from the systems or the decisions they produce or influence.

Existing EU laws intend to restrict the use of automated decisions, with Article 22 of the General Data Protection Regulation ('GDPR') providing 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') has a 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*'.

As we noted in our submission to the EU Commission's consultation on the White Paper, there are three issues with 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.

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 nor the 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, but 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*'.[3] These provisions give extremely wide discretion to 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 of 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.[4] Likewise, the meaning of '*human intervention*' is open to interpretation. LED provides some guidance on who should be carrying out the human intervention,[5] but there needs to be greater clarity on what meaningful human intervention entails in different contexts.

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.[6] 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.

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 through legislation (if appropriate) to further clarify the circumstances in which Member States are allowed to deploy AI systems for criminal justice proceedings.

There is also a need for procedural safeguards that ensure 'meaningful' human oversight; no individual should be subject to an automated decision which engages their human rights without meaningful human input. There must also be greater clarity on what meaningful human intervention entails in different contexts, to prevent mere administrative (e.g 'box-ticking') human involvement in such automated decisions, without either any or sufficient engagement or consideration, from overriding protections in legislation, such as notification to an individual where they have been subject to an automated decision.

Procedural safeguards which can be put in place to ensure meaningful human input include:

---

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

[4] 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

[5] Recital 38, Law Enforcement Directive

[6] Article 29 Data Protection Working Party, 'Guidelines on Automated individual decision-making and profiling for the purposes of Regulation 2016/679' (3 October 2017)

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.**

Presumption of innocence

The right to be presumed innocent in criminal proceedings is a basic human right, and one that is expressly recognised in, and safeguarded by EU law under Directive 2016/343 (the 'Presumption of Innocence Directive').[7] 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,[8] with a variety of predictive policing tools that aim to facilitate preventative policing measures and to deter crimes before they have taken place developed and deployed across Europe.[9]

These predictive tools can be regarded as part of a broader trend in law enforcement that moves away from 'reactive' policing, and towards 'preventative' or 'proactive' policing. These tools intend to pursue legitimate objectives of preventing, or reducing harm,[10] 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.[11] 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 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 the risks of criminality of certain ethnic and racial groups.

---

[7] 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

[8] 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

[9] 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

[10] 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

[11] 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

Many AI and algorithmic systems used in criminal justice systems are statistical models, based on data. 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 criminal justice data is representative of structural biases and inequalities in the societies which the data represents, and is always comprehensively lacking in the kind of detail that is needed to make truly 'accurate' predictions or decisions. 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 and discriminatory policing patterns, and its use in these AI systems merely results in a reinforcement and re-entrenchment of those inequalities and discrimination in criminal justice outcomes.[12]

While it is clear that certain types of predictive policing infringe the presumption of innocence from a moral and ethical viewpoint, it is unclear whether these systems are considered to violate the *legal* presumption of innocence under EU law and international human rights law as currently formulated and applied. 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.

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.[13] These types of high impact, fact-sensitive decisions should never be delegated to automated processes, either wholly or partly, particularly those which ultimately operate by identifying correlations rather than causal links between an individual's characteristics and their likely behaviour.

In order to achieve this, and ensure these protections are respected by law enforcement and criminal justice agencies, strong and clear regulation is needed. There must be clear legal requirements to ensure that AI systems respect the presumption of innocence and do not operate in a way which pre-designates an individual as a criminal before trial, nor allow or assist the police to take unjustified, disproportionate measures against individuals without reasonable suspicion.

[12] 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

[13] ECHR, Article 6(2)

Discrimination

One of the most frequent and well-evidenced 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. As noted above, there are fundamental issues with the way AI systems are designed and created which can lead to bias. AI built on data embedded with such biases and used to assist, inform, or make decisions in the criminal justice system, can expand and entrench the biases represented in the data.[14] This conclusion has been echoed by the UN Rapporteur on racism, who stated in her most recent report that emerging digital technologies driven by big data and artificial intelligence "are entrenching racial inequality, discrimination and intolerance".[15]

The dangers of the failure to adequately regulate the use of AI to prevent such discrimination are clear and have been witnessed in Europe. The 'Crime Anticipation System' ('CAS'), a predictive policing software 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.[16] 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.

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.[17] It is clearly discriminatory, as well as entirely unfair and an affront to equality of arms, for people living in certain areas to be punished more severely than others in different areas for the same crimes.

These examples illustrate the need for regulations to ensure that AI systems are designed to be non-discriminatory, and to exclude categorisations and classifications that deepen and legitimise social biases and stereotypes.

---

[14] 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

[15] https://www.ohchr.org/Documents/Issues/Racism/SR/HCR44_Statement_15_July_2020.pdf

[16] 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

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

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, have clearly stated aims or objectives, and be carried out pre-deployment to reduce the likelihood of individuals being affected by discriminatory profiling and decisions. There is, therefore, a clear need for a legislative solution to mandate these requirements.

Transparency

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 is uninterpretable for most, if not all humans. Criminal procedure should enable the full disclosure of all aspects of AI systems that are necessary for suspects and accused persons to contest their findings.

In addition, AI systems' decisions, or decisions they have influenced, also 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 automated decision by an AI system, they will not have the ability to challenge that decision, or the information that the decision was based on. There must be a requirement for individuals to be notified, not just for "purely automated" decisions, but whenever there has been an automated decision-making system involved, assistive or otherwise, that has or may have impacted a criminal justice decision.

In order to ensure the necessary transparency and notification in relation to AI systems in criminal justice, regulation is needed, so that individuals are notified of AI decisions, and the decisions can be understood, scrutinised and challenged by their primary users, suspects and accused persons, as well as the general public.

Roadmap policy options

In order to regulate the current and future use of AI in criminal justice proceedings, as well as closing the gaps in existing data protection laws, the EU must, at a minimum, set the following standards:

1) to govern the design and deployment of AI systems in criminal justice systems;

2) to make sure that AI systems are used in accordance with human rights standards and prevent discrimination in criminal justice proceedings; and

3) to guide Member States in governing the deployment of AI systems and monitor their subsequent use.

Fair Trials believes that these requirements can only be achieved by **Option 3** in the Roadmap, an EU legislative instrument establishing mandatory requirements in relation to AI applications.

The **'Baseline'** of no EU policy change, the current status quo, is clearly not acceptable in this context, given the very real and increasing threats to the right to a fair trial posed by AI and algorithms in the criminal justice system.

Similarly, **Option 1** and **Option 2** are not sufficiently robust to provide the protections needed, and ultimately, there is too much at stake in the criminal justice context – and indeed others – to leave it up to self-regulation by private companies.

A soft-law approach as per Option 1, relying on industry to take the necessary action will clearly not be appropriate in the context of criminal justice, in as much as it would rely on industry to set standards or take a leading role in criminal justice systems. Nor would a 'voluntary labelling scheme' as under Option 2 be sufficient, as the protections and safeguards needed in the criminal justice system cannot be on a 'voluntary' basis.

Industry has proven in the context of criminal justice that they are not able to meet even the minimum requirements needed to protect and safeguard individuals subjected to AI systems. 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. Due to commercial or proprietary interests, these minimum requirements are often not met, with companies selling AI systems to law enforcement and criminal justice agencies citing proprietary interests when refusing to be transparent about the system or how it reaches decisions.[18] Ultimately, there is too much at stake in the criminal justice context – and indeed others – to leave it up to self-regulation by private companies.

As the Roadmap notes, soft-law may only "marginally facilitate implementation of fundamental rights legislation", whereas "binding requirements will strengthen the respect of existing fundamental rights for all AI systems covered".[19]

As a result, **Option 3**, an EU legislative instrument establishing mandatory requirements in relation to AI applications, is clearly the most appropriate option within those proposed in the Roadmap. Of the sub-options given, **sub option a** is potentially too narrow and restrictive to deal sufficiently with the

[18] 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

[19] https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=PI_COM:Ares(2020)3896535&rid=3

range of AI systems being used in the criminal justice system beyond such specific categories as biometric identification systems, though such systems do represent a significant threat to fundamental rights. The framing of **sub-option b** is too vague for us to be able to consider whether all those AI and algorithmic systems in the criminal justice system would definitively be classified as 'high-risk', although it is likely, but it is possible that a narrower definition with subjective interpretation may not sufficiently capture all those systems which could pose a threat to the right to a fair trial. A legislative act which covers all AI such as that proposed by **sub-option c** may be able to ensure proper and comprehensive protection, encompassing the range of AI and algorithmic automated decision-making systems, and provide a clear and sufficient solution to the challenges and problems posed by AI in criminal justice.