EU COMMISSION 'AI ACT' CONSULTATION

FAIR TRIALS' RESPONSE

AUGUST 2021

ABOUT FAIR TRIALS

Fair Trials is an international human rights NGO that campaigns for fair and equal criminal justice systems. Fair Trials' team of experts expose threats to justice and identify practical changes to fix them. The organisation produces original research, campaigns to change laws, support strategic litigation, reform policy and develop international standards and best practice.

Fair Trials supports movements for reform and building partnerships with lawyers, activists, academics and other NGOs. It is the only international NGO that campaigns exclusively on the right to a fair trial, providing a comparative perspective on how to tackle failings within criminal justice systems globally.

INTRODUCTION & SUMMARY

We welcome the fact that the EU is taking a much-needed legislative approach to regulate and limit the use of artificial intelligence (AI), and that it recognises that the use of AI in law enforcement and criminal justice can have serious implications for fundamental rights.

AI systems are used by European law enforcement and criminal justice authorities to predict and profile people's actions and assess risk, leading to surveillance, stop and search, questioning, arrest and detention, and sentencing and probation, as well as other non-criminal justice punishments such as the denial of welfare, housing, education or other essential services. In doing so, these systems engage and infringe fundamental rights, including the right to a fair trial, privacy, and data protection rights, as well as reproducing and reinforcing discrimination on grounds including but not limited to race, socio-economic status and nationality. As more and more countries are turning to AI in law enforcement and criminal justice, it is more crucial than ever that the EU takes this opportunity to become a leading standard-setter in this area, ensuring the protection of fundamental rights.

However, while recognising the risks to some extent, the AI Act does not go nearly far enough to prevent certain fundamentally harmful uses, particularly in relation to law enforcement and criminal justice, which will have damaging consequences across Europe for generations.

In order to do this meaningfully, the Act must prohibit AI used by law enforcement, and judicial and criminal justice authorities used to predict, profile or assess people's risk or likelihood of 'criminal' behaviour, generate reasonable suspicion, and justify law enforcement or criminal justice action, such as surveillance, stop and search, arrest, detention, pre-trial detention, sentencing and probation. No amount of safeguards, short of a full statutory prohibition, will protect against these fundamental harms effectively.

In the absence of a full prohibition, and to prevent additional harms, uphold the rule of law and safeguard justice systems, there are several bare minimum safeguards and requirements that can be adopted to lessen the fundamental rights impact of these AI systems. In particular:

a) The AI Act contains vague and non-specific 'bias' requirements, none of which will prevent discrimination and bias. AI systems used in law enforcement and criminal justice contexts must be subject to mandatory, independent bias testing, but the feasibility of such testing depends on the availability of criminal justice data that is severely lacking in the EU;
b) The AI Act must require more openness, transparency, and explainability of AI systems and their use, the decisions that are made, and significantly, must focus not just on ensuring transparency to the *users* of the systems, but also to those individuals impacted by AI or AI-assisted decisions;
c) Given that AI can have a significant impact on individuals when it is used in law enforcement and criminal justice, it is crucial that there are effective avenues for individuals to challenge not just the AI decisions, but also the system itself. However, the AI Act does not facilitate or provide clear routes for challenge or redress for individuals attempting to contest or challenge AI systems, or their decisions; and
d) The AI Act includes several exemptions for uses of AI from safeguards. This means that there is a lack of protection against a technology that can engage and infringe fundamental rights, including the right to a fair trial, privacy and data protection rights, as well as result in discrimination based on race, socio-economic status or class and nationality.

1. PROHIBITED ARTIFICIAL INTELLIGENCE PRACTICES MUST INCLUDE LAW ENFORCEMENT AND CRIMINAL JUSTICE USE OF AI

The AI Act must prohibit the use of AI and automated decision-making systems in law enforcement and criminal justice which predict, profile and assess people's 'risk' of criminality or certain behaviour classed as criminal, generate reasonable suspicion and justify law enforcement action, arrest, pre-trial detention, sentencing and probation.

Articles 6 and 7 of the AI Act classify AI systems for use in law enforcement and the administration of justice as 'high-risk' applications that are explicitly permissible, but subject to various requirements under Chapter 2. As set out in Annex III, permissible applications include predictive and profiling tools used by law enforcement and criminal justice authorities, similar to those already deployed in the EU, which have been proven to produce discriminatory outcomes.

Fair Trials welcomes the fact that the Commission recognises the risks associated with the use of AI in law enforcement and the administration. However, there are serious fundamental issues with the use of AI in law enforcement and criminal justice which cannot be prevented against by safeguards, other than the full prohibition of their use under the AI Act.

These include:

(i) the hard-wiring of discrimination via the use of criminal justice and non-criminal justice data, with predictions, profiles and risk-assessments based on correlations between an individual and the group they are assessed to most closely resemble;

(ii) the same kind of profiling completely undermining the presumption of innocence, through pre-emptive decision-making based on predictions, profiles and risk-assessment likely to infringe the right to a fair trial; and

(iii) technological barriers to transparency with certain types of machine-learning systems that prevent people from understanding, and where necessary, challenging decisions that affect them.

Article 5 of the AI Act recognises that the negative impact of certain uses of AI, such as real-time biometric identification, is potentially so great that the only appropriate regulatory response is a statutory prohibition. Fair Trials believes that given the significant risks to the right to a fair trial, the considerable societal implications of worsening discrimination in the criminal justice system, and the serious human impact of arrest, deprivation of liberty, and criminal convictions, the use of AI in law enforcement and criminal justice to predict, profile and assess people's 'risk' of criminality or certain behaviour classed as criminal, or profile geographic or physical areas, should similarly be prohibited in the Act.

In terms of specifying which uses should be prohibited, those 'Law enforcement' and 'Administration of justice and democratic practices' uses recognised as high-risk within the 'high-risk' framework in Annex III provide a starting point, specifically 7(a), (d), (e), (f), and (g), and 8(a), and should be included in Article 5. The prohibition in Article 5 should also include AI and automated decision-making systems which predict and profile levels of crime or certain criminal actions in geographic or physical locations, based on historic data, for the purpose of making law enforcement decisions, as well as the use of such systems to predict, profile or assess risk in the context of pre-trial detention, sentencing or probation by either law enforcement, judicial or other criminal justice authorities.

Hard-wiring of discrimination in AI systems

The way in which predictive and risk-assessment AI systems are designed, created and operated means that they are predisposed to produce biased and discriminatory outcomes, further entrenching existing biases that lead to the over-representation of certain racial, ethnic, and socio-economic groups in the criminal justice system.

AI systems are created, trained and operated using data, upon which the system analyses and produces outcomes. Where an AI system is based on biased data or is operated using biased data, bias

can result from over-representations in the data that is used to train it, or the data upon which the system carries out analysis. Some automated systems known as machine-learning algorithms 'learn' how to make assessments or decisions based on analysis of data. However, the data used to train the machine-learning system might be incomplete, inaccurate, or biased, and this could lead to the system producing inaccurate or biased outcomes.

The type of AI designed or created for use in the criminal justice system will almost inevitably use data which is heavily reliant on or entirely made up of law enforcement data, crime records or other criminal justice authorities' data. These data and records do not represent an accurate record of criminality but merely a record of law enforcement or prosecutorial or judicial decisions – the crimes, locations and groups that are policed, prosecuted and criminalised within that society, rather than the actual occurrence of crime. The data may not be categorised or deliberately manipulated to yield discriminatory results, but it will reflect the structural biases and inequalities in the society which the data represents.

For example, law enforcement actions or judicial decisions resulting from or influenced by racial or ethnic profiling, or the targeting of less economically advantaged people, can result in biased data concerning certain groups in society,[1] and the systematic under-reporting and systematic over-reporting of certain types of crime and in certain locations will also be represented in crime records and data.[2] Similarly, prosecutorial or judicial decisions will often disproportionately affect those sae groups, who are drawn into criminal justice systems as a result of biased policing, and the same structural racism which influences law enforcement decisions will influence prosecutorial and judicial decisions.

Although AI and automated decision-making systems are intended to profile an individual, predict an or assess the 'risk' of an individual's likelihood of certain (criminal) behaviour in future, or the likelihood of criminal activity occurring in a specific area, these tools are, in reality, incapable of making individualised predictions. These systems use and analyse data from many individuals or areas, and then forecast *aggregate group* or area risk.[3] A person-oriented prediction or risk assessment merely indicates that a person shares traits with a group who did or did not carry out certain (criminal) behaviour at a certain rate, but the prediction or assessment cannot and will not provide individualised information about how that individual will behave or act.[4] As law enforcement records are most accurately a record of law enforcement action, predictive or risk-modelling AI tools will most accurately provide predictions of future law enforcement action, such as the targeting of groups or

---

[1] Oosterloo, Serena and van Schie, Gerwin 'The Politics and Biases of the "Crime Anticipation Systems" of the Dutch Police' (2018) Proceedings of the International Workshop on Bias in Information, Algorithms, and Systems co-located with 13th International Conference on Transforming Digital Worlds (iConference 2018), http://ceur-ws.org/Vol-2103/paper_6.pdf.

2 Lum, Kristian and William, Isaac 'To Predict and Serve?' (2016) Significance 13 (5): 14–19 https://rss.onlinelibrary.wiley.com/doi/full/10.1111/j.1740-9713.2016.00960.x; Bennett Moses, Lyria and Chan, Janet, 'Algorithmic prediction in policing: assumptions, evaluation, and accountability' (2016) Policing and Society 28:7, https://www.tandfonline.com/doi/10.1080/10439463.2016.1253695; Barocas, Solon and Selbst, AndrewD., 'Big Data's disparate impact' (2016) *California Law Review*, 104, 671, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2477899.

[3] Buskey, Brandon and Woods, Andrea 'Making Sense of Pretrial Risk Assessments' (2018) The Champion Issue June 2018, https://www.nacdl.org/Article/June2018-MakingSenseofPretrialRiskAsses.

[4] ibid.

areas that have historically been targeted by law enforcement.[5] The same is often true for prosecutorial or judicial records.

It is well documented that there is significant discrimination in criminal justice systems in the EU.[6] The EU Fundamental Rights Agency (FRA) has found clear disparities in stop and search practices against people from ethnic and national minority groups in several Member States in studies conducted over 5 years in both 2010 and 2018.[7] Across Europe, there are also clear and widespread racial and ethnic disparities in pre-trial detention. In Council of Europe states, 40% of all 'foreign nationals' in prison were being held in pre-trial detention, compared to 25% of all prisoners. People from certain ethnic or racial groups face worse outcomes in their cases: they face longer prison sentences and are not granted non-custodial sanctions such as fines,[20] and data from across Europe shows that they are also disproportionately over-represented in prison, relative to the percentage of the population they represent.[21]

When law enforcement and criminal justice data representing such disparities is used in predictive, profiling and risk-assessment AI systems in criminal justice, it will inevitably flag up people, whose profiles fit those who are over-represented in that data, as being higher risk. AI built and operated on data embedded with such biases, which then assists or informs, law enforcement or criminal justice decisions, can also result in a re-enforcement and re-entrenchment of those biases.[26] This leads to a self-fulfilling prophecy whereby the predictions become true because police target those groups following the prediction or risk-assessment, which leads the system to strengthen its correlation between those groups and criminal justice outcomes, leading to self-perpetuating 'feedback loops' which reinforce patterns of inequality.[27]

These AI systems are increasingly used in European and EU Member States' law enforcement and criminal justice systems. For example, in 2015 the Amsterdam Municipality started the 'Top 400', a

---

[5] ibid.

[6] Justicia European Rights Network 'Disparities in Criminal Justice Systems for Individuals of Different Ethnic, Racial, and National Background in the European Union' (November 2018); and Namoradze, Zaza and Pacho, Irmina 'When It Comes to Race, European Justice Is Not Blind' (2018) Open Society Justice Initiative, https://www.justiceinitiative.org/voices/when-it-comes-race-european-justice-not-blind.

[7] Fundamental Rights Agency, 'EU-MIDIS II Second European Union Minorities and Discrimination Survey' (2018); and European Union Agency for Fundamental Rights 'Data in Focus Report – Police Stops and Minorities' (2010), https://fra.europa.eu/sites/default/files/fra_uploads/1132-EU-MIDIS-police.pdf.

[20] Fair Trials, Disparities and Discrimination in the European Union's Criminal Legal Systems, January 2021. https://www.fairtrials.org/sites/default/files/publication_pdf/Disparities-and-Discrimination-in-the-European-Unions-Criminal-Legal-Systems.pdf

[21] Hilde Wermink, Sigrid van Wingerden, Johan van Wilsem & Paul Nieuwbeerta, Studying Ethnic Disparities in Sentencing: The Importance of Refining Ethnic Minority Measures, in Handbook on Punishment Decisions 239-264 (L Ulmer & M Bradley 1ed 2019); Samantha Bielen, Peter Grajzl and Wim Marneffe, Blame Based on One's Name? Extralegal Disparities in Criminal Conviction and Sentencing, SSRN Electronic Journal, (2008); and Virginie Gautron & Jean-Noël Retière, La décision judiciaire : jugements pénaux ou jugements sociaux ?, 88 Mouvements 11-18 (2016).

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

27 ibid; Ensign, Danielle et al. 'Runaway Feedback Loops in Predictive Policing' (2017) Cornell University Library, 29 June 2019, https://arxiv.org/abs/1706.09847; Bennett Moses, Lyria and Chan, Janet, 'Algorithmic prediction in policing: assumptions, evaluation, and accountability' (2016) Policing and Society 28:7, https://www.tandfonline.com/doi/10.1080/10439463.2016.1253695.

risk-modelling profiling system, in partnership with police and social services.[28] It attempts to profile the 'top 400' young people, under the age of 16, who it considers are most at risk of committing crime in future. The criteria used by Top400 to assess the risk of someone committing crime includes whether someone has been arrested – but not charged or convicted – as a suspect for certain crimes, whether they are associated with 'gang' activity, as well as whether they have been absent from school or changed schools regularly, or have previously been subjected to police surveillance. None of these indicators refer to objective evidence of criminality as tested by a criminal court,[29] and include the use of arrest and 'suspicion' data and reports, as well as police hearsay and opinion 'intelligence', without any objective evidence of involvement in crime, as well as a child's school history. There is a clear discriminatory outcome of this system, as 1/3 of the people on the Top400 are young Dutch Moroccan men.

Even seemingly legitimate data can act as a 'proxy' for other factors, for example race or ethnicity, such as the use of school data as above, or, as is a common example, home addresses or area codes.[30] AI systems will seek out correlations between area codes and the risk of reoffending – in other words, to identify which area codes have 'higher-risk' residents than others.[31] As there is very pronounced ethnic residential segregation in many European countries,[32] it is highly probable that in practice AI systems will inadvertently establish a link between ethnic origin and risk. Many other forms of data can also act as proxy for race or ethnicity, such as financial information on income, data on access to welfare or benefits or other public services. The use of financial data can also lead to targeting people based on their socio-economic background which is also problematic and discriminatory.

AI systems and the presumption of innocence

Predictive and risk assessment AI and automated decision-making systems target individuals and profiles them as criminals, resulting in arrest, pre-trial detention as a 'preventative' measure, or a longer sentence, before they have carried out the criminal action for which they are being profiled. In essence, the very purpose of these systems is to undermine the fundamental right to be presumed 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'),[42] the Charter of Fundamental Rights,[43] as well as the European Convention on

[28] Amsterdam Municipality 'Tweede halfjaarmonitor Top400' (2016) https://docplayer.nl/113213400-Tweede-halfjaarmonitor-top400.html.
[29] Gemeente Amsterdam 'Top 400', https://www.amsterdam.nl/wonen-leefomgeving/veiligheid/top400/.
[30] Zuiderveen Borgesius, Fredreik 'Discrimination, artificial intelligence, and algorithmic decision-making' (2018) Directorate General of Democracy, Council of Europe.
[31] Oswald, Marion et al. 'Algorithmic risk assessment policing models: lessons from the Durham HART model and 'Experimental' proportionality' (2018) Information & Communications Technology Law. 27:2, 223-250.
[32] E.g. Sweden. See Malberg, Bo 'Residential Segregation of European and Non-European Migrants in Sweden' (2018) Eur J Popul 34(2): 169-193 https://link.springer.com/article/10.1007/s10680-018-9478-0; FRA, 'Summary Report – The State of Roma and Traveller Housing in the European Union – Steps towards Equality' (2010).
[42] 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.
[43] ibid., Article 48.

Human Rights.[44] People should not be judged on the basis of their past actions by predictive and risk assessment tools. This destroys the very concept of the presumption of innocence. Further, people should also not be pre-judged on the actions of other people who share similar characteristics or backgrounds, such as race or ethnicity, socio-economic status, residential address or other factors. However, this is exactly what these systems are trained to do – to find correlations between different characteristics and backgrounds, and predict, profile and assess 'risk' on the basis of these characteristics and backgrounds.

Although the use of these tools in law enforcement and criminal justice do not directly 'convict' people, they allow police to treat legally innocent individuals as pseudo-criminals, resulting in surveillance and monitoring, questioning and harassment, and even arrest, as well as further non-criminal justice action, such as denial of access to public services, or reporting to social services – effectively 'punishing' them on account of their profiles. When used to decide on bail or pre-trial detention, they can influence serious decisions on the deprivation of liberty, potentially resulting in someone being detained awaiting trial as a result.

The use of AI and automated decision-making systems to make such pre-trial predictions, profiles and risk-assessments damages the fundamental human rights principle that the a matter of guilt or innocence can only be determined by means of a fair and lawful criminal justice process.[47] AI systems cannot respect the presumption of innocence, if they are used to pre-designate an individual as a criminal before trial, or they are used to assist law enforcement or other criminal justice authorities to take unjustified or disproportionate measures against individuals without reasonable suspicion.

Technological barriers to transparency, explainability and accountability

AI systems that are used to assist law enforcement and criminal justice decisions through individualised predictions and risk-assessments tend to have technological barriers that prevent effective and meaningful scrutiny. Certain types of AI systems, including machine-learning systems which use neural networks, are built in ways that makes it impossible to decipher how exactly decisions and outcomes are made. This not only makes it extremely difficult to regulate these systems, but it also prevents people from understanding how decisions that affect them are made.

Transparency and accountability are principles that should be central to all criminal justice systems. It is only where policing powers and criminal justice decisions are open to effective scrutiny that there is proper oversight of the criminal justice system, and fundamental fair trial rights can be exercised in practice.

Fair Trials would emphasise that AI and automated decision-making systems must not only be 'transparent', but also explainable and intelligible.[48] The 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.[49] In principle, this is an essential and very useful requirement, but it is also

[44] Article 6(2), ECHR.
[47] ECHR, Article 6(2).
[48] Article 32, Toronto Declaration.
[49] Recital 71, GDPR.

one that seems difficult to implement in practice, given that both 'explainability' and intelligibility are highly subjective concepts.

It is crucial that people affected by AI or AI-assisted decisions in the policing or criminal justice context are able to challenge the accuracy and lawfulness of those decisions. The effective 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.

However, machine-learning algorithms using neural networks, often described as 'black boxes' for the inability to see, analyse or understand what is going on inside the network, cannot meet these essential requirements. Some machine-learning algorithms are simply too complex to be understood to a reasonable degree of precision,[50] and this is especially the case where AI systems incorporate 'neural networks'. 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.[51] 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 human capabilities.[52]

2. SAFEGUARDS IN THE AI ACT PROVIDE INADEQUATE PROTECTIONS AGAINST DISCRIMINATION AND BIAS

The AI Act contains vague and non-specific 'bias' requirements, none of which will prevent discrimination and bias. AI systems used in law enforcement and criminal justice contexts must be subject to mandatory, independent bias testing, but the feasibility of such testing depends on the availability of criminal justice data that is severely lacking in the EU.

As considered above, the law enforcement and criminal justice data used to create, train and operate AI and automated decision-making systems is reflective of systemic, institutional and societal biases which result in Black people, Roma, and other minoritised ethnic people being overpoliced, and disproportionately detained and imprisoned across the EU. Fair Trials believes that these challenges are so fundamental and ingrained, that it is questionable whether any such system would not produce such outcomes.

The safest and surest way to prevent discrimination is through prohibition, but in the event that a full prohibition is not forthcoming, a rigorous testing regime is the bare minimum required to lessen the risk of discrimination and ensure equality before the law and non-discrimination under the Charter of Fundamental Rights.[55]

Current safeguards in the AI Act

[50] 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)
[51] Ibid.
[52] Ibid.
[55] ibid., Article 48.

The AI Act includes various safeguards that attempt to address the risk of bias and discrimination. These include:

a) A requirement to have a 'risk management system' to monitor a high-risk AI system during its entire lifecycle (Article 9);
b) A requirement to ensure that data used for the training, validation, and testing of AI systems are subject to 'appropriate data governance and management practices (Article 10);
c) A requirement to ensure that high-risk AI systems achieve an appropriate level of accuracy (Article 15).

These safeguards are clearly well-intended, but none are capable of addressing the risk of bias and discrimination effectively.

Article 9: 'Risk management system'

Fair Trials agrees that high-risk AI systems should be subject to ongoing monitoring. However, we have serious doubts that 'risk management systems' will be able to sufficiently address the risk of bias and discrimination in high-risk AI systems.

Article 9(4) sets out that risks should be eliminated "*as far as possible*" through "*design and development*" but considers that some risks "*cannot be eliminated*", and that some risk is deemed acceptable "*provided that the system is used in accordance with its intended purpose or under conditions of reasonably foreseeable misuse*". Discriminatory outcomes from AI systems are not just a risk. They are also often a fundamental element of AI systems used in law enforcement and criminal justice, and often something which occurs *during* the design and development phase.

Article 10: Data and data governance proposals

The high-risk AI system data governance and management proposals in Article 10 will not prevent data with over-representations and disproportionate representations being used, leading to discriminatory outcomes. It is positive that Article 10(5) of the Act recognises the need that special category data shall be required to be processed in order to make assessments on bias. However, without a meaningful framework, to ensure for rigorous monitoring and testing, this ability carries little weight.

Article 10(2)(f) mentions the examination of training, validation and testing data for bias, but does not set out in any detailed or meaningful way how that should occur, nor does it make it a mandatory requirement, as opposed to a "*practice*" that may or may not occur. Moreover, there is no mention of any examination of the actual operation which AI systems would be analysing when deployed.

A training, validation or testing dataset can be "*relevant, representative, free of errors and complete*" as required by Article 10(3), yet still discriminatory. It is incomprehensible what *"representative*" might mean for a criminal justice dataset, as structural racism means that where criminal justice data is available, such as in the United Kingdom, Black people and other minoritized ethnic people are disproportionately over-represented.[57] The same applies to the requirement in Article 10(4) requirement that training, validation and testing data sets "*shall take into account… the characteristics*

[57] The Lammy Review: an independent review into the treatment of, and outcomes for, Black, Asian and Minority Ethnic individuals in the Criminal Justice System, 2017. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/643001/lammy-review-final-report.pdf

and elements that are particular to the geographical, behavioural or functional settings within which the high-risk AI system is intended to be used". The dataset used may well take into account all the characteristics and elements and yet still be deeply discriminatory.

Article 15: Accuracy, robustness, and cybersecurity

Article 15 of the Act also states that high-risk AI systems are required to be designed and developed to ensure an "*appropriate level of accuracy*" which is to be established "*in the light of their intended purpose*". There is no clarity on what that "*appropriate level*" of accuracy is. It would not be appropriate for AI systems used in law enforcement and criminal justice, and the resulting consequences of such decisions, to be any less than certain, not least to ensure non-discrimination and that fundamental rights are protected

The need for mandatory independent bias testing

The only effective way in which AI systems may be regarded as non-discriminatory is if they are subject to rigorous mandatory independent testing for biases, both in the design and pre-deployment phase, and continuously post-deployment. These types of tests need to be carried out throughout the life-cycle of an AI system, as is set out in Article 9, analysing the potential impact of the AI system pre-deployment, as well as continuing 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, such tests are just not feasible in many Member States, where there is often a lack of criminal justice data, and where local laws prohibit the collection of racially-disaggregated data in any circumstances.

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 among Member States more generally questions whether they currently have adequate legal and policy foundations for introducing AI systems responsibly into criminal justice processes. 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.[58]

A further challenge to assessing and analysing discriminatory police practices in Europe is that in most EU Member States the official collection of criminal justice data disaggregated by ethnicity, race and/or nationality is not available because it is either forbidden by law, or not standard practice, and there are not consistent practices for collecting and analysing such data across the EU.[59]

Without the relevant data to analyse AI and automated decision-making system outcomes, there can be no way of detecting whether there is bias or discrimination. 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. Data needed for monitoring and evaluation purposes will, of course, need

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

[59] 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

to have been collected starting from well before the introduction of the AI system, so that a proper pre- and post- analysis comparison can be made.

3. THE AI ACT MUST ENSURE FAR GREATER TRANSPARENCY AND EXPLAINABILITY

The AI Act must require more openness, transparency, and explainability of AI systems and their use, the decisions that are made, and significantly, must focus not just on ensuring transparency to the *users* of the systems, but also to those individuals impacted by AI or AI-assisted decisions.

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,[60] and to be given the opportunity to comment on the evidence used against them.[61] 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.[62]

However, the AI Act falls far short of ensuring sufficient transparency that would ensure that the right to a fair trial and the right to liberty can be exercised effectively. In particular, it is disappointing that while the AI Act recognises the importance of transparency, it prioritises the interests of the users of AI systems, and commercial interests, over the fundamental rights of those affected by these systems.

AI systems must be transparent for both users and affected individuals

Article 13 provides for the transparency and provision of information to users in relation to high-risk AI systems. This includes that they must be designed and developed in such a way to ensure that *"their operation is sufficiently transparent"* and to *"enable users to interpret the system's output and use it appropriately"*. Information must also be provided on its intended purpose, the "*level of accuracy robustness*" and "*accuracy*", potential fundamental rights issues and human oversight measures, and details about input data and training data used.

This information provision is a positive measure, but this attempted transparency safeguard fails at a key juncture – ensuring meaningful transparency and explainability to those *impacted* by the system, not just the user, so that individuals subjected to AI systems' decisions are given information about the decision, are able to understand that information, and are able to use that in any contestation or challenge to that decision.

By contrast, the requirements set out in Article 60 and Annex VIII for what information is to be provided on the publicly-accessible EU database for stand-alone high-risk AI systems, do not replicate the same level of transparency as that which is intended to be provided to AI system users. The database does not require the same level of information to be publicly available, such as information on the level of accuracy, performance and intended subjects, potential risks, fundamental rights issues, human oversight measures and the training and testing data sets used.

