



HARVARD Kennedy School

**MOSSAVAR-RAHMANI CENTER**  
for Business and Government

# **Future for AI, Machine Learning and Assistive Technology in the Justice System: A Principled and Practical Approach**

**Rt Hon Sir Robert Buckland KBE KC**  
**Harvard Kennedy School**

May 2025

## **M-RCBG Associate Working Paper Series | No. 257**

The views expressed in the M-RCBG Associate Working Paper Series are those of the author(s) and do not necessarily reflect those of the Mossavar-Rahmani Center for Business & Government or of Harvard University. The papers in this series have not undergone formal review and approval; they are presented to elicit feedback and to encourage debate on important public policy challenges. This paper is copyrighted by the author(s). It cannot be reproduced or reused without permission. Pursuant to M-RCBG's Open Access Policy, this paper is available to the public at [hks.harvard.edu/centers/mrcbg](https://hks.harvard.edu/centers/mrcbg) free of charge. Papers may be downloaded for personal use only.

# **THE FUTURE FOR AI, MACHINE LEARNING AND ASSISTIVE TECHNOLOGY IN THE JUSTICE SYSTEM: A PRINCIPLED AND PRACTICAL APPROACH**

**BY RT HON SIR ROBERT BUCKLAND KBE KC, SENIOR FELLOW, MOSSAVAR-  
RAHMANI CENTER FOR BUSINESS AND GOVERNMENT, HARVARD  
KENNEDY SCHOOL**

## **INTRODUCTION**

In my first paper, I aimed to set the scene by focusing, not on AI technology itself, but upon judges and judgement; in other words, what is the ethical and practical context that judges find themselves in, including the challenges of bias and deepfakery.<sup>1</sup> In the year or so since its publication, the AI world has moved on at even greater pace, leading me to the conclusion that, due to the extrinsic changes being made to our way of life by AI, the justice system has neither the luxury of time nor of choice when it comes to dealing with AI.

The aim of this paper is to lay the groundwork for the widespread deployment of AI in justice systems with the aim of improving access to justice and shortening the time it takes to resolve cases. I look to my own jurisdiction, England and Wales as a place where I advocate some new approaches and governance structures, based upon some general principles laid down in this paper.

The paper will, first, identify the current challenges facing the judicial system and the potential of AI to solve them, second, lay down a set of principles to guide the use of AI to resolves cases, and third, address the potential unintended consequences of widespread AI use and measures that can be taken to alleviate them.

If I needed reminding of the point that AI technology is all around us already and is part of our system of policing and justice, I received one just last month as I walked, with justified trepidation, towards the Principality Stadium at Cardiff to see my beloved Wales rugby team get thrashed by England. As I did, I passed signs telling me that the South Wales Police were using facial recognition technology in the City Centre that day as part of their crime prevention and investigation measures. The use of this technology has been controversial to

---

<sup>1</sup> Robert Buckland, "AI, JUDGES AND JUDGEMENT: SETTING THE SCENE", M-RCBG Working Paper No. 220, <https://www.hks.harvard.edu/centers/mrcbg/publications/awp/awp220>.

say the least. It was South Wales Police whose initial deployment of facial recognition landed them up in the England and Wales Court of Appeal, where it was held that its use was unlawful, on the basis that its potential impact on minority communities had not been evaluated prior to its use, and that this was a breach of the Police authority's statutory duty under the Equality Act.<sup>2</sup> In other words, the question of bias had not been properly addressed.

Since then, with the necessary work and assessments having been done, this technology is reappearing. The London Borough of Croydon will soon be installing permanent facial recognition infrastructure, in order, it says, to help detect crime. The arguments about bias go on. There is no specific modern legislative structure underpinning the use of this technology, and it may well be that this is an indicator of how things will develop elsewhere.

In the ensuing months, I like many others have struggled to keep up with the pace of change. I have benefitted from conversations with experts in AI like Professor Richard Susskind, who, rightly, take the view that the rise of AI isn't just another "bolt on" to traditional practices, but a complete transformation of our lives and of our relationships with existing institutions and systems. It is in that spirit of transformation, then, that I approach my second and final paper.

A growing and real question for us now is: will people still want to use conventional court litigation systems if they can access private dispute resolution processes that are cheap and fast? Does increasing familiarity with AI mean that more and more people will readily consent to automated decision-making in justice? I think the answer is a resounding yes, but that the consequences for the existing system and our rule of law do not have to be a zero-sum game.

Instead, state systems of justice can, at their heart, enshrine principles of fairness, human rights and independence of decision-making that will be the "kite mark" or gold standard of a justice system that has integrity. This will continue to be of particular importance when it comes to crime and punishment, family arrangements for children, and reputation-affecting disputes.

---

<sup>2</sup> *R (on the application of Bridges) v South Wales Police* [2020] EWCA Civ 1058.

Let us not forget the practical realities facing many of our court systems. Looking at England and Wales, the current court backlog in the Crown Court is now over 70,000 cases.<sup>3</sup> This is having a huge and negative impact on complainants, witnesses, and defendants alike, whilst the wider public interest in justice being seen to be done and relatively swiftly is not being served. In March 2024, an independent review of the laws and rules on disclosure of material in criminal cases headed by Jonathan Fisher KC recommended an overhaul of the law to reflect the necessity of using AI and assistive technologies to sift and catalogue material relevant to the issues in the case but not relied upon by the Prosecution.<sup>4</sup>

The purpose of this paper is to try to set out some principles for the use of AI technology in our systems, and to propose a type of governance to oversee their safe and ethical use. We have now gone far past the stage where the pace and scale of change should be governed by current limitations in LLMs (large language models) such as hallucinations. This paper is written on the assumption that AGI (artificial general intelligence) will become a reality within a decade. The real question for us is whether to take small incremental steps now (e.g., automated decision-trees) or to make a great leap forward. The latter option may well be difficult to achieve because of the complexity of state procurement processes, but the former option risks placing justice systems in a constant state of “catch-up” with the rest of the world.

## PART ONE

### **I. Big picture: the current problems and the potential of AI to solve them**

The challenging situation faced by many jurisdictions across the world when it comes to caseload pressures continues to be a reality. As mentioned above, in England and Wales, the Crown Court criminal case backlog is now a record high of over 70,000 and rising, compared to a caseload of about 45,000 cases in the years before the pandemic.<sup>5</sup> Victims and witnesses

---

<sup>3</sup> Claire Brader, “Reducing the Crown Court backlog”, House of Lords Library, 13 March 2025, <https://lordslibrary.parliament.uk/reducing-the-crown-court-backlog/#:~:text=The%20crown%20court%20backlog%20reached,their%20cases%20to%20be%20resolved.>

<sup>4</sup> Jonathan Fisher KC, “Independent Review of Disclosure and Fraud Offences”, 20 March 2025, [https://www.gov.uk/government/publications/independent-review-of-disclosure-and-fraud-offences.](https://www.gov.uk/government/publications/independent-review-of-disclosure-and-fraud-offences)

<sup>5</sup> Brader (n 3).

face waits of several years before being able to give evidence, with a resultant decline in the quality of evidence and lack of resolution for defendants, complainants, and their families.

One of the main causes of this delay, apart from lack of available courts, barristers, and judges, is the sheer time taken to manually analyse digital data that increasingly forms the evidential backdrop to all cases. The proper requirement for the Defence to be able to see all relevant material in an investigation to ensure that the prosecution is proceeding fairly, and that the defendant can fully defend themselves, means that cases take far longer to get to trial than in a pre-digital age. Although the Crown Court uses a digital document system, the rest of the investigative process has not seen a significant deployment of automated technologies that would lead to greater speed and efficiency.

In the criminal courts, the uses of AI and automated systems are many and varied. In Annex 1 of this paper, I outline some proposed practical uses for assistive technology in the administration of criminal justice in a joint submission to the Leveson Review of Criminal Procedure that I made with William Rees, former UK Barrister and a Deployment Strategist at Palantir Technologies. Some of the technology that should be deployed is most accurately described as assistive technology, rather than AI in its full sense, but the self-generation of important administrative emails and other messages and the drafting of court orders without either any or any initial human input would involve the use of AI.

## **II. Sectors in which AI has already been / should be deployed**

At the risk of sounding facile, it must be said that the potential impact of AI is extraordinary and beneficial. While the focus of this paper is adjudication (i.e., the work of judges and the operation of the courts), it is important to stress that AI presents serious opportunities for the entire legal system. The purpose of this section is to highlight the individual sectors in addition to judges and the courts which stand to benefit from widespread AI usage, with a particular focus, where it exists, on work that is already underway. The hope is that they will follow the lead of the judiciary and adopt sector-specific policies to ensure safe AI use and improve access to justice.

## A. Judiciary

Judicial Guidance for the Responsible Use of AI in Courts and Tribunals was issued in December 2023.<sup>6</sup> The Guidance is not overlong, which should make it more accessible for daily use, and pulls no punches when it comes to US influences!

In summary, it firstly asks Judges to ensure they have a “basic understanding” of AI tools and their capabilities. Important limitations are noted: users are reminded that public AI chatbots produce the most likely combination of words, not the most accurate answer. AI tools are not a good way of obtaining new and unverified information but are more useful as confirming material the user would recognise as correct. The quality of answers will depend on the nature of the prompts received, so it may be inaccurate, biased, or incomplete. There is a warning that current LLMs will have a “view” on law that is based heavily on US law.<sup>7</sup>

The Guidance makes it clear that judges should not enter information into a public AI chatbot that isn’t in the public domain, or which is private and confidential. Chat histories should be disabled and that no permissions should be given to access information on devices being used.<sup>8</sup> Any disclosure should be reported as a data incident.<sup>9</sup>

The accuracy of information provided to judges via AI tools should always be checked, and that errors and biases in the training data used by LLMs need to be understood and corrected.<sup>10</sup>

When it comes to extrinsic issues, judges are reminded to be ready to remind lawyers of their obligations to independently verify the accuracy of their research via AI.<sup>11</sup> Litigants in person, who rarely have the necessary verification skills, will need to be asked about whether accuracy checks have been made.<sup>12</sup> Additionally, continuous training for judges and lawyers

---

<sup>6</sup> Courts and Tribunals Judiciary, “Artificial Intelligence (AI): Guidance for Individual Office Holders”, 12 December 2023, <https://www.judiciary.uk/wp-content/uploads/2023/12/AI-Judicial-Guidance.pdf>.

<sup>7</sup> *Ibid*, p 3.

<sup>8</sup> *Ibid*.

<sup>9</sup> *Ibid*, p 4.

<sup>10</sup> *Ibid*.

<sup>11</sup> *Ibid*, p 5.

<sup>12</sup> *Ibid*.

on the limitations and proper use of AI tools is essential. Implementing robust oversight mechanisms can further ensure the integrity of AI generated information in legal proceedings.

Sensibly, the Guidance makes specific reference to fake material, whether it be text, images or video.<sup>13</sup> Whilst forgery has always been an issue for courts to handle, the new challenges posed by deepfake technology must be recognised.

The examples given for potential uses and risks of Generative AI fall into two basic categories: AI can be useful in summarising text and administrative tasks, but risky when it comes to legal research and analysis. The Guidance is a useful first step along the road, but much more will need to be done to fully train judges and tribunal chairs as to the opportunities and risks of using AI.

Legal research databases are now notably leveraging generative AI, specifically retrieval-augmented generation (RAG), to enhance the accuracy and reliability of their outputs. RAG works by combining text generation with document retrieval, ensuring that the generated context is comprehensive and verifiable. This process involves the AI generating text based on input queries and simultaneously retrieving relevant legal documents, case law, and statutes from vast databases. By integrating these sources into the generated text, RAG helps ensure that the information provided is accurate and supported by authoritative references.

## **B. Legal Professionals**

When it comes to legal practitioners themselves, professional representative bodies are already generating advice for members. For solicitors in my jurisdiction, for example, in August 2024 the Law Society of England and Wales issued updated advice on generative AI.<sup>14</sup> In summary, whilst the guidance emphasises the opportunities of increased cost efficiency as these increasingly affordable and sophisticated tools become available, it also sets out examples of both new and pre-existing risks, which are a helpful guide to lawyers.

---

<sup>13</sup> *Ibid.*

<sup>14</sup> The Law Society, "Generative AI: the essentials", 7 August 2024, <https://www.lawsociety.org.uk/topics/ai-and-lawtech/generative-ai-the-essentials>.

Firstly, the question of the infringement of intellectual property and copyright, trademark, patent and related rights, plus the disclosure or misuse of confidential information. Secondly, data protection and privacy risks. Thirdly, cybersecurity and hacking risks. Fourth, the generation of misleading, inaccurate or false outputs by generative AI. Fifth, concerns about bias and ethics, leading to unfair or discriminatory results, plus ESG considerations. Sixth, reputational or brand damage to the lawyer or firm concerned.<sup>15</sup>

Helpfully, the Law Society sets out a checklist for practitioners considering the use of generative AI. The list starts with a fundamental question: for what purpose is the AI tool being used, before going on to outline the desired outcome of using it. At all times, professional obligations under the Solicitors' Code of Conduct must be observed when it comes to the use of such tools. It is recommended that practitioners check the data management and security standards of the AI tool, and that rights over training data, AI prompts and outputs are established.<sup>16</sup>

The onus is on the lawyer to check whether the AI tool is a closed system within the boundaries of the law firm or whether it operates as a training model for third parties. This means that the data and information processed by the AI tool are kept private and secure within the firm's internal systems, without being shared externally. It is also recommended that prior to use of the tool, there is a discussion with the client as to its use with particular consideration given to the risks mentioned earlier in the Guidance. Client engagement letters must now include further provisions for the use of AI tools, ensuring transparency and client consent. These provisions should disclose AI tool usage, assure data privacy and security, discuss potential risks, and obtain explicit consent. What is the insurance position too-that needs to be established by the lawyer. This potentially involves confirming their professional liability insurance covers potential risks associated with AI usage, such as data breaches or errors in AI-generated outputs. Finally, it is advised that the outputs are reviewed for accuracy, factual correctness and bias.<sup>17</sup>

So far, for England and Wales solicitors, their Regulatory Authority has not developed any specific guidance on generative AI use or disclosure of use for clients. This is a situation that

---

<sup>15</sup> *Ibid.*

<sup>16</sup> *Ibid.*

<sup>17</sup> *Ibid.*

Regulators should be addressing urgently, bearing in mind the rapid pace of development and use here. The Law Society, however, have developed a useful document that has applicability not only to the profession but also to those procuring AI tools for the administration of justice itself.