---

[60] ECtHR, *Beraru v Romania* App. No. 40107/04 (Judgment of 18 March 2014)
[61] ECtHR, *Kuopila v Finaland*, App. No. 27752/95 (Judgment of 27 April 2000)
[62] Access to Information Directive, Article 7(1); ECtHR, *Wloch v Poland*, App. No. 27785/95 (Judgment of 19 October 2000)

AI systems must be explainable to both users and affected individuals

Article 13(3) that states that AI systems must *"enable users to interpret the system's output and use it appropriately"*. This is a vague requirement that is widely open to interpretation, but Fair Trials welcomes the importance given to ensure explainability of AI systems.

However, this requirement gives no thought to the individual impacted by the system decision, or whether they might be able to interpret the system's output appropriately. The right to a fair trial and liberty can only be exercised effectively in practice, where suspects and defendants have the facilities and capabilities to challenge decisions regarding them. The inability to understand how a decision has been made completely undermines defence rights. Further, the AI Act gives no consideration of how people with little or no digital literacy, let alone the level of technical expertise often required to analyse AI systems and their decisions, would be able to interpret the output of an AI system and prepare a case or defence in relation to it.

The threshold for whether an AI system's output can be interpreted and used appropriately should not be the user of the system, but the person or people who are impacted by the system. It is their fundamental rights and freedoms which may be impacted, particularly in relation to law enforcement and criminal justice uses of AI, and in order to protect those fundamental rights.

There should be transparency requirements for human overseers of AI systems

Article 14 provides that high-risk AI systems must have human oversight while the system is in use, and be "*effectively overseen*". The human overseer must "*fully understand the capabilities and limitations*" of the AI system, "*remain aware*" of automation bias, be able to interpret the systems output, and importantly, be able to ignore, disregard or "*reverse*" the systems output. These are positive requirements.

However, the Act must go further in what it requires from decision-makers. Decision-makers should be required to show how and in what way decisions were influenced, by a broader range of factors other than the AI system, through fully reasoned, case-specific, written decisions. 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.

Trade secrets should never be a barrier to transparency

The "*confidentiality*" requirements in Article 70(2) seriously restricts meaningful transparency of AI systems and their use, with its emphasis on protecting "trade secrets" and "confidential business information". We note that Article 70(3) suggests that intellectual property does not affect obligations under '*criminal law of Member States*', but it needs to be clarified whether or not the effect of Article 70 is to protect 'trade secrets' from being disclosed to criminal suspects and defendants, even where that information is relevant for challenging decisions made regarding them.

Proprietary software and commercial concerns should never be a barrier to the openness and transparency necessary in criminal justice proceedings. It is unclear how technical documentation might be made accessible for use in criminal justice proceedings, as the provision specifically states

that confidential information on law enforcement uses of high-risk AI shall not be disclosed, and only staff of the market surveillance authority holding the appropriate level of security clearance are allowed to access technical documentation which must remain on the premises of the market surveillance authority.

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.[64]

Individuals affected by AI or AI-assisted decisions should be notified

There is a requirement under Article 52 to ensure that AI systems which interact with people are designed and developed in such a way that individuals are informed that they are interacting with an AI system. However, the requirement does not apply to AI systems "*authorised by law to detect, prevent, investigate and prosecute criminal offences*".[65]

We appreciate that the primary purpose of this exemption might be to enable law enforcement authorities to use tools whose effectiveness depends on discreet deployment. However, the exemption is far too broad, and it could potentially cover almost any type of AI system used to assist law enforcement authorities and criminal justice decision-making. Defendants and/or their representatives have the right to know where there has been an AI or 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.

4. THE AI ACT MUST MAKE PROVIDE CLEAR AVENUES FOR REDRESS FOR INDIVIDUALS IMPACTED BY AI DECISIONS

Given that AI can have a significant impact on individuals when it is used in law enforcement and criminal justice, it is crucial that there are effective avenues for individuals to challenge not just the AI decisions, but also the system itself. However, the AI Act does not facilitate or provide clear routes for challenge or redress for individuals attempting to contest or challenge AI systems, or their decisions.

Another serious omission from the Act is a clear provision for individuals to contest or challenge an AI decision. It is essential that individuals subjected to AI systems' decisions are able to contest and challenge them. There does not appear to be any such provision in the Act at all. Instead, the Act only allows for certain authorities to provide oversight and accountability of the use of AI systems. Article 63 appears to put all enforcement power for requirements in the Act on high-risk AI systems in the hands of the "*market surveillance authority*", which for law enforcement uses are required to be a data protection supervisory authority or national competent authority.[78] The Act makes sets out a route for "providers" of AI systems to report "incidents" or "malfunctioning" to the market surveillance authority under Article 62, but no comparable route for individuals subjected to those incidents or malfunctioning. The power to take "*corrective action*" under Article 22, where an AI

[64] Access to Information Directive, Article 7(1),(2), and (4)
[65] Article 52(1),
[78] Article 63(5)

system is "*not in conformity*" with the Act, also lies with the provider, signifying a clear conflict of interest, and there is no apparent route for an individual impacted to raise relevant concerns. The only apparent route to challenge an AI system appears to be via the certification process, set out in Articles 44 and 45, whereby "*parties having a legitimate interest*" can appeal against the decision of a notified body to award a certificate of conformity with the obligations under the Act.

5. WIDE-RANGING EXEMPTIONS IN THE ACT SERIOUSLY UNDERMINE SAFEGUARDS IN THE ACT

The AI Act includes several exemptions for uses of AI from even these safeguards. This means that there is a lack of protection against a technology that can engage and infringe fundamental rights, including the right to a fair trial, privacy and data protection rights, as well as result in discrimination based on race, socio-economic status or class and nationality.

For example, Article 52, which, as considered above, exempts AI systems which are used to detect, prevent, investigate, and prosecute criminal offences – effectively law enforcement and criminal justice purposes – from the requirement to ensure someone is informed they are dealing with an AI system. Article 47 provides for derogation from the Article 43 "*conformity assessment procedure*", an albeit weak attempt to harmonise commercial standards on AI, allowing high-risk AI systems which are used for "*public security*" or "*the protection of life*" to be placed on the market without the need to follow the Article 43 requirements.

Article 54 specifically permits the bending of data protection rules so that data lawfully collected for one purpose can be used for another – training and testing "*innovative AI systems*" for law enforcement and criminal justice purposes, and for *"public security*". While this does require that any processing of personal data done under this requirement does not lead to "*measures or decisions affecting the data subjects*", it is uncertain how effectively this will be enforced. The concept of a 'trial' or 'pilot' of AI and automated decision-making systems has become meaningless in recent years as some European countries have begun to trial and test these systems in real-life circumstances, such as trials of live facial recognition and predictive bail systems in the UK, which resulted in stops, searches, and arrests, on citizens of (at the time) a Member State country, during 'trials' between 2016 and 2019.[79]

Europol and other international law enforcement bodies

There is a specific exemption in the AI Act in relation to Europol and other "*international organisations*" working on law enforcement and judicial cooperation, which means that these organisations and any AI systems they use will not be subject to what little safeguards there are contained in it.

Article 2(4) of the AI Act which sets out its 'scope', states:

> *"This Regulation shall not apply to public authorities in a third country nor to international organisations falling within the scope of this Regulation pursuant to paragraph 1, where those*

[79] https://www.wired.co.uk/article/london-met-police-facial-recognition

authorities or organisations use AI systems in the framework of international agreements for law enforcement and judicial cooperation with the Union or with one or more Member States"

Recital 11 of the pre-amble also specifically mentions Europol in this context:

> *"... **this Regulation should not apply to** public authorities of a third country and **international organisations when acting in the framework of international agreements concluded at national or European level for law enforcement and judicial cooperation** with the Union or with its Member States. Such agreements have been concluded bilaterally between Member States and third countries or between the European Union, **Europol** and other EU agencies and third countries and international organisations*." (emphasis added)

Europol already collects and holds a vast amount of extremely sensitive personal data in its databases and information systems, such as the Europol Information System (EIS), Europol Analysis System (EAS) which hosts and uses a number of analytical tools including Serious and Organised Crime Threat Assessment (SOCTA), Internet Organised Crime Threat Assessment (IOCTA), EU Terrorism Situation and Trend Report (TE-SAT) and the Scanning, Analysis and Notification (SCAN) team.[80]

Europol also has numerous AI research and innovation projects underway. It recently established an innovation laboratory, which will monitor emerging technologies and their usefulness for law enforcement and participate in projects aiming to develop new ways of using those technologies for the police. This includes AI, machine learning, big data and augmented reality.[81]

The general exemption in Article 2, alongside the above exemptions, intends to give Europol *carte blanche* to use potentially harmful AI systems, as well as developing new ones free from scrutiny, oversight and apparently even adherence to the law, completely undercutting any attempts to safeguard against discrimination or protect the right to a fair trial, especially in the context of cross-border law enforcement cooperation. These wholesale exemptions within the proposed AI Act for Europol – and other agencies – to use AI, mean that there are minimal or no safeguards to protect fundamental rights and protect against discrimination.

[80] https://www.europol.europa.eu/activities-services/services-support/strategic-analysis

[81] https://cordis.europa.eu/project/id/767542/de; https://www.statewatch.org/news/2020/november/eu-police-seeking-new-technologies-as-europol-s-innovation-lab-takes-shape/