### C. Tech Companies

AI is already being utilised in the practice of law, assisting with tasks ranging from document review to legal research to client communication. This utility is only going to increase as products become more sophisticated and tailored to the responsibilities of legal professionals.

In addition to general-purpose tools (i.e., tools trained to generate human-like responses to generic prompts such as ChatGPT and Claude), tech companies are developing specialised tools, specifically designed to assist in the research or practice of law.

These include **Harvey**, a domain-specific AI tool for law firms and legal professionals launched by former attorney Winston Weinberg and AI researcher Gabriel Pereyra in 2022; **Kira**, a machine learning-based AI tool specifically designed to review legal documents; **Oliver**, an AI-driven personal assistant launched in September 2024 trained to conduct legal research, draft contracts, and analyse legal documents; **Reggi**, a generative AI tool solely trained on a database of over 16 million laws and regulations to advise users on regulatory compliance across a range of jurisdictions; and **Spellbook**, an AI platform that uses ChatGPT and other LLMs to draft and review legal documents.

In their 2023 State of Practice survey, however, Bloomberg Law found that while most law firms reported internal talks to better understand the utility of AI, only 11% had purchased or invested in generative AI technology, only 7% were developing their own generative AI technology, and only 7% were encouraging the use of generative AI in practice.<sup>18</sup> This stems, at least in part, from concerns about the safety and reliability of AI (e.g., deepfakes, hallucinations, breach of privacy, and data breach), a particular concern considering the strict professional responsibilities legal practitioners are subject to. A 2024 study by a group of

---

<sup>18</sup> Bloomberg Law, “2023 State of Practice: Practice in the New Era”, <https://pro.bloomberglaw.com/insights/technology/2023-state-of-practice-practice-in-the-new-era/>.

researchers at Stanford University’s RegLab and Human-Centred Artificial Intelligence found that LLMs using a retrieval augmentation system or RAG created by Thompson Reuters and Lexis Nexis, two well-established players in legal research tools and publications, still hallucinated (i.e., generated inaccurate information) in general terms about 18% of the time, even though the RAG system is designed to reduce such errors.<sup>19</sup>

These mistakes have real life consequences. In one particularly prominent case, a lawyer of the New York firm Levidow, Levidow & Oberman faced sanction for submitting to federal court a 10-page brief in a personal injury airline dispute.<sup>20</sup> The brief cited more than half a dozen ostensibly relevant legal authorities, none of which, it transpired, existed, having been erroneously generated by ChatGPT. A major task for tech companies aiming further to penetrate the legal market, therefore, is to achieve a level of legal sophistication sufficient to assuage these concerns.

#### **D National and International AI Regulation:**

If law and justice systems are going to await sector regulation for action by national and international bodies, it is going to be a long one. At Annex 2, I attach a recent lecture I delivered in London about the current state of AI regulation internationally. The fragmented national and international regulatory landscape means that unless individual sectors such as justice and legal services act on oversight and regulation, then the proliferation of AI will carry considerable risk. This is not an argument against the use of machines in legal processes, but a warning that without proper thought, then the benefits of automation will be outweighed by the detriments.

To establish some parameters for action, it is worth looking back for a moment at the effect of previous generations of technology on the administration of justice. In summary, I contend that the “devolve and forget” approach of previous generations when it comes to automation and reliance on tech systems with no effective supervision, is riddled with risks and failures that can cause serious personal, economic and political problems.

---

<sup>19</sup> Varun Magesh et al. “Hallucination Free? Assessing the Reliability of Leading AI Research Tools”, *Empirical Legal Studies*, (forthcoming 2024).

<sup>20</sup> Sara Merken, “New York lawyers sanctioned for using fake ChatGPT cases in legal briefs”, June 26 2023, <https://www.reuters.com/legal/new-york-lawyers-sanctioned-using-fake-chatgpt-cases-legal-brief-2023-06-22/>.

## PART TWO

### **I. Inhuman systems and dodgy IT programs: lessons learned from pre-existing automated systems in England & Wales.**

What evidence do we have so far of the impact of automated systems on existing justice systems? The Single Justice Procedure (SJP) in England and Wales, which was introduced by means of the Criminal Justice and Courts Act 2015, allows prosecutors to deal with cases via a special procedure, where the cases involve adult defendants accused of lesser offences that cannot result in a prison sentence, such as speeding, driving without insurance, TV license evasion and train fare evasion.<sup>21</sup>

Defendants receive a notice containing the charge via post, with a statement setting out the facts of the offence and guidance on what steps to take, including rights to a lawyer. There is the option to plead guilty online or by post, or to ask for a court hearing. If they wish to plead not guilty, there must be a court hearing, save for cases where Defendants seek to argue in person against a disqualification from driving for example.

If they plead guilty or do not respond within the 21-day time limit, their case will be dealt with via the SJP. SJP cases are dealt by a single magistrate, advised by a professional lawyer as they deal with the case on the papers, with prosecutor or defendant present.

Many aspects of the SJP, such as trial in absence or the use of written statements as evidence, are not new, but the new elements are that one magistrate deals with the case, rather than a minimum of two; it need not be heard in open court and there is no fixed date for the hearing as it can be heard any time from 21 days after the notice was served on the defendant.

From April 1, 2019, to September 30, 2023, the SJP received 3,102,392 criminal cases, including 609,164 via the digital service.<sup>22</sup> There has been an issue regarding public access and journalistic scrutiny, but HM Courts and Tribunals Service argues that pending court lists can be viewed online by the public and press and media contacts receive detailed lists daily,

---

<sup>21</sup> UK Government, "Single justice procedure notices", <https://www.gov.uk/single-justice-procedure-notice>.

<sup>22</sup> UK Government, "Fact sheet: Single Justice Service".

plus weekly results of all SJP cases. It is intended to further expand use of this automated procedure.

The main criticism of the SJP system is one related not to its internal operation, but to external human factors, leading to unfairness when dealing with vulnerable defendants. Very recently, the former Lord Chief Justice of England and Wales, Lord Thomas of Cwmgiedd said that the SJP was “*tilted against fairness*” and should be overhauled and that time is taken to deal properly with all cases.<sup>23</sup> He noted that mitigation letters from defendants, outlining their difficulties, were routinely not read by prosecutors. “*You should never convict someone in a conveyor belt system*”, he said. “*Slowing it down a bit and letting the journalists have a good look over the system is the easiest and simplest way to address public concerns.*”<sup>24</sup>

An Evening Standard investigation in 2024 found that a pensioner with Alzheimer’s was prosecuted over £3.34 owed to the DVLA under the SJP, a domestic abuse victim was fined under SJP when her controlling ex-boyfriend didn’t insure the car, and an unwell pensioner was prosecuted for failing to pay a TV licence when caring for his ill wife.<sup>25</sup> All these controversies are real and concerning, but they once again relate more to the human operation of an automated system rather than the automated system itself. Failures by human magistrates to exercise discretion and to read relevant documents have led to concerns about the system itself.

In 2024, the UK Parliament took the unprecedented step of passing legislation that has the effect of removing or quashing of criminal convictions for theft, fraud, and false accounting recorded against hundreds of sub-postmasters who were responsible for local branches of the Post Office network in the UK.<sup>26</sup> This measure was taken after the scale and extent of an underlying IT failure in an accounting system used across the network became increasingly

---

<sup>23</sup> Charles Hymas, “Secret court hearings are unfair and in need of reform, says former chief justice”, *The Telegraph*, 19 August 2024, <https://www.telegraph.co.uk/news/2024/08/19/secret-court-hearings-unfair-need-reform-lord-thomas/>.

<sup>24</sup> *Ibid.*

<sup>25</sup> See e.g., Tristan Kirk, “Pensioner, 93, with dementia convicted in fast-track SJP courts for not having car insurance”, *Evening Standard*, 7 March 2025, <https://www.standard.co.uk/news/crime/single-justice-procedure-dvla-pensioner-dementia-care-conviction-car-insurance-b1215124.html>.

<sup>26</sup> Post Office (Horizon System) Offences Act 2024.

apparent after a long-running campaign by aggrieved postmasters and a number of England and Wales Court of Appeal cases where convictions were overturned.

The reason for this were flaws or bugs within the software system that had been developed by ICL, now Fujitsu, in the late 1990s. This was compounded by an insistence by the Post Office and its leadership that there was no issue with the IT system, which meant that there were also failures of disclosure and scrutiny via the criminal justice court process.<sup>27</sup> This specific issue is being addressed by a Public Inquiry led by Sir Wyn Williams, a former High Court Judge,<sup>28</sup> and which is yet to report but which no doubt will provide a full analysis of the problem, but it is already clear that the criminal justice process, which for years has allowed the product of IT to be admissible without further verification or testing, has emerged poorly from this scandal.

One of the key questions is why there was no forensic examination by suitable experts of the system and its outputs. The question of the ability to challenge IT evidence is one that has had a wider impact on public confidence both in machine processes and the ability of the system itself to be transparent and to acknowledge failure. The showing of a major TV drama based upon the scandal, “Mr Bates and The Post Office” over Christmas 2023 attracted millions of UK viewers.<sup>29</sup> This, therefore, is not an issue of marginal public importance, and highlights the human context in which machines and technology old and new must operate.

It is important, as such, to draw distinctions between intrinsic and extrinsic factors. What can be said is that the speedy nature of automated procedures means there is less opportunity for there to be an enlightened intervention and more risk of error being compounded. In the case of the SJP, the need to proceed more carefully when a Defendant is vulnerable is clear. The challenge will be how to effectively identify such cases without losing the benefits of greater efficiency. In the case of Horizons, the initial failures in the system were ignored and then covered up, with consequent damage to a major public organisation and to the court system itself. The lessons for the future use of AI systems are clear: without careful planning as to

---

<sup>27</sup> See generally, Mark Sweeney, “What is the Post Office Horizon IT scandal all about?”, *The Guardian*, 7 January 2024, <https://www.theguardian.com/business/2024/jan/07/what-is-the-post-office-horizon-it-scandal-all-about>.

<sup>28</sup> See Post Office Horizon IT Inquiry, <https://www.postofficehorizoninquiry.org.uk/>.

<sup>29</sup> Mr Bates v The Post Office, <https://www.pbs.org/wgbh/masterpiece/shows/mr-bates-vs-the-post-office/>.

precisely how they are to be integrated, they are less likely to deliver justice in the eyes of the public.

As things stand, the observations made by Tim Clement-Jones in *“Living with the Algorithm”*, neatly sum up the position when it comes to AI governance in the UK public sector.<sup>30</sup> When summing up the findings on the UK Committee For Standards In Public Life’s 2020 report: *“It found that, despite the GDPR (General Data Protection Regulation), the data ethics framework, the OECD principles and the guidelines for using artificial intelligence in the public sector, the Nolan principles of openness, accountability and objectivity were not, but should be, embedded in AI governance in the public sector.”* The Nolan Principles of Public Life, published in 1995 are: honesty, integrity, objectivity, accountability, selflessness, openness and leadership, and underpin the required standards of UK public office holders.

In early 2025, JUSTICE, the UK’s leading legal research charity, produced a report on AI and justice which set some helpful suggested standards and parameters that could be the benchmark for the safe application of AI systems in the administration of justice.<sup>31</sup> In summary, they propose two clear requirements for those looking to use AI in justice systems:

1. **Goal led:** ensure the tool is clearly aimed at improving one or more of the justice system’s core goals of access to justice, fair and lawful decision-making and transparency.
2. **Duty to Act Responsibly:** ensure all those involved in creating and using the tool take responsibility for ensuring that the rule of law and human rights are embedded in each stage of its design, development and deployment.

I will focus in on the question of adjudication and the administration of justice, to set out some principles and applications that could be used.

## **II. Guiding principles for adjudication: developing the right criteria for determining which cases are suitable for automation versus human adjudication**

---

<sup>30</sup> Tom Clement-Jones, *Living with the Algorithm: Servant or Master?: AI Governance and Policy for the Future*, Oxford, 2024.

<sup>31</sup> JUSTICE, “AI in our justice system”, January 2025, <https://justice.org.uk/ai-in-our-justice-system/>.

As AI continues its onward march, will our courts, currently used to the murmur of advocates conferring and the rustle of case files, soon be complemented by the quiet hum of servers processing routine cases? Maybe, but how easy is this going to be? The efficiency gains of automated systems are appealing, especially in jurisdictions facing significant backlogs such as England and Wales but requires criteria that protect the fundamental principles of justice and human rights. Balancing innovation with ethical considerations is crucial to maintaining public trust in the legal system.

Consider the magistrates' courts of England and Wales on a Monday morning. The list is filled with a variety of cases. Already, thanks to the SJP discussed above, straightforward cases—a speeding ticket or a failure to pay a television licence—are processed through an automated system. Can we go one stage further and fully automate the sentence outcomes too, by automatically drawing on a legal precedent and sentencing guidelines that will swiftly and consistently handle these routine cases?

As we move through the court list, however, cases of increasing complexity may appear. Picture a case involving a young offender charged with shoplifting, but whose circumstances reveal a history of neglect and mental health challenges. It is in these instances where the nuanced judgement of a human magistrate, judge, or juror in Crown Court cases remain important. The ability to perceive the subtle cues in a defendant's demeanour, to weigh the intangible factors that may influence a just outcome, these are qualities that, as yet, remain beyond the reach of even our most sophisticated AI systems. My understanding is that emotional responses are increasingly being sought and gained from machines, but this is technology that is still developing, and which will take time to command public confidence.

In considering the integration of AI, it is paramount that we establish a set of rules to assess the suitability of cases for automated processing. These rules must be grounded in the fundamental principles of justice, fairness, and transparency that underpin our legal traditions. They should also be flexible enough to adapt to the rapidly evolving landscape of AI technology, whilst steadfast in their commitment to protecting the rights of all individuals who come before the courts.

The first and perhaps most crucial rule in assessing AI suitability is the principle of **'algorithmic humility'**. This rule stipulates that any AI system deployed in a judicial context must be programmed with an acute awareness of its own limitations. The system should be capable of recognising when it is operating at the edges of its training data or encountering scenarios that fall outside its realm of competence. For example, if an AI system processing a routine traffic offence detects language in the defendant's statement suggesting mental health concerns or unusual circumstances, it should immediately flag the case for human review. This self-awareness is not just a safeguard against erroneous judgments, but a fundamental ethical requirement for any AI system entrusted with matters of justice. Incorporating this principle ensures that AI complements rather than compromises the integrity of judicial processes.

Consider the case of Mr Johnson, who contests an automatic disqualification for a drink driving offence on for the “Special Reason” (a concept known to English criminal law in this regard) that he was rushing his dangerously ill wife to a very nearby hospital. An AI system, no matter how sophisticated, may struggle to fully appreciate the nuances of such a situation. The presence of keywords like 'emergency' or 'hospital' in Mr Johnson's statement should trigger the AI to recognise that this case falls outside its purview and requires human evaluation.

This demonstrates the AI's awareness of its limitations and ensures that cases requiring human empathy and judgment are appropriately handled. Context is crucial in judicial matters, as it allows for a deeper understanding of the circumstances surrounding each case, ensuring that justice is administered fairly and compassionately. In LLMs and generative AI, context is likewise extremely important, as it enables these systems to generate more accurate, relevant, and coherent responses by understanding the nuances and subtleties of the input they receive.

The second rule relates to the principle of **'opt-in consent'**, or “**informed choice**”. In the initial stages of AI integration, participation in automated judicial processes should be on a strictly voluntary basis. Defendants should be given a clear and informed choice between traditional human-led proceedings and AI-assisted adjudication, including the benefits and limitations of each. This opt-in approach serves several purposes. Firstly, it respects individual autonomy, allowing those who are comfortable with AI systems to benefit from

potentially faster processing times, whilst ensuring that those who prefer human adjudication are not forced into an automated system against their will. Secondly, it provides opportunity for data collection and system refinement. By comparing outcomes between AI-processed and human-processed cases, we can continuously improve the accuracy and fairness of the AI systems.

However, the opt-in process must be carefully managed to avoid creating a two-tiered justice system. Clear information must be provided about the nature of the AI system, its decision-making process (often referred to as ‘algorithmic transparency’), and the rights of appeal. For example, a defendant opting for AI adjudication of a parking fine should be informed that while the process may be quicker, they retain the right to appeal the decision to a human judge if they are unsatisfied with the outcome. This transparency is important in maintaining public trust in the justice system as it evolves to incorporate new technologies.

The third rule in our framework is the principle of '**contextual sensitivity**'. AI systems must be capable of recognising and flagging cases where broader societal or systemic issues may be at play. For example, if an AI system processing traffic offences notices a statistically significant increase in speeding tickets issued at a particular location, it should not simply process these cases in isolation. Instead, it should flag this pattern for human investigation, as it may indicate issues with road design, signage, or faulty speed cameras. This ability to identify potential systemic issues is crucial in ensuring that AI systems do not inadvertently perpetuate or exacerbate existing inequalities or flaws in the justice system.

Moreover, contextual sensitivity extends beyond just traffic offences. For instance, in the health and social care sector, an AI system might detect a rise in certain medical conditions within a specific demographic. Rather than merely diagnosing and treating these cases individually, the system should flag this trend for further investigation to determine if there are underlying environmental, social, or economic factors contributing to the increase. This could lead to more effective public health interventions and policies.

Consider the case of a neighbourhood where a disproportionate number of parking fines are being contested. An AI system processing these cases individually might miss the broader context – perhaps unclear signage or a recent change in parking regulations that hasn't been well communicated. By flagging this pattern, the AI enables human authorities to investigate

and address the root cause, potentially preventing unnecessary penalisation of residents and fostering a fairer application of the law. This will be a distinct improvement on the current situation, which largely depends on the anecdotal experience of individual judges and is nowhere as near as comprehensive or systematic as an AI system of this nature.

AI systems could also track the outcomes of contested fines to identify any biases in the adjudication process. For example, if a particular demographic is more successful in contesting fines, this could indicate potential disparities in how fines are issued and contested. By flagging such trends, the AI system can help ensure that the enforcement of parking regulations is equitable and just. In essence, the AI's ability to contextualize and analyse data from multiple angles ensures that it not only addresses immediate issues but also contributes to long-term improvements in policy and enforcement.

The fourth rule in our framework is the '**principle of continuous human oversight**'. While AI systems may be entrusted with certain decision-making processes, there must always be a clear chain of human responsibility and the possibility of human intervention. This rule mandates regular audits of AI decisions, random sampling of cases for human review, and clear processes for appealing AI decisions to human judges. For example, in a system processing minor civil claims, a certain percentage of cases should be randomly selected for review by human judges, regardless of whether the parties involved have requested an appeal. This ongoing oversight serves to maintain the integrity of the system and provides a mechanism for identifying and correcting any systematic biases or errors that may emerge over time.

The fifth rule is the '**principle of ethical transparency**'. Any AI system deployed in a judicial context must be open to scrutiny, with its decision-making processes explainable in clear, non-technical language. This transparency is crucial not only for maintaining public trust but also for ensuring that defendants can effectively challenge decisions if necessary. For instance, if an AI system recommends a particular sentence in a minor criminal case, it should be able to provide a clear explanation of the factors it considered and how they influenced its recommendation. This explanation should be comprehensible to the defendant, their legal representation, and the general public. Additionally, ethical transparency entails regular audits and assessments of AI systems to ensure they remain fair, unbiased, and aligned with the principles of justice.

The sixth and final rule in our framework is the '**principle of adaptive learning**'. While AI systems must operate within strictly defined parameters, they should also have the capacity to learn and improve over time based on feedback from human oversight. However, this adaptive capability must be carefully managed to prevent the emergence of unintended biases or drift from established legal principles. Any significant changes to the AI's decision-making processes should be subject to rigorous testing and approval by a panel of legal experts before implementation.

These six rules – algorithmic humility, opt-in consent, contextual sensitivity, continuous human oversight, ethical transparency, and adaptive learning – will form a robust framework for assessing the suitability of AI integration in our justice system.

#### **When automation can be used: some principles:**

Implementation should begin with a carefully selected pilot programme in a handful of jurisdictions, focusing initially on low-stakes, high-volume cases. However, rather than simply automating these processes wholesale, a system of 'augmented decision-making' that combines AI analysis with human oversight at every stage should be implemented. In determining which cases are suitable for automation versus human adjudication, a taxonomy of case types can be created and then subsequently evaluated with a strong emphasis on cases which should not have any element of automation and be flagged for human review. This would include for example cases involving mental health issues, such as competency to stand trial, mental health defences, and sentencing considerations.

I propose a multi-faceted approach that considers the nature of the case, the potential consequences, and the technological capabilities of AI systems. At the heart of this framework is an initial triage process. This serves as a defence against the potential pitfalls of over-automation, ensuring that cases requiring human discernment are identified early and routed appropriately. The triage process is not a binary sorting mechanism, but a nuanced evaluation conducted by trained legal professionals assisted by AI tools:

**Tier 0: Direct Human Adjudication.** This tier is reserved for cases that demand full human oversight due to their complexity, potential impact, or precedent-setting nature. Examples

include serious criminal cases, constitutional challenges, and cases involving vulnerable individuals such as minors or those with mental health concerns. The case of *R v McNaughton* (1843), which established the insanity defence in English common law, and associated legislation relating to mental capacity and the criminal law exemplifies the type of nuanced evaluation that necessitates an assessment of expert psychiatric evidence by a court.

**Tier 1: AI-Assisted Human Adjudication.** Cases in this tier benefit from AI support in research and analysis, but ultimate decisions remain firmly in human hands. This might include moderate criminal offences, civil cases above a certain monetary threshold, or employment tribunals. In the case of an unfair dismissal claim, AI can conduct sentiment analysis on employee communications to detect potential biases or discriminatory language, cross-referencing company policies with legal standards to identify any procedural violations. In addition, AI systems could analyse years of company records and employment law precedents, but the final weighing of evidence and credibility of witnesses would be conducted by a human judge.

**Tier 2: Human-Overseen AI Adjudication.** This tier represents a more significant role for AI, but with robust human safeguards. Minor traffic cases or small claims court cases might fall into this category. However, as noted above, even seemingly straightforward cases can have hidden complexities. For example, a contested parking fine might not just be about a £60 penalty, but stem from a belief that the parking signage in an area is unclear and potentially discriminatory towards those with visual impairments. Such a nuanced complaint, which might be overlooked by an AI system simply processing the binary fact of a parking violation, would be promptly flagged for human review. This ensures that the subtleties and broader implications of the case are thoroughly examined, maintaining fairness and justice in the adjudication process.

**Tier 3: Fully Automated Processing.** Reserved for the most routine and uncontested matters, this tier would handle cases like uncontested parking fines or automatic statutory penalties. It represents a level of adjudication where AI operates independently, handling straightforward cases with efficiency and speed. However, even at this level, safeguards are important. Every automated decision must be accompanied by a clear and concise explanation of the right to appeal for human review, ensuring that the individuals are aware of their options should they wish to contest the decision. Additionally, regular audits must be conducted to maintain the

integrity of the system, ensuring that it functions correctly and fairly. These measures are essential to uphold transparency and trust in the fully automated processing tier, guaranteeing that justice is served even in the most routine cases.

Therefore, the success of the framework hinges foremost on its ability to recognise its own limitations. This self-awareness must be built into the system at every level. For instance, if the AI system processing a Tier 2 case detects language in the defendant's statement suggesting mental health concerns, it would automatically escalate the case to Tier 0 for direct human review. This ensures that sensitive issues are handled with the necessary human empathy and understanding. Similarly, if an unusual pattern of decisions emerges in Tier 3 cases, the system would flag this for human investigation, potentially uncovering systemic issues that require address. This proactive measure could potentially uncover systemic issues that require immediate attention and resolution.

Critics might argue that such a system risks creating two-tiered justice, with 'real' justice for serious cases and 'automated' justice for minor ones. However, I would contend that by maintaining human oversight at all levels and ensuring easy access to appeals, we are not creating a two-tiered system, but rather optimising resource allocation. This could lead to more thorough consideration of all cases, as judges are freed from routine matters to focus on complex issues that truly require human insight and judgement. By doing so, the system ensures that every case, regardless of its perceived simplicity, receives the attention it deserves, thereby upholding the principles of fairness and justice across the board.

Others might express concern about the potential for AI systems to perpetuate biases present in historical case data. This is indeed a serious consideration, but one that this framework could be designed to address through its multi-layered approach with human oversight at various stages, providing multiple opportunities to identify and correct biases. For instance, at Tier 2, human adjudicators can review cases flagged for potential bias, ensuring that any problematic patterns are promptly addressed. Additionally, the framework could incorporate regular bias audits, where independent experts analyse the AI's decision-making process to detect and mitigate any emerging biases. I deal in more detail with ways in which to mitigate AI system bias later in this paper.

Furthermore, the open-source nature of the AI systems allows for continuous scrutiny and

improvement by a diverse community of experts. This transparency ensures that the AI's algorithm and data sets are constantly reviewed and refined, incorporating feedback from a wide range of perspectives. By engaging a broad spectrum of stakeholders, including legal professionals, ethicists, and technologists, the system can evolve to be more equitable and just. In England and Wales, **a joint board of judges, lawyers and technologists** could be given the task of reviewing and monitoring AI systems applying this framework, which recognises that justice is not just about efficiency, but about fairness, accountability, and maintaining public trust in our legal institutions.

### III. How explainability can be used to increase confidence in verdicts

The integration of AI into the justice system presents a double-edged sword. On one hand, we have the pressing need for transparency and accountability in AI-driven legal decisions. On the other, we face the risk of providing too much information, potentially arming unscrupulous actors with the tools to manipulate the system. This tension creates a complex landscape that requires careful navigation.

As referred to in my first paper, in the case of *State v. Loomis*,<sup>32</sup> the Wisconsin Supreme Court upheld the use of the COMPAS risk assessment tool in sentencing decisions. The court's ruling hinged partly on the condition that the risk scores would not be the determining factor in sentencing. However, the proprietary nature of the COMPAS algorithm meant that neither the defendant nor the court could scrutinise its inner workings. This lack of transparency raised significant due process concerns and exemplifies the need for explainable AI in legal contexts.

Now, imagine a future where AI systems used in sentencing are fully explainable. Defence barristers could potentially use this information to craft narratives that play to the AI's known decision-making patterns. This is not a hypothetical concern; we have already seen similar behaviours in other domains where AI plays a significant role. For example, in search engine optimisation (SEO), professionals have long sought to understand and manipulate Google's algorithms to improve their clients' search rankings. In the legal context, the stakes are considerably higher.

---

<sup>32</sup> 881 N.W.2d 749 (Wis. 2016).

Consider a hypothetical scenario where an AI system is used to recommend sentencing in criminal cases. If the system is known to place significant weight on factors such as community ties and employment status, defence lawyers might coach their clients to emphasise these aspects, potentially even encouraging them to take on community roles or short-term employment shortly before their trial. While these factors may indeed be relevant to sentencing decisions, their deliberate manipulation could lead to outcomes that do not truly reflect the spirit of the law.

The trade-off becomes clear: increased explainability enhances fairness and allows for meaningful challenge of AI decisions, but it also provides a potential playbook for system manipulation. This dilemma is not unique to the legal sector. In the financial industry, the use of explainable AI models for credit scoring has led to the emergence of 'credit repair' services that advise clients on how to artificially boost their scores. While some of these services offer legitimate financial advice, others border on fraud, exploiting the known parameters of the scoring algorithms. This raises important ethical questions about the fairness and integrity of AI-assisted judicial processes, highlighting the need for robust safeguards and oversight mechanisms to ensure that justice is served equitably.

Suppose an AI sentencing support system is implemented in a jurisdiction, and its explainability features reveal that it places significant weight on expressions of remorse and plans for rehabilitation. In a drug possession case, two defendants with similar offences and backgrounds receive markedly different sentence recommendations. The AI recommends a lighter sentence for Defendant A, who submitted a lengthy letter of apology and a detailed plan for drug rehabilitation. Defendant B, who maintained his innocence, receives a harsher recommendation. This underscores the importance of models being able to capture and comprehend context, as the nuances of each case and the genuine intentions behind defendant's actions are critical for ensuring just and appropriate sentencing outcomes.

On the surface, this might seem fair – remorse and willingness to reform are legitimate factors in sentencing. However, this transparent system could lead to a situation where genuine remorse becomes indistinguishable from coached performances. Moreover, it could unfairly disadvantage defendants who maintain their innocence or come from cultural backgrounds where public displays of emotion are less common.

To address these challenges, I propose a framework of "Calibrated Transparency" for legal AI systems:

1. Provide different levels of explanation to different stakeholders. For instance, judges and legal professionals might receive detailed breakdowns of the AI's decision-making process, while defendants and the public receive more general explanations focused on the key factors influencing the decision. The varying levels of explanation should all be published, in accordance with the principle of Open Justice.
2. Regularly update the specific weights and thresholds used by the AI system, making it harder for individuals to game the system over time. This could be analogous to how social media platforms frequently update their algorithms to prevent manipulation, providing a form of algorithmic transparency and effectively an audit trail.
3. Implement systems that look at patterns of behaviour over time, rather than just at the moment of trial. This could help distinguish between genuine life changes and last-minute attempts to influence the AI.
4. Require human judges to provide their own reasoning alongside the AI's recommendation, especially when they deviate from it. This maintains the role of human judgment in the legal process.
5. Develop programs to train legal professionals on how to ethically interact with AI systems, emphasising the importance of genuine representation over "system gaming".

The trade-off here is clear: we sacrifice some degree of absolute transparency for a system that is more robust in combating manipulation. The key is to find the right balance – enough explainability to ensure fairness and accountability, but not so much that it undermines the integrity of the legal process. As we implement these systems, we must remain vigilant and adaptive. Regular audits, both of the AI's performance and of case outcomes, will be crucial.

We must be prepared to adjust our approach as new forms of system gaming emerge or if we find that the balance between explainability and system integrity needs recalibration.

However, the implementation of such systems must be balanced against the right of defendants to present their best case. We must be careful not to create a situation where legitimate defence strategies are unfairly penalised. This is where the importance of explainability comes to the fore. Explainability techniques such as integrated gradients and Shapley value analysis offer tools for understanding the decision-making processes of complex AI systems. These methods allow us to attribute importance to different input features, providing insight into which factors most influenced a particular decision. In the context of legal AI, these techniques could be used to create detailed, comprehensible explanations of how a verdict or sentence recommendation was reached.

For example, in a case like *R v L and Others*,<sup>33</sup> where the Court of Appeal had to consider the complex interplay of factors in a modern slavery case, an explainable AI system could provide a breakdown of how it weighted different elements such as the vulnerability of the victims, the duration of the offending, and the defendant's level of involvement. This explanation could then be scrutinised by human judges, ensuring that the AI's reasoning aligns with legal principles and precedents.

To implement this effectively, there would be a standardised format for AI systems to present their decision-making processes, designed to be readily understandable by legal professionals and, importantly, translatable into layperson's terms for defendants and the public. This would include several key components: (i) a hierarchical breakdown of the factors considered, showing primary, secondary, and tertiary influences on the decision; (ii) quantitative measures of each factor's importance; (iii) comparisons to similar cases, highlighting where the current case aligns with or deviates from precedent; and (iv) identification of any unusual or outlier factors that significantly influenced the decision.

This protocol would serve multiple purposes. First, it would improve transparency, allowing all parties to understand and, if necessary, challenge the AI's reasoning. Second, it would act as a deterrent to overly simplistic attempts at gaming the system, as the explanation would

---

<sup>33</sup> [2013] EWCA Crim 991.

reveal if certain factors were being given undue weight. Third, it would provide valuable data for ongoing refinement of the AI system, helping to identify and correct any emerging biases or flaws in its reasoning. However, we must acknowledge that increased explainability is not without its risks. Detailed explanations of AI decision-making could potentially provide a roadmap for more sophisticated attempts at system manipulation.

Moreover, we must consider the broader implications of highly explainable AI in the justice system. There is a risk that over-reliance on AI explanations could lead to a kind of "algorithmic jurisprudence", where human judges increasingly defer to AI reasoning without applying their own critical thinking. To mitigate this, AI explanations should be treated as advisory rather than determinative, with a requirement for human judges to articulate their own reasoning, especially where they deviate from the AI's recommendation.

The strategic advantages sought by litigators and the need for explainability in AI verdicts are two sides of the same coin. By developing sophisticated systems for detecting potential gaming, alongside protocols for explaining AI decisions, we can create a more transparent, fair, and trustworthy integration of AI in the justice system. In England and Wales, the joint board or committee that I have already recommended be set up in Part III could be given a key role here too.

#### **IV. Compiling training data: a question of attribution**

##### **A. Introduction**

The compilation of training data is a critical factor in ensuring the accuracy, fairness, and ethical integrity of automated judicial decision-making systems. We begin by outlining the entities that have the capability to compile the necessary data and evaluating the merits of assigning this responsibility to them. It is essential to recognise that the compilation of training data for generative AI based systems is intrinsically linked to the development of these systems. Compiling the data and developing the system are not tasks undertaken in isolation. This linkage presents both explicit and implicit challenges.

Explicitly, the focus must be on the capability, bias, accuracy, relevance, and ethical standards adhered to by the entities and the individuals responsible for data compilation. Implicitly, it will be a requirement of the entities tasked with this responsibility to be active

participants in the development process of the system. This deep involvement is crucial because the process is not merely one of passing data to developers; it necessitates bridging the expertise gap due to the subjective nature of the data, the responsibilities involved, and externalities associated with the system. For instance, data compilers must work closely with developers to ensure that the nuances and context of the data are fully understood and appropriately integrated into the AI's algorithms. Closing the full cycle of the considerations, an important consideration is that entities will inevitably influence the development of the system itself. Their active participation ensures that the system evolves in a way that accurately reflects the complexities and subtleties of the real-world scenarios it is designed to adjudicate.

As such, when evaluating the capabilities, advantages, and disadvantages of each entity in the data compilation process, it is crucial to consider the methodology for training and testing the data with the involvement of these entities. This section will begin by outlining the entities capable of compiling the required training data, followed by a discussion of the metrics and consideration of assigning this responsibility to each. Finally, it will propose the most appropriate entity or combination of entities to be assigned this role, alongside an examination of the methodologies for training and testing the data to ensure the system's efficacy and fairness.

## **B. Stakeholders Responsible for Compiling Test Data**

Involving the judiciary in compiling data strengthens public trust by ensuring that AI systems adhere to legal standards. The judiciary's oversight role can address concerns regarding fairness and transparency. Hartzog and Selinger highlight that such transparency is key to building institutional credibility, which is essential in AI-based judicial processes.<sup>34</sup> Moreover, the legal expertise of the judiciary ensures that AI systems comply with established laws, as Calo points out, emphasizing the need for specialists to guide AI development.<sup>35</sup>

---

<sup>34</sup> Woodrow Hartzog and Evan Selinger, "Surveillance as a Loss of Obscurity," 72 *Wash. & Lee L. Rev.* 1343 (2015).

<sup>35</sup> Ryan Calo, "Artificial Intelligence Policy: A Primer and Roadmap", 51 *UC Davis Law Review* 399 (2017).

However, there are limitations to depending solely on the judiciary. Relying on historical judicial data can exacerbate existing biases, particularly affecting marginalized communities. Eubanks underscores how biases embedded in historical data can be carried over into AI systems.<sup>36</sup> Additionally, the judiciary may not have the necessary technical knowledge to handle the intricacies of AI development, as Baker (2021) suggests, underlining the importance of interdisciplinary collaboration. This reliance on historical data can perpetuate and even amplify systemic inequalities. For instance, if judicial data reflects past biases against certain groups, an AI system trained on this data might disproportionately target these groups in future predictions or decisions. To mitigate this, AI systems must be designed with robust bias detection and mitigation mechanisms.

### **Interdisciplinary Commission**

An interdisciplinary commission brings together professionals from various sectors, such as technology, law, and ethics, to ensure a well-rounded and balanced approach to data management. Kitchin (2014) argues that collaboration between different fields is vital for handling data-driven systems in an ethical and efficient manner. By incorporating diverse viewpoints, these commissions are better equipped to identify and address biases within AI systems. Benjamin (2019) stresses the risk of systemic biases being reinforced by algorithms unless diverse teams actively work to counteract them.

However, despite the benefits, creating and sustaining an interdisciplinary commission requires substantial time and financial resources. The AI Now Institute (2018) highlights the resource-intensive nature of such oversight, which can lead to slower decision-making and can complicate the development process.

### **Strategic Public-Private Partnership (PPP)**

A Strategic Public-Private Partnership (PPP) provides a practical approach by merging the accountability of government with the innovative capabilities of the private sector. Harvard's Health Systems Innovation Lab emphasizes the advantages of PPPs in pooling resources within health systems while maintaining ethical standards through shared responsibility. Wirtz et al. argue that PPPs foster innovation by blending the flexibility of the private sector

---

<sup>36</sup> Virginia Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*, New York, 2018.

with the governance of the public sector, creating a setting where both innovation and accountability can thrive.<sup>37</sup>

However, balancing the diverse interests of public and private stakeholders poses challenges. Zuboff points out that private interests can sometimes clash with public goals, potentially affecting transparency.<sup>38</sup> Pasquale warns that when private entities control data and algorithms, accountability can diminish, especially when proprietary interests overshadow the need for public transparency.<sup>39</sup>

### **C. Methodology**

#### **Open Source and Transparency**

In this context, the term open-source refers to a transparent and accessible process rather than fully open-source large language models (LLMs). This approach ensures that the models and data are reviewed by key stakeholders, such as an interdisciplinary commission, without being made fully available to the public or general law firms. By adopting a quasi-open-source method, where access is limited to the interdisciplinary commission until proper validation is achieved, transparency is maintained without exposing proprietary data or incomplete models. Pasquale supports balancing transparency with the protection of proprietary interests.<sup>40</sup>

Additionally, incorporating third-party validation services such as start-ups like Warden AI can ensure that the data used is of high quality and free from biases before being applied in legal settings. These external evaluations enhance the system's credibility and ensure adherence to ethical standards, as outlined by Mittelstadt et al.<sup>41</sup>

It is also essential to clarify that a judiciary-specific LLM will not be developed. Instead, existing models such as OpenAI's GPT or Anthropic's Claude can be tailored for legal use,

---

<sup>37</sup> Bernd Wirtz, Jan Weyerer, and Carolin Geyer, "Artificial Intelligence and the Public Sector – Applications and Challenges", 42(7) *International Journal of Public Administration* 596 (2019).

<sup>38</sup> Shoshana Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*, New York, 2019.

<sup>39</sup> Frank Pasquale, *The Black Box Society: The Secret Algorithms that Control Money and Information*, Cambridge (MA), 2015.

<sup>40</sup> *Ibid.*

<sup>41</sup> Brent Mittelstadt et al., "The ethics of algorithms: Mapping the debate", 3(2) *Big Data & Society* (2016).

with oversight from the interdisciplinary body to ensure these adaptations align with legal principles. Calo stresses the importance of customizing domain-specific AI while balancing innovation and ethical oversight.<sup>42</sup>

### **Principles and Approaches for Testing Models**

Establishing clear guidelines for testing models like OpenAI, Anthropic, Mistral, and Gemini is essential. Rather than relying on binary outcomes, testing should accommodate growing complexities. Defining what qualifies as 'true' or 'false' results is key to maintaining rigorous and context-specific evaluations. Floridi provides ethical guidelines that can inform this process, ensuring that models are tested in real-world scenarios.<sup>43</sup>

A solid methodology for testing different versions of these models must also be in place to ensure that each new iteration is thoroughly assessed. This ensures the models are prepared for the complex and evolving demands of judicial applications. Wirtz et al. underscore the need for AI models to adapt and meet new challenges.<sup>44</sup>

### **Continuous Monitoring and Updating**

Once deployed, models need to be continuously monitored to ensure they remain effective in addressing new cases and adapting to evolving legal requirements. Gibney warns of the risks associated with AI systems deteriorating over time if not properly overseen.<sup>45</sup> Regular updates to both the models and their training data are essential to ensure that the AI's responses reflect legal advancements and societal changes. Mittelstadt et al. emphasize that ongoing updates are particularly important in sensitive fields like law, where outdated models could result in serious errors.<sup>46</sup>

### **Synthetic Data Risks**

Training AI models on synthetic data can cause a swift decline in model quality. Gibney, writing in *Nature*, highlights the dangers of relying too much on AI-generated data without

---

<sup>42</sup> *Ibid* (no 35).

<sup>43</sup> Luciano Floridi, "What the Near Future of Artificial Intelligence Could Be", *Philosophy & Technology* (2020).

<sup>44</sup> *Ibid* (no 37).

<sup>45</sup> Elizabeth Gibney, "AI models fed AI-generated data quickly spew nonsense", *632 Nature* 18 (2024).

<sup>46</sup> *Ibid* (no 41).

proper human oversight, cautioning against the risk of "model collapse".<sup>47</sup> This is especially concerning in judicial systems, where maintaining the integrity of data is crucial for ensuring accuracy and fairness.

### **PART THREE**

**The potential unintended consequences of widespread AI use and the measures that can be taken to minimise or alleviate them.**

#### **Part I: Some Unintended Consequences:**

##### **Extrinsic change: Synthetic media and deepfakes**

As mentioned in my first paper, another fundamental extrinsic challenge posed to the justice system by AI is the growing prevalence of synthetic media and deepfakes: namely, fabricated image, video, or audio recording created or altered using AI tools to appear to the reasonable observer to be genuine. This challenge is pressing—the courts are already grappling with its consequences—and so demands a prompt and concerted effort to develop systems capable of ensuring the authentication of evidence and, in turn, maintaining public and professional trust in the judicial process.

The sophistication of deepfake is rapidly improving such that it is becoming increasingly difficult to determine whether an image or other audiovisual is real or fake. This development poses a serious threat to the operation of the adjudicatory process and, specifically, court proceedings, which can be broken down into two interrelated issues. The first, is the provision of fabricated evidence generated using deepfake technology, creating the possibility of unfair trials and wrongful outcomes. The second, is the prospect of lawyers exploiting the uncertainty surrounding fabricated evidence and, in turn, placing an undue burden on opposing counsel to demonstrate evidence is not fabricated with the consequence of crucial evidence being rendered inadmissible. This is a potential practical consequence of the so-called “liar’s dividend”, mentioned in my first Paper.

These issues are not merely academic. The courts are already dealing with deepfake-generated evidence which has disturbed, or at least threatened to disturb, the administration

---

<sup>47</sup> *Ibid* (no 45).

of justice, both in the U.S. and the UK. To put it another way, synthetic media such as deepfake “has [already] begun to invade legal proceedings”.<sup>48</sup> The first high profile case involving deepfake evidence arose in the context of custody proceedings in the UK at some point in 2019 (under UK law, such proceedings occur behind closed doors).<sup>49</sup> An audio recording submitted to the court on behalf of the mother appeared to show her husband using threatening comments which, on their face, served to demonstrate the father’s incapacity to look after their children moving forward. As it transpired, the recording was fabricated using deepfake technology trained to falsify voices. While metadata analysis provided on behalf of the father enabled the court to identify and dismiss the fabricated evidence, the case sends a clear warning of the potential for deepfake seriously to undermine the legal process.

The second high profile case arose in the U.S. in 2021 and serves to demonstrate the damage that can be done not only by the existence of deepfake but also by its mere allegation.<sup>50</sup> Raffaella Spone, a resident of Pennsylvania, was arrested and charged for allegedly creating and distributing deepfakes to undermine rivals of her cheerleader daughter. The claim included allegations that Spone created fake videos of the teenage rivals vaping and altered their social media accounts to make it appear as if they had been acting irresponsibly by drinking and smoking. Spone proceeded to instruct a firm of digital forensics experts who in her defence determined on the balance of probabilities that the material was authentic rather than deepfake. The prosecution subsequently dropped their case, but not before Spone’s reputation had been destroyed.

The threats posed by deepfake demand a pressing and concerted response to uphold the right to a fair trial and protect the integrity of the administration of justice.

As Delfino and others have noted, the evidentiary procedures and rules of evidence currently in operation in the U.S. and UK were constructed prior to the development of deepfake

---

<sup>48</sup> Rebecca Delfino, “Deepfakes on Trial: A Call To Expand the Trial Judge’s Gatekeeping Role To Protect Legal Proceedings from Technological Fakery”, 74(2) *Hastings Law Journal* 293 (2023).

<sup>49</sup> Gabriella Swerling, “Doctored audio evidence used to damn father in custody battle”, *The Telegraph*, 31 January 2020, <https://www.telegraph.co.uk/news/2020/01/31/deepfake-audio-used-custody-battle-lawyer-reveals-doctored-evidence/>.

<sup>50</sup> Christina Morales, “Pennsylvania Woman Accused of Using Deepfake Technology to Harass Cheerleaders”, *The New York Times*, 14 March 2021, <https://www.nytimes.com/2021/03/14/us/raffaella-spone-victory-vipers-deepfake.html>.

technology and, as such, are incapable alone of resolving the challenges posed by the presentation of fabricated evidence in court.<sup>51</sup> That is not to say that the territory is completely uncharted. Since their inception, the courts have been required to screen evidence submitted by counsel before allowing it to be relied upon, with the onus falling on the proponent to demonstrate that the evidence is real (i.e., is an authentic representation of what it ostensibly shows). That said, there is a crucial difference between traditional evidence and evidence generated, or allegedly generated, by deepfake: while the former can efficiently and dependably be appraised by natural human intelligence, the latter is rapidly becoming indistinguishable, at least to the human eye, from reality. The consequence is that the existing framework for evidence regulation is not equipped to deal with deepfake and other forms of synthetic media, requiring the introduction of additional, tailored rules and procedures.

Delfino proposes a two-pronged approach: first, investing in self-authenticating technology specifically developed to distinguish between real and deepfake-generated audiovisuals while in the meantime utilising existing technology currently being used vis-à-vis scientific and other audiovisual evidence such as witness authentication and digital forensic evidence; and second, placing the responsibility for authenticating evidence solely with judges (rather than with juries), who the author suggests are more capable than lay persons at resisting cognitive fallacies and have the capacity to develop expertise over time.<sup>52</sup>

However, the efficacy of this approach is uncertain: in particular, the capability of authenticating technology reliably to identify deepfake-generated evidence. As Apolo (2024) explains, such technology is expensive and time-consuming—a specific problem for the criminal justice system, which is strapped for cash and currently faced with a significant case backlog—and in any case is persistently outpaced by the deepfake technologies it purports to regulate.<sup>53</sup> The consequence, at least from the author’s perspective, is the need for an alternative approach which does not rely on human actors to undertake the increasingly difficult task of distinguishing between real and deepfake content.

---

<sup>51</sup> *Ibid* (no 55).

<sup>52</sup> *Ibid*.

<sup>53</sup> Yvonne Apolo, “Beyond a Reasonable Doubt? Audiovisual Evidence, AI Manipulation, Deepfakes and the Law”, 5 *IEEE Transactions on Technology and Society* (2024).

One such approach is the introduction of laws criminalising the use of harmful content in the courtroom, placing the onus on lawyers to ensure the authenticity of the evidence they submit on behalf of their client. (See e.g., the Deepfakes Accountability Act, a federal bill proposed in the U.S. in 2019, which would have, *inter alia*, proscribed the circulation of deepfakes without a digital watermark and/or a prominent statement disclosing the extent of the moderation.) In addition, Ahmad et al. (2020),<sup>54</sup> Alruwaili (2021),<sup>55</sup> and others endorse the use of blockchain technology, analogous to the regulation of DNA evidence, to ensure a ‘chain of custody’ (i.e., a chronological record documenting the custody) to determine whether a given piece of evidence has been fabricated.

### **Extrinsic Change: do State Courts become increasingly redundant?**

As human society becomes more and more accustomed to the use of AI and machine technologies in everyday life, from shopping to driving to automation of application processes, there will be an inevitable increase in confidence and a reduction in anxiety when it comes to AI. As swifter, cheaper processes become the expected standard of performance from private and public entities, what happens to a court system that remains AI-free? Is there a real risk that, when it comes to dispute resolution, the first choice of claimants will be to use private dispute resolution processes that will very often involve automated decision-making. Could this mean that, when it comes to private contractual, tortious and other types of disputes, such cases simply disappear from our state court systems?

Before those responsible for administering justice heave a collective sigh of relief, several important things must be considered. Firstly, if there is a huge drop off in private cases, what does that mean for the wider court system? Public cases will continue to engage state court systems, from criminal public prosecutions through to judicial review of government and its agencies. Does it mean that a swathe of judges who had been appointed to deal with civil cases have to be re-deployed and re-trained? Or does it mean that more and more judges with this expertise will leave the Bench and, as happens more often these days, become private

---

<sup>54</sup> Liza Ahmad, Salam Khanji, Farkhund Iqbal, and Faouzi Kamoun, “Blockchain-based chain of custody: towards real-time tamper-proof evidence management”, *Association for Computing Machinery* No 48 (2020).

<sup>55</sup> Fahad Alruwaili, “CustodyBlock: A Distributed Chain of Custody Evidence Framework”, *12(2) Information* 88 (2021).

dispute adjudicators or arbitrators in cases that do not involve automatic decision-making processes?

What, too, of the future of caselaw? The benefit of having private contractual and other civil disputes being litigated in our public courts is that the law can be developed by our judges in a public way, with transcripts of judgments at first instance in the High Court and then at appellate level being available to lawyers and the public via the National Archive. Without a flow of cases, how is the law to develop and evolve? Waiting for Parliament to reform the law via statute can be a frustratingly long experience. Primary legislation, without the sort of Henry VIII amending provisions that cause legislators huge concern about accrual of Executive power, is not the most flexible of instruments. Is a wider question as to the adequacy of Parliamentary process in the age of AI now being begged? This issue is, sadly, not for me to fully address in this paper, but the question of how our law is to evolve is a live one that must be considered sooner rather than later.

**Extrinsic Change: does the increasing use of automated systems by Governments pose a challenge to justice and accountability?**

As Governments increasingly move to the deployment of automated systems to replace human decision-makers for example when it comes to the award of benefits or the issue of required documents for driving, then the question of accountability and explicability of decisions starts to loom large. In any Judicial Review application in the UK, the Government has a Duty of Candour, requiring it to disclose all relevant material surrounding the particular decision that is being challenged.

This duty allows applicants to fully understand and challenge the rationale used for a particular decision, applying tests of lawfulness that have been set down in caselaw and statute. Without the deployment of technology that explains a decision, then the machine will present itself as an inscrutable black box, with nothing that can be disclosed or explained. Whilst the need for speed and certainty of decision-making is entirely understandable in these times of constrained public spending, the need for an explanation remains.

**Extrinsic Change: enforcement of judgments: can AI do all the heavy lifting, or will injustice ensue?**

There is an understandable argument that, once the judicial or court process has been completed, then the next stage of enforcement of judgments or orders could and should be entirely handed over to automated processes, which is the increasing norm in many other parts of our lives. The current situation in England and Wales is far from satisfactory. In many civil and family cases, enforcement can prove to be a difficult challenge, as parties seek to evade their liabilities or responsibilities either by withholding payment in financial cases or by not complying with family court orders relating to children, for example.

The Civil Justice Council of England and Wales has this month produced a Report on Civil Enforcement <https://www.judiciary.uk/wp-content/uploads/2025/04/CJC-Report-on-Civil-Enforcement-April-2025.pdf> which recommends a new unified digital enforcement court in civil claims, with a portal that can track the financial position of defendants in a way that will eliminate or reduce the need for separate investigations and which would still offer protection to defendants facing genuine difficulties in satisfying judgments. Importantly, the Report does not advocate taking the enforcement process away from the courts and allowing it to become a purely administrative one. This is because, rightly, the view is taken that enforcement must be part of a continuing course of justice, to ensure that unfairness and abuse does not occur.

Whilst this Report is a good start, much more work will need to be done to firstly create databases about the financial status of defendants that are reliable and accessible to the relevant agencies. If as I suspect the enforcement of judgements will remain a process to be controlled ultimately by the Courts, then the deployment of AI has to be done in a way that enhances, rather than diminishes, the reputation of the system.

## **Part II: Mitigating Measures**

### **A. Requirements for models to be able to detect and mitigate bias**

The stakes in legal decision-making are exceptionally high, with potential life-altering consequences for individuals and far-reaching implications for society. Therefore, the strategy must be both rigorous and adaptable, capable of evolving alongside rapid advancements in AI. I propose the implementation of a protocol for AI models in the justice system. This protocol would operate on four interconnected levels: Data Scrutiny, Model Transparency, Continuous Monitoring, and Societal Impact Assessment.

The first layer, **Data Scrutiny**, involves a deep dive into the training data used to develop AI models. This process goes beyond simple demographic checks to include a comprehensive analysis of the historical and societal context of the data. For instance, when examining historical case data, we must consider not just the raw outcomes, but the societal norms and potential systemic biases prevalent at the time these cases were decided. This might involve collaborating with social scientists and legal historians to provide context for the data. Consider, for example, the landmark case of *Brown v. Board of Education*<sup>56</sup> in the United States. If we were to train an AI model on pre-1954 education law cases without proper context, it might perpetuate the "separate but equal" doctrine that was prevalent at the time. This underscores the need for a nuanced understanding of historical context in our data scrutiny process.

To illustrate this point, let's examine a more recent example. The England and Wales case of *R (on the application of Begum) v Headteacher and Governors of Denbigh High School*<sup>57</sup> dealt with the complex issue of religious dress in schools. An AI system trained on historical UK cases relating to religious freedom might struggle to appropriately weigh evolving societal attitudes towards multiculturalism and religious expression that this case reflected. This highlights the need for Data Ethics Boards to include not just legal experts and data scientists, but also sociologists and other experts who can provide context to culturally sensitive cases.

The creation of Data Ethics Boards for each jurisdiction implementing AI in their justice system is a major step, but their composition and mandate would require careful consideration. These boards must be representative of the communities they serve and should have the authority not just to audit existing data, but to proactively shape the data collection process. For example, they might mandate the collection of more granular data on factors like socioeconomic status, education level, or neighbourhood characteristics in criminal cases. This could help identify and address systemic biases that might be masked by more surface-level demographic data.

---

<sup>56</sup> 347 U.S. 483 (1954).

<sup>57</sup> [2006] UKHL 15.

A Data Ethics Board may, for example, also be empowered to commission targeted research to fill gaps in the existing data. If it is found that there is insufficient data on the outcomes of cases involving defendants with mental health issues, the board could initiate a comprehensive study on this topic. This approach ensures that the models are trained on data that is not just historically accurate, but also relevant to current societal needs and challenges. Furthermore, the boards may seek to establish guidelines for data provenance and quality. Each piece of data used to train the AI would be traceable to its source, with documentation of any preprocessing or anonymization steps taken. This level of transparency is important for maintaining public trust and allowing for effective audits.

However, we must also be mindful of the potential pitfalls of this approach. There's a risk that in our effort to correct for historical biases, we might inadvertently introduce new ones. For instance, if we overly emphasise certain factors to try to correct perceived injustices, we might create an AI system that is biased in the opposite direction. Moreover, the creation of an ethics board raises important questions about governance and accountability. Who appoints the board members? How long do they serve? How are disagreements within the board resolved? These are questions that need to be addressed to ensure the legitimacy and effectiveness of such bodies.

The second layer, **Model Transparency**, addresses the 'black box' problem often associated with complex AI systems. Here, we must go beyond simple interpretability features to develop what I call 'Legal AI Explainers' (LAIEs). These would be specialised systems designed to translate the decision-making processes of the primary AI models into clear, legally relevant explanations. The LAIEs would not only provide a step-by-step breakdown of how a decision was reached but would also link each step to relevant legal principles and precedents.

Reminding ourselves of the *Compas* case and the dangers of using opaque systems (*State v. Loomis*<sup>58</sup>) in Wisconsin, referred to above and in my previous paper, we can envision LAIEs that go beyond simple risk scores or binary outputs and which deal with opacity. For example, in a sentencing context, an LAIE might provide a detailed breakdown of the factors

---

<sup>58</sup> 881 N.W.2d 749 (Wis. 2016).

considered, their relative weights, and how they interact to produce a final recommendation. This explanation would be anchored in established legal principles and relevant case law.

Let's look at a hypothetical case where an AI system recommends a particular sentence for a drug possession offence, for example, in English criminal sentencing law when it comes to a defendant who has pleaded Guilty to possession of a Class A drug with intent to supply.

The LAIE might explain: *"The recommendation of a 24-month suspended sentence is based on the following factors: (1) The quantity of drugs involved (5 grams of cocaine) which, according to the Sentencing Council Guidelines for Drugs Offences, is a Category 4 level of seriousness and with a lesser role means a starting point of 18 months imprisonment with a range from a community sentence to 3 years imprisonment. (2) There are no aggravating factors and when it comes to mitigating factors, the defendant's lack of previous convictions, which suggests a lower likelihood of reoffending. (3) The defendant's demonstrated steps towards rehabilitation in his pre-sentence report are also relevant. (4) He pleaded guilty at the earliest opportunity and is therefore entitled to a reduction in his sentence of one third. (5) The starting point for this case is therefore one of 15 months imprisonment, meaning that a sentence of 10 months, suspended for two years, is imposed."*

This level of detail not only provides transparency but also allows for meaningful challenge and debate. It enables judges, lawyers, and defendants to understand and potentially contest the AI's reasoning, ensuring that the human element of judicial discretion is maintained.

To implement this process, we would need to establish a new field of study combining law, computer science, and linguistics. This interdisciplinary approach would focus on developing AI systems capable of generating explanations that are not only technically accurate but also legally sound and comprehensible to both legal professionals and laypersons.

One promising avenue for developing LAIEs is the use of "attention mechanisms" in neural networks, a technique that has shown success in creating more interpretable AI models. For example, researchers at MIT have developed a system that can highlight which parts of an input (such as specific words in a legal document) were most important in reaching a decision.

Adapting this technology to legal contexts could allow LAIEs to visually demonstrate how different parts of a case file influenced the AI's decision. Another potential approach is the use of "counterfactual explanations", a concept explored by researchers at the Alan Turing Institute. In a legal context, this might involve the LAIE providing alternative scenarios: "If the defendant had no prior convictions, the recommended sentence would have been X. If the value of stolen goods had exceeded £10,000, the recommendation would have been Y." This approach not only explains the actual decision but also helps to delineate the boundaries of the AI's decision-making process.

However, the development of LAIEs is not without challenges. One significant hurdle is the potential for these explanations to be used to "game" the system. If the exact weighting and interaction of factors are made explicit, it could lead to attempts to manipulate inputs to achieve desired outcomes. This concern has been raised in other contexts where AI decision-making processes have been made transparent, such as in credit scoring systems. To address this, LAIEs might need to incorporate a degree of randomness or variability in their explanations, like the concept of "differential privacy" used in data protection. This would involve providing explanations that are accurate in aggregate but contain small, random variations in individual cases, making it difficult to reverse-engineer the exact decision-making process.

Another challenge lies in ensuring that LAIEs can keep pace with the evolving nature of law. Legal principles and precedents are constantly being refined and reinterpreted. For example, the case of *R v Jogee*,<sup>59</sup> demonstrates how important rules of evidence such as joint enterprise can evolve. LAIEs would need to be designed with the flexibility to incorporate such changes rapidly and accurately.

The third layer, **Continuous Monitoring**, involves the implementation of a real-time bias detection system. This system would go beyond periodic checks to provide vigilance against emerging biases. To achieve this, **Bias Detection Neural Networks** (BDNNs) would be designed to operate alongside the primary AI system, analysing its outputs for potential biases.

---

<sup>59</sup> [2016] UKSC 8.

The architecture of these networks would be inspired by recent advancements in anomaly detection and adversarial machine learning. For example, we might implement a system like the one described by Zhang et al. in their 2018 paper "Mitigating Unwanted Biases with Adversarial Learning".<sup>60</sup> Their approach uses adversarial debiasing techniques to remove unwanted information from a model's internal representations. In our context, we could adapt this method to create BDNNs that are trained to identify patterns of bias across various demographic factors in real-time.

Consider a practical example: In a system that processes bail applications, the BDNN might detect that over a period of weeks, the primary AI has begun recommending slightly higher bail amounts for defendants from a particular postcode. This bias might be subtle enough to escape notice in periodic reviews, but the BDNN, trained on principles of counterfactual fairness, would flag this emerging pattern immediately. The use of counterfactual fairness techniques is particularly relevant in legal contexts. As defined by Kusner et al. in their 2017 paper "Counterfactual Fairness", this approach ensures that a decision remains the same in a counterfactual world where an individual's protected attributes are changed.<sup>61</sup> In our bail application example, the BDNN would check whether the AI's decision would remain consistent if the defendant's postcode (which might be a proxy for race or socioeconomic status) were different. Moreover, the continuous monitoring layer would incorporate dynamic data re-weighting mechanisms. This approach is inspired by works like that of Jiang and Nachum on "*Identifying and Correcting Label Bias in Machine Learning*".<sup>62</sup> As biases are detected, the system would automatically adjust the importance of different data points in real-time, ensuring that the model's decision-making process remains as fair and unbiased as possible.

The implementation of such a system is not without challenges. One significant hurdle is the need for these adjustments to be made rapidly and in a context-aware manner. This requires the development of sophisticated algorithms capable of understanding the nuanced legal context of each decision. Moreover, we must also be cautious about potential unintended

---

<sup>60</sup> Brian Zhang, Blake Lemoine, and Margaret Mitchell, "Mitigating Unwanted Biases with Adversarial Learning" (2018). Available at: <https://dl.acm.org/doi/10.1145/3278721.3278779>.

<sup>61</sup> Matt Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva, "Counterfactual Fairness" (2017). Available at: <https://dl.acm.org/doi/10.5555/3294996.3295162>.

<sup>62</sup> Heinrich Jiang and Ofir Nachum, "Identifying and Correcting Label Bias in Machine Learning" (2020). Available at: <https://proceedings.mlr.press/v108/jiang20a.html>.

consequences of such dynamic adjustments. There's a risk that rapid changes in the decision-making process could lead to inconsistencies in legal outcomes over short periods. To mitigate this, changes would need to be made incrementally, and their impacts closely monitored. Furthermore, the continuous monitoring system must be designed with transparency in mind. Each adjustment made by the system should be logged and explainable, allowing for human oversight and ensuring that the pursuit of fairness doesn't inadvertently introduce new forms of bias or unfairness.

The implementation of such a sophisticated continuous monitoring system represents a significant technical challenge, but it's one that's crucial to address if we're to ensure the ongoing fairness and integrity of AI-driven legal decision-making.

The fourth layer, **Societal Impact Assessment**, acknowledges that bias in the justice system doesn't exist in a vacuum but has real-world consequences that can perpetuate or exacerbate societal inequalities. I propose the establishment of an **AI Justice Impact Unit (AIJIU)** maybe within a relevant government department or as an independent oversight body. This unit would be responsible for conducting regular, comprehensive assessments of how AI-driven legal decisions are affecting different communities and demographic groups.

To demonstrate the potential impact of AIJIUs, let us consider a hypothetical scenario based on real-world concerns. Imagine an AI system implemented to assist in bail decisions across a large metropolitan area. After six months of operation, initial data suggests that the system has led to more consistent bail decisions and a reduction in overall pre-trial detention rates. However, a comprehensive review by the AIJIU reveals a more complex picture. Through statistical analysis, the AIJIU might discover that while overall pre-trial remand in custody rates have decreased, they have disproportionately decreased for defendants from affluent areas, while remaining relatively stable for those from economically disadvantaged areas.

This disparity might not be immediately apparent from aggregate data alone, highlighting the importance of granular, demographic-specific analysis. Qualitative research conducted, including interviews with defendants, lawyers, and community leaders, could uncover the reasons behind this disparity. For example, the AI system might be placing significant weight on factors such as stable employment and fixed address, which could inadvertently disadvantage individuals from communities with higher rates of informal employment or

housing insecurity. Community engagement efforts might also reveal further nuances. For example, cultural differences in communication styles could be influencing how the AI interprets statements made by defendants from certain ethnic backgrounds, potentially leading to biased risk assessments. This echoes real-world concerns raised about AI language models, such as GPT-3, which have been shown to exhibit biases against non-standard English dialects.

The AIJIUs would use a combination of statistical analysis, qualitative research, and community engagement to build a holistic picture of the AI system's impact. They would have the power to recommend systemic changes, not just to the AI models themselves, but to the broader application of AI in the justice system. For example, they might recommend the development of community-based support programs to help defendants meet the criteria that the AI system considers indicative of lower flight risk, thus addressing the root causes of the disparity rather than merely adjusting the algorithm.

-

Bias, to some degree, is inherent in any human-designed system, including our current justice frameworks. The goal, therefore, is not to achieve perfect neutrality - a likely impossible standard - but to systematically reduce bias to a level that society deems acceptable and that represents a significant improvement over current human-driven process.

One approach to establishing an acceptable threshold of bias is to look at existing standards in other fields. For example, the US Equal Employment Opportunity Commission uses a "four-fifths rule" in employment decisions. This rule states that if the selection rate for a protected group is less than 80% of the rate for the group with the highest selection rate, this may be regarded as evidence of adverse impact. While this provides a clear numerical threshold, it's important to note that it has been criticised for being overly simplistic and potentially allowing for significant disparities to persist.

The UK Equality Act 2010 introduces the concept of "*proportionate means of achieving a legitimate aim*" when considering indirect discrimination. This more flexible approach allows for some level of disparate impact if it can be justified as necessary and proportionate. Translating this into the realm of AI-driven legal decisions, we might consider a system acceptable if it can demonstrate that any disparities in outcomes are proportionate to legitimate legal factors and not based on protected characteristics. However, both of these

approaches, while providing useful reference points, may not be sufficient for the high-stakes environment of the justice system.

A system that is transparently and continuously improving, with clear mechanisms for identifying and addressing biases, is preferable to my mind to one that claims perfect neutrality but whose inner workings are opaque and unexamined.

In conclusion, the protocol provides a potential framework for detecting and mitigating bias in AI models used in the justice system. However, it is important to recognise that this is not a one-time implementation but an ongoing process that requires constant vigilance, adaptation, and societal engagement. As we continue to integrate AI into our justice systems, we must remain committed to the principles of fairness and equality under the law, using technology as a tool to enhance, rather than undermine, these vital tenets of justice.

## **CONCLUSION**

In summary, the rapid rise of AI and machine learning means that justice systems across the world are already adapting. Thus far, the use of AI has largely been limited to administrative and research-oriented tasks, but some jurisdictions are deploying automated decision-making to deal with a range of mainly civil disputes. The prospects of a fully international set of rules to govern the deployment of automated technologies remain remote, and in many countries, there is no domestic legislation or systematic regulation of the use of AI in the various realms of human activity.

Rather than adopting a purely reactive approach to both the intrinsic and extrinsic use of AI, I believe that justice systems should develop their own set of ethical and professional standards, with mitigating, monitoring, and evaluation functions being carried out jointly by the judiciary, the legal profession, and technologists. In this paper, I have focused on some potential changes in my jurisdiction of England and Wales, but whether this is in the form of one joint body or several bodies dedicated to the oversight, assessment, and monitoring of the use of AI in the justice system is of less significance than the principle of “do no harm”, which should underpin such supervisory activities.

There is no doubt whatsoever that the rise of AI is going to change the concept of justice itself. I believe that millions more people should and will have access to legal information and assistance as the “justice gap” is closed by automated and well-designed systems. If the approach to automation is, however, one of “devolve and forget”, then injustices and miscarriages of justice will inevitably occur, with resultant pressure on policymakers to take corrective action.

A legal and judicial sector that takes up the mantle of leadership on the use of AI and its integration into justice processes, realising its own knowledge gaps and limitations, should work with technological experts to maintain a healthy balance between efficiency and justice. In this paper, I have sought to identify some means of doing just that.

### **ACKNOWLEDGEMENTS**

I am indebted to the following people for their direct input and continuing advice in the compilation of this paper: my Research Assistant, Angus Taylor, an LLM Student at Harvard Law School, Michael Bryan, formerly of Harvard Medical School and now a DPhil Student at the Centre for Immuno-Oncology, Oxford University, Dr Carlo Ross, Master in Public Health, Harvard University, Amanda Chaboryk, head of Legal Data and Systems at PWC and Will Siebert, my former Parliamentary Research Assistant and now a Masters student at King’s College London.

### **BIBLIOGRAPHY**

Ahmad, L., Khanji, S., Iqbal, F., and Kamoun, F., “Blockchain-based chain of custody: towards real-time tamper-proof evidence management”, *Association for Computing Machinery* No 48 (2020).

Alruwaili, F., “CustodyBlock: A Distributed Chain of Custody Evidence Framework”, 12(2) *Information* 88 (2021).

Apolo, Y., “Beyond a Reasonable Doubt? Audiovisual Evidence, AI Manipulation, Deepfakes and the Law”, 5 *IEEE Transactions on Technology and Society* (2024).

Bloomberg Law, “2023 State of Practice: Practice in the New Era”. Available at: <https://pro.bloomberglaw.com/insights/technology/2023-state-of-practice-practice-in-the-new-era/>.

Brader, C., “Reducing the Crown Court backlog”, House of Lords Library, 13 March 2025. Available at: <https://lordslibrary.parliament.uk/reducing-the-crown-court-backlog/#:~:text=The%20crown%20court%20backlog%20reached,their%20cases%20to%20be%20resolved.>

*Brown v. Board of Education of Topeka* 347 U.S. 483 (1954).

Buckland, R., “AI, JUDGES AND JUDGEMENT: SETTING THE SCENE”, M-RCBG Working Paper No. 220. Available at: <https://www.hks.harvard.edu/centers/mrcbg/publications/awp/awp220>.

Calo, R., “Artificial Intelligence Policy: A Primer and Roadmap”, 51 UC Davis Law Review 399 (2017).

Clement-Jones, T. (2024) *Living with the Algorithm: Servant or Master?: AI Governance and Policy for the Future*. Oxford: Unicorn Publishing Group.

Courts and Tribunals Judiciary, “Artificial Intelligence (AI): Guidance for Individual Office Holders”, 12 December 2023. Available at: <https://www.judiciary.uk/wp-content/uploads/2023/12/AI-Judicial-Guidance.pdf>.

Delfino, R., “Deepfakes on Trial: A Call To Expand the Trial Judge’s Gatekeeping Role To Protect Legal Proceedings from Technological Fakery”, 74(2) *Hastings Law Journal* 293 (2023).

Eubanks, V. (2018) *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St Martin’s Press.

Fisher, J., “Independent Review of Disclosure and Fraud Offences”, 20 March 2025. Available at: <https://www.gov.uk/government/publications/independent-review-of-disclosure-and-fraud-offences>.

Floridi, L., “What the Near Future of Artificial Intelligence Could Be”, *Philosophy & Technology* (2020).

Gibney, E., “AI models fed AI-generated data quickly spew nonsense”, 632 *Nature* 18 (2024).

Hartzog, W. and Selinger, E., “Surveillance as a Loss of Obscurity”, 72 Wash. & Lee L. Rev. 1343 (2015).

Hymas, C., “Secret court hearings are unfair and in need of reform, says former chief justice”, *The Telegraph*, 19 August 2024. Available at: <https://www.telegraph.co.uk/news/2024/08/19/secret-court-hearings-unfair-need-reform-lord-thomas/>.

Jiang, H. and Nachum, O., “Identifying and Correcting Label Bias in Machine Learning” (2020). Available at: <https://proceedings.mlr.press/v108/jiang20a.html>.

JUSTICE, “AI in our justice system”, January 2025. Available at: <https://justice.org.uk/ai-in-our-justice-system/>.

Kirk, T., “Pensioner, 93, with dementia convicted in fast-track SJP courts for not having car insurance”, *Evening Standard*, 7 March 2025. Available at: <https://www.standard.co.uk/news/crime/single-justice-procedure-dvla-pensioner-dementia-care-conviction-car-insurance-b1215124.html>.

Kusner, M., Loftus, J., Russell, C., and Silva, R., “Counterfactual Fairness” (2017). Available at: <https://dl.acm.org/doi/10.5555/3294996.3295162>.

Magesh et al. “Hallucination Free? Assessing the Reliability of Leading AI Research Tools”, *Empirical Legal Studies*, (forthcoming 2024).

Mittelstadt, B., et al., “The ethics of algorithms: Mapping the debate”, 3(2) *Big Data & Society* (2016).

Morales, C., “Pennsylvania Woman Accused of Using Deepfake Technology to Harass Cheerleaders”, *The New York Times*, 14 March 2021. Available at: <https://www.nytimes.com/2021/03/14/us/raffaella-spone-victory-vipers-deepfake.html>.

Pasquale, F. (2015) *The Black Box Society: The Secret Algorithms that Control Money and Information* Cambridge, MA: Harvard University Press.

Post Office (Horizon System) Offences Act 2024.

Post Office Horizon IT Inquiry. Available at: <https://www.postofficehorizoninquiry.org.uk/>.

*R (on the application of Begum) v Headteacher and Governors of Denbigh High School* [2006] UKHL 15.

*R (on the application of Bridges) v South Wales Police* [2020] EWCA Civ 1058.

*R v Jogee* [2016] UKSC 8.

*R v L and Others* [2013] EWCA Crim 991.

Sara Merken, “New York lawyers sanctioned for using fake ChatGPT cases in legal briefs”, *Reuters*, June 26 2023. Available at: <https://www.reuters.com/legal/new-york-lawyers-sanctioned-using-fake-chatgpt-cases-legal-brief-2023-06-22/>.

*State v. Loomis* 881 N.W.2d 749 (Wis. 2016).

Sweeney, M., “What is the Post Office Horizon IT scandal all about?”, *The Guardian*, 7 January 2024. Available at: <https://www.theguardian.com/business/2024/jan/07/what-is-the-post-office-horizon-it-scandal-all-about>.

Swerling, G., “Doctored audio evidence used to damn father in custody battle”, *The Telegraph*, 31 January 2020. Available at: <https://www.telegraph.co.uk/news/2020/01/31/deepfake-audio-used-custody-battle-lawyer-reveals-doctored-evidence/>.

The Law Society, “Generative AI: the essentials”, 7 August 2024. Available at: <https://www.lawsociety.org.uk/topics/ai-and-lawtech/generative-ai-the-essentials>.

UK Government, “Fact sheet: Single Justice Service”.

UK Government, “Single justice procedure notices”. Available at: <https://www.gov.uk/single-justice-procedure-notice>.

Wirtz, B., Weyerer, J., and Geyer, C. “Artificial Intelligence and the Public Sector – Applications and Challenges”, 42(7) *International Journal of Public Administration* 596 (2019).

Zhang, B., Lemoine, B., and Mitchell, M. “Mitigating Unwanted Biases with Adversarial Learning” (2018). Available at: <https://dl.acm.org/doi/10.1145/3278721.3278779>.

Zuboff, S. (2019) *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. New York: PublicAffairs.

## **ANNEX 1**

### **AI IN THE CRIMINAL PROCESS: AN URGENT OPPORTUNITY**

*Authors:*

Rt Hon Sir Robert Buckland KBE KC

William Rees, Palantir Technologies UK

## **1. Summary**

1. The adoption of AI must be considered a critical component of any recommendations for a more efficient criminal court system and improved timeliness for victims, witnesses and defendants. As a technology its revolutionary benefits are being proven daily. Across the UK it is being employed to deliver better, faster, and smarter care in hospitals; in schools to cut down the 15+ hours a week they spend on lesson planning and marking in pilots; and throughout professional services to reduce the time drafting structured reports and forms by 20-80%<sup>63</sup>. It is no longer a novel innovation but a baseline standard for efficacy and we must urgently embrace the opportunities it offers in criminal justice, not least because it can provide direct efficiency gains immediately and without requiring either primary legislation or more sensitive changes to legal processes and rights.
  
2. In this submission we will provide an overview of the current capabilities of AI before moving on to discuss specific opportunities for their deployment to improve efficiency in the criminal courts. This will include looking at areas including:
  - Compliance tracking, enforcement and improvement;
  - Reducing the judicial burden of administrative work;
  - Proactive case progression;
  - Case retrieval, analysis and drafting;
  - Indictment drafting;
  - Consistency in charging and review decisions; and
  - Improving allocations.
  
3. We will then proceed to discuss the principles and best practices that must be applied to any application of AI in the criminal courts and explain how AI can be deployed successfully and swiftly through the creation of a dedicated AI office.

4. Further information and walkthroughs of any of the workflows discussed herein, including those described as sensitive, can be provided on request, subject in some cases to restrictions on publication. Please contact us for more information.

### **1.1. Authors' Contributions**

5. Rt Hon Sir Robert Buckland KBE KC is former Lord Chancellor and Secretary of State for Justice, former Secretary of State for Wales, Minister of State for Justice and Solicitor General of England and Wales. He is currently a Senior Fellow at the Mossavar-Rahmani Center for Business and Government at Harvard Kennedy school. He is a member of Foundry Chambers, Senior Counsel at Payne Hicks Beach LLP and is a member of the Policy Unit at DAC Beachcroft. He is not affiliated with or employed by Palantir Technologies and is contributing to this submission in an individual capacity as an interested party.
6. William Rees is a Deployment Strategist at Palantir Technologies UK ("Palantir") and former barrister. He is contributing to this submission on behalf of Palantir an industry-leader in the deployment of AI and data technologies both in the government and the commercial settings. Please see more information about Palantir at [palantir.com/uk/](https://palantir.com/uk/)

## Contents

AI in the Criminal Process: An Urgent Opportunity .....	<b>Error! Bookmark not defined.</b>
1. Summary .....	48
1.1. Authors' Contributions .....	49
2. Capabilities and Opportunities.....	51
2.1. Capabilities .....	51
2.2. Opportunities.....	52
3. Deployment, Timelines and Adoption.....	58
3.1. Deployment.....	59
3.2. Timelines.....	61
3.3. Adoption .....	61
4. Ethics and Principles.....	62
4.1. Ethics.....	62
4.2. Principles.....	63
5. AI Office .....	65
6. Financial and Operational Considerations .....	67
7. Conclusion .....	67

## 2. Capabilities and Opportunities

7. In this section we will discuss the current state of AI, its capabilities and where they are applied in industry and government. We will then move on to discuss discrete opportunities for the deployment of AI within the criminal process to deliver improved efficiency and timeliness.

### 2.1. Capabilities

8. While AI has a long history, its adoption has increased exponentially since 2021, driven primarily by the release of a new generation of Generative AI models such as OpenAI's GPT or Claude by Anthropic. In simple terms these models can receive user input in the form of text, images, or audio, and generate increasingly sophisticated, realistic, accurate, and useful text. This means that AI systems can answer questions, respond to dialogue, and interface intuitively with humans and other systems, but at a vastly greater speed and scale. Indeed, the most recent models can ingest and respond to up to 100 pages of text in seconds at a cost of less than a couple of pounds.
9. Moreover, as platforms such as Palantir's Artificial Intelligence Platform have shown, AI is not just constrained to interpretation and generation of text, but also planning and execution of actions. From writing and sending emails and text messages, to creating and scheduling work orders, AI agents have moved beyond chat to interact with the world around them.
10. As a result of these capabilities AI has reached a high level of public recognition and become a key strategic priority for industry and governments alike; over 300m people use OpenAI's ChatGPT<sup>64</sup>, an application which allows users to interact with a general purpose chatbot, on a weekly basis; approximately 60% of Fortune 500 companies referenced AI in their most recent earnings calls<sup>65</sup>; and the UK government has recently announced it is adopting more than 50 recommendations from its "AI Opportunities Action Plan" to turbocharge adoption of AI in the private and public sectors.

---

<sup>64</sup> <https://www.theverge.com/2024/12/4/24313097/chatgpt-300-million-weekly-users>

<sup>65</sup> <https://deloitte.wsj.com/cio/largest-u-s-companies-disclose-ai-as-material-risk-1ee3ba07>

11. The potential for AI is clear, and crucially leaders in industry and government are already proving the model for successful adoption. To provide some tangible examples of where AI is already delivering in the most sensitive public spheres, the authors have direct experience of AI deployments:
- 11.1. in hospitals to synthesize information and institutional knowledge from multiple sources to generate clinical documents, saving clinician time while also increasing reliability and standardisation;
  - 11.2. at government agencies to collect and interpret information generated across multiple sources in real-time, to ensure that activities align with priorities and that resources are used effectively, and to support improved, expedited decision-making and collaboration; and
  - 11.3. at government agencies to allow users to access and interact with organisation-specific context, streamline repetitive manual tasks, and provide secure access to large language models to support innovation. This tooling can be deployed via a chat interface for users to rapidly query and retrieve information, draft documents and coordinate actions at scale, such as the successful management of large-scale international events.
12. Due to their sensitive nature, information regarding these applications cannot be disclosed in a public forum but for more information and to see these applications, please contact the authors.

## **2.2.Opportunities**

13. Leveraging this experience of successful applications, it is possible to quickly identify several key areas in which AI can be quickly deployed to great benefit in the criminal process.
14. In focusing on these areas, we have been informed by some of the areas for improvement identified in the 2015 review of efficiency in criminal proceedings<sup>66</sup> (“the 2015 Review”), as well as direct experience of the justice system, and have prioritized those application

---

<sup>66</sup> <https://www.judiciary.uk/wp-content/uploads/2015/01/review-of-efficiency-in-criminal-proceedings-20151.pdf>

that would not require primary legislation and which could be deployed without confronting operational, procedural or legal constraints. However, as you will see the principles and capabilities set out can easily be generalised to a wide range of applications not detailed herein.

15. Finally, it is important to note that while we have sought to identify discrete areas in which to target the deployment of AI, the benefits derived should not be considered in isolation. Many of the use-cases complement each other and AI should be seen not as a collection of point solution but an ecosystem of agents, each able to interact with relevant data, tools, users and each other. As more workflows are implemented, we can expect the benefits to compound as coordination between users and AI agents increases.

### ***2.2.1. Compliance tracking, enforcement and improvement***

16. Compliance with the Criminal Procedure Rules (“CrimPR”) and judicial orders is crucial for the efficiency and effectiveness of the criminal justice system, and was an issue highlighted in the 2015 Review. The rules were designed to foster a culture of responsibility for case progression, but they are ineffective if not observed.
17. Indeed, ensuring compliance is challenging. A significant issue is the development of a 'culture of failure' within the courts, where there is an expectation that deadlines will not be met. This culture undermines the authority of the court and the rule of law.
18. Here AI can reduce the burden of identifying and recording non-compliance, provide proactive responses to encourage correction and compliance, and support the process of accountability while minimizing the impact on court resources.
19. An AI “Compliance Agent” can be configured and deployed to:
  - 19.1. continuously review documents uploaded to the Crown Court Digital Case System;
  - 19.2. compare against the CrimPR and orders to identify and record non-compliance and its impact;
  - 19.3. coordinate the response thereafter, raising issues for review by Case Progression Officers, Listing Officers and Judges as appropriate; and

- 19.4. support the process of remedying the breach and/or creating some sanction.
20. Most importantly, the agent and the compliance history of specific court users can also be utilised to increase accountability and compliance, not just reduce the burden of remedying breaches on the courts. There is a myriad of approaches to increasing compliance that could be pursued but to give some examples the agent may:
- 20.1. pro-actively facilitate compliance by interacting with parties to give notifications and relevant, specific guidance on compliance with relevant orders and the CrimPR;
- 20.2. helping oversight users to more efficiently and accurately identify the individuals, organisations or units most frequently in non-compliance and coordinating appropriate responses. These may include reminders of duties and guidance on improving compliance, attendance at compliance hearings, written notices requiring response or escalation to bodies; and
- 20.3. create a culture of compliance by generating periodic compliance feedback and assessments for court users that recognises good performance, identifies most commonly occurring issues, sets out best practice, and highlights the professionalism, importance and impact of good compliance.

### ***2.2.2. Reducing the judicial burden of administrative work***

21. While resources are constrained in multiple areas, there is a consensus that sitting time is presently the key constraint in the court system. After a reduction in the overall limit during the preceding decade, Sir Robert Buckland, in agreement with the then LCJ removed the annual limit on Crown Court sitting days at the beginning of the Covid crisis to try to reduce the court backlog. The pandemic proved to be a constraint on sitting days, but even now, it is apparent that there are some challenges, such as the supply of judges and barristers, to how much sitting time can be increased further and how quickly.
22. Consequently, increasing the efficiency of judges' time should not just be considered a matter of efficiency, but also an urgent priority. One such opportunity for doing this through AI, can be found in its capability to coordinate and accelerate the resolution of administrative matters such as paper applications, or case management orders.

23. There are numerous ways in which an AI “Administrative Assistant” could be deployed to reduce the amount of time judges spend on administrative work. The key functionality required would be:
- 23.1. to identify when the need for some order or direction arises, either in response to an application, at the initiation of the court or in anticipation of an issue;
  - 23.2. retrieve and synthesize the relevant information from the case, CrimPR, case law and legislation to inform the order or direction; and
  - 23.3. generate a draft order, utilising these materials as well as any precedents, templates or forms that may be appropriate together with a clear explanation of its reasoning.
24. Thereafter the agent will be able to collaborate with the judge and parties as appropriate, whether to encourage early resolution through agreement, support determination on the papers, or facilitate the swift determination through hearing. In each instance, supporting edits and iterations of the draft order and learning from the exercise for future deployments.
25. We would anticipate that such support may not only make the process of completing administrative work less time-consuming, but in many cases, by seeking to anticipate issues and then facilitate agreement, could avoid burdening court or judge time altogether.

### ***2.2.3. Proactive case progression***

26. The 2015 Review identified the need and benefits of early engagement between the parties, a pro-active approach to case progression and dedicated Case Progression Officers who can ensure that all cases proceed to completion without delay and ensure that all parties comply with the relevant rules.
27. As we can generalise from the compliance workflow above, there are numerous opportunities for an AI “Case Progression Assistant” to support the most effective vision of these measures through continuous review of cases, identification of issues at the earliest possible opportunity, and increased coordination between all relevant parties.

28. This may encompass not just flagging procedural issues at the earliest stage but also ensuring that all practical measures are in place to ensure that matters proceed swiftly. For example, an AI “Case Progression Assistant”, may generate and automate outreach to confirm attendance of required persons, or facilitate engagement between parties to resolve scheduling issues.
  
29. As with all our use-cases it is important to stress that the AI Case Progression Assistant is not intended to replace the role of dedicated Case Progression Officers, but rather to allow them to scale up their impact. The AI agents will act not as a replacement but a force multiplier.

#### ***2.2.4. Indictment drafting***

30. The current system of preparing indictments in criminal proceedings was identified as a significant burden on court time in the 2015 Review but remains a requirement of the Crown Court process to date.
  
31. To summarise the nature of the issue, when a case reaches the Crown Court, the original charge or summons is withdrawn and replaced by an indictment. The process of drafting and preferring an indictment is governed by specific legal requirements and formalities, which are both time-consuming and administratively burdensome and so the preparation of thousands of indictments each year imposes a significant administrative load on both the prosecution and the courts.
  
32. However, AI agents can be deployed to automatically draft indictments based on the case documents precedents and law, coordinate reviews and edits from all parties, perform compliance checks and reviews of drafts, and manage the process of acquiring the signature.
  
33. This is a simple workflow that could be triggered on transfer of a case to the Crown Court and would not only reduce the administrative burden of indictments and accelerate the

progression of cases through the courts but also improve the consistency of indictments by applying the same principles and approaches across all indictments.

34. Indeed, even if the requirement for drafting an indictment on transfer to the Crown Court is removed by future reforms, the early creation, update and coordination of more consistent, legally compliant charges by AI would promise considerable opportunities for efficiency gains.
35. Naturally, AI-supported indictment drafting would need to be transparent, explainable, and reliable. The process should be subject to ongoing evaluation for accuracy and fairness, and the system should empower officials to rapidly and effectively audit generated text. We believe that this bar can be met, and suggest strategies for doing so in our Ethics section.

#### ***2.2.5. Consistency in charging and review decisions***

36. Inconsistent charging decisions, in particular overcharging, is a source of key inefficiency in the court system. It leads to more cases being allocated to the Crown Court than should be the case, a burden on court time in pursuing or correcting charges, and a higher rate of ineffective trial dates due to cracked cases.
37. AI can assist in the improving the consistency and accuracy of charges in several ways:
  - 37.1. by providing guidance on applicable principles and precedents to improve rigour and consistency;
  - 37.2. by applying real-time review and oversight of charging decisions, continuously re-evaluating them as information changes, and flagging outliers for review;
  - 37.3. by maintaining a history of charging decisions to assess accuracy in real-time, identify outliers, and providing updated guidance.

#### ***2.2.6. Improving allocations***

38. The 2015 Review identified several issues with the current allocation process for either-way offences, which can be tried either in the Magistrates' Court or the Crown Court. These included inconsistency in allocation decisions, with different courts interpreting the

guidelines differently and over-allocation to the Crown Court where the eventual sentences fell within the power of the Magistrates' Court.

39. Pro-active case progression as outlined above can of course assist with improving allocations by encouraging early clarification of the issues. However, an AI "Allocations Assistant" could assist with improving decision-making in several ways:

39.1. the agent could assist the court in reaching an appropriate sentence estimation for alternative factual bases based on the guidelines and comparable cases;

39.2. the agent can also ensure that there is increased rigour in the application of guidelines, by ensuring that the court walks through the reasoning behind the decision;

39.3. by applying consistent principles and approaches in assisting different courts, the agent will be able to prompt more consistent decision making; and

39.4. by maintaining a history of decisions made, the agent will be able to identify which courts may be outliers in terms of allocation decisions.

40. Indeed, whatever future reforms may occur, the importance of allocating cases appropriately and consistently will remain. If, for example, we move to a unified criminal court system, it may become more important to ensure that decisions on allocation are still taken with rigour as formal administrative divisions are removed. The authors of this report support proposals to create a unified criminal court, and believe that the introduction of the technologies outlined here will further remove practical obstacles to effective implementation of such a change.

### **3. Deployment, Timelines and Adoption**

41. In this section, we will consider the steps to deploying AI in pursuit of the opportunities outlined in part 2 above. We will discuss what data is required, and how ontologies create an operational layer for your AI agents, the anticipated timelines within which applications like those described in part 2 above can be deployed, and some considerations around increasing adoption and minimising operational change.

### **3.1.Deployment**

42. There are several areas to consider in the deployment of AI workflows but the key to success lies three steps, first identifying the data your AI will access, secondly building your Ontology, and thirdly employing a “test and learn” approach for rapid iteration with end-users.

#### ***3.1.1. Data***

43. Just as any professional would need access to the right information and knowledge to accomplish some tasks, AI must also be provided with access to that same information. However, care should be taken to distinguish between providing information for your agent to access and use and providing data to “train” an LLM.

44. The latter describes the process of creating the core model which takes some media input and provides text output, and this requires considerable expertise, time, infrastructure, and expenditure. It is not necessary to for the applications set out above and indeed would be counterproductive. Instead, it’s best to use existing generative AI models in a secure environment, interacting with curated data to retrieve appropriate context to perform defined tasks.

45. The data requirements for the applications discussed in section 2 would be quite simple and could be easily integrated into any mature AI platform in hours or days. They might include:

- 45.1. an integration to the Crown Court Digital Case System (“DCS”) for case documents;
- 45.2. the CrimPR and relevant legislation accessed via gov.uk or uploaded manually;
- 45.3. case law; and
- 45.4. sentencing guidelines and principles from the Sentencing Council.

46. These sources should all be available under Crown Copyright or through HMCTS/MOJ.

### 3.1.2. *Ontology*

47. An Ontology is an operational layer that sits on top of the digital assets integrated into the platform, such as datasets and AI models, and connects them to their real-world counterparts. These counterparts can range from physical resources like judges or courts to abstract concepts like cases or orders.
48. It is critical to operational AI workflows as it provides the environment in which the agents and applications you configure can interact with the data and actions needed to accomplish some goals, as well as the auditability, governance and security controls required for any application of AI in the legal system.
49. For this context, the Ontology will serve as a digital twin of the criminal process, containing both semantic elements (objects, properties, links) and kinetic elements (actions, functions, dynamic security) needed to enable the use cases. Key components of the Ontology may include:
  - 49.1. Objects such as Cases, Judges, Court Officers, Representatives, Documents, Orders, Criminal Procedural Rules, Sentencing Guidelines, Legislation, Case Law and the links between them;
  - 49.2. Action Types and Functions: tools to retrieve documents from source systems, query and edit data, send emails, draft orders or upload documents to DCS; and
  - 49.3. Security and Governance Controls: permission defining which users and agents may access cases and documents, who may run specific actions, and full auditability of when actions have been run and data accessed or edited.
50. This Ontology will be configured to allow agents and users alike to easily navigate the knowledge and information within the case documents to engage effectively with key users, all while complying with governance, privacy and security requirements. Once established the Ontology can easily be extended and re-used to create new workflows at diminishing incremental cost.

### **3.1.3. Test and Learn**

51. Once the data and Ontology is established it is essential to adopt a “*test and learn*” approach to demonstrate and improve the efficacy of any workflow. This encompasses:
  - 51.1. Identifying how the users can best interact with the workflow, accounting for operational constraints;
  - 51.2. Testing the logic and functionality of the AI Agents;
  - 51.3. Testing the usability and effectiveness of the developed workflow; and
  - 51.4. Tracking and quantifying efficiency gains.
52. This should be done in rapid iterations to maintain a fast feedback loop and ensure that the workflow will be effective.

### **3.2. Timelines**

53. In deploying AI, you should always aim to be delivering value in weeks not months or years. For the defined applications, the data integration and Ontology set up could be completed within days and thereafter each application set out above can be configured, tested and deployed in weeks. None is dependent on complex integrations, and in each case the only constraint will be how quickly you can iterate with end-users when refining the agents and interfaces. Overall, the above workflows could be deployed and delivering efficiency gains within ~12 weeks.

### **3.3. Adoption**

54. The final hurdle in successful deployments is achieving adoption by end-users. In most cases this requires not only build tools and agents that the user values, but also minimising disruption, and navigating operational and procedural constraints.
55. Traditional approaches emphasise the importance of training and change management, but with AI we can aim to create end solutions that are natural and intuitive, and as much as possible to “meet the user where they are”. This may mean utilising existing tools such as email as the interface for many users and otherwise designing applications that are easy to use and intuitive such as chat interfaces. Users should be able to benefit from the tools you deploy with minimal instruction, and no extra skills or training.

56. Moreover, with the applications we have set out in section 2 above we have consciously focused on deployments where the agent would be acting within the current operational process, acting only as an advisor or support assistant to parties or coordinating their response to issues. This will effectively minimise any operational or procedural barriers and avoid the need for lengthy training or operational change management.
57. Finally, successful government AI initiatives often excel through a "*field-to-learn*" or "*test-fix-test*" approach. This strategy is not only an effective approach to implementation of workflows, but it also promotes rapid experimentation and iteration, enabling end-users and operators to implement AI in contexts pertinent to their missions. By involving operators directly in testing and refining AI solutions, this approach facilitates both technical innovation and ethical boundary-setting. It provides technologists, ethicists, policymakers, and end-users with a deeper understanding of AI deployment challenges, moving beyond theoretical proposals detached from operational realities. Encouraging agencies to adopt such programs can foster responsible and iterative AI innovation within the courts and provide the environment for broader adoption of AI in a responsible and principled way.

#### **4. Ethics and Principles**

58. In this section we consider several concerns frequently raised in relation to the deployment of AI within criminal justice. First, we consider the ethical implications of adopting AI, and why it is imperative that we consciously deploy AI in ways that are consistent with established ethical principles and rather than deploy technology and shift our principles to accommodate its capabilities. Second, we consider some risks around AI that must be managed in any deployment. Finally, we set out the principles by which AI can be deployed effectively, safely and securely.

##### **4.1.Ethics**

59. There are a range of critical ethical concerns to consider in the deployment of AI to criminal justice: the fear that algorithms and models applied will be biased against certain characteristics and unable to see past probabilistic generalisations to the specific facts of a

case; the objection that defendants and witnesses will be reduced to statistics, judged not as individuals but data to be processed; the fundamental concern that introducing AI will undermine the critical humanity, respect and dignity of the justice process.

60. Such concerns must be considered carefully. It would be antithetical to any concept of justice to deploy statistical machine learning to make judgements based on predicted outcomes. It would be contrary to the purpose of the justice system to replace judges with AI or have AI make any decisions that might impact the outcome of a matter whatsoever.

61. However, as we have shown through the application defined in part 2, it is possible and indeed much more common and effective, to deploy AI to support and elevate the users' judgment, rather than to replace it. In those circumstances utilising AI as a tool is an effective and necessary instrument of justice creating more consistent, robust and transparent decision-making. Indeed, as the court backlog grows and resources become increasingly strained, it is more urgent than ever to embrace such opportunities to continue to deliver swift justice and maintain public confidence in our justice system.

62. To resolve the two positions requires an approach to AI that embraces its opportunities while remaining vigilant to the effects of its misapplication. Our experience with technology demonstrates that this is a continuous struggle. It requires a deep understanding of the technology but more importantly the discipline to embed established principles within the systems we build, and determination to shape technology to the ethical framework rather than compromise in pursuit of false promise.

#### **4.2.Principles**

63. As such, it is crucial to set out the key principles that should be applied to any deployment of AI in criminal justice.

#### **4.2.1. Governance**

64. Data systems used in sensitive domains like justice should implement privacy-by-design, and should feature the full range of governance tools necessary to ensure respect for the rights and reasonable expectations of affected parties.
65. User access to data and tools must be subject to controls, and these must permeate through to any AI agents to prevent unauthorised disclosure or actions. This must be a core component and consideration of any platform, not an afterthought. The addition of AI agents to a justice system should in no way compromise established data protection structures and conventions.
66. Every deployment should have clear operational controls. A dedicated team should evaluate workflows to ensure compliance with criminal procedure, human rights and the rule of law. Core values should be instilled to appropriate users and developers through training and compliance monitoring.
67. Every deployment of AI should contain checks that ensure a human user has evaluated the output, applied their own judgment, and taken accountability for outcomes. Checkpoints and user acknowledgments should be enabled on all actions to ensure compliance with this principle.

#### **4.2.2. Transparency**

68. All actions taken by agents and users should be tracked within the platform, and the reasoning applied throughout should be clear to users and available for review if required. Audit chain should exist for any action taken in the platform, with clear lines of accountability to end-users.

#### **4.2.3. Evaluation**

69. All agents should be evaluated during configuration and then on a continuous basis. This will consist not only of reviewing output, but also checking reasoning to ensure that agents

are not only reaching the right conclusion but doing so for the right reasons. This should be built into the platform and operational process. Any platform used must have the capability to manage this requirement at scale without creating significant overhead.

70. A rigorous testing and evaluation programme should make use of automated checks, tracking the evolution of key metrics over time, and regular escalation of changes for expert human review.

### **An Oversight Mechanism**

71. The safe and ethical use of AI processes as part of the administration of justice requires good governance, transparency and evaluation. We propose that an oversight body or committee, led by senior judiciary but with membership including technical experts and recognised ethical specialists, should be set up in order to regularly monitor the situation in order to avoid the traps that the failure to properly monitor the use of other, older technologies have led the criminal justice system into on previous occasions.

## **5. AI Office**

72. We recommend that the England and Wales criminal courts establish a Chief Artificial Intelligence Officer (“CAIO”) to ensure the responsible adoption of AI technologies in a manner that enhances trustworthiness and accountability. Introducing a CAIO position within existing court structures—rather than external authorities—and ensuring that the office collaborates closely with current officials and organisations, will align AI initiatives with existing governance and institutional mandates effectively.
73. Drawing from extensive experience with technology roles in similar government settings, we propose the following recommendations to empower and make CAIOs effective in their roles.

### **Empowerment and Authority**

74. It is crucial that the CAIO position is endowed with significant seniority and operational authority to prevent it from becoming merely ceremonial. CAIOs should report directly to senior judiciary and ministers, to ensure their involvement in mission-critical projects and

decision-making processes. Their placement should allow them to influence technology acquisition and development substantively.

### **Focused Initiatives and Credibility**

75. The effectiveness of senior technology officers is often determined by the focus and framing of their initial goals. Officers who start with a specific project will achieve a better understanding and credibility regarding the operational realities and technology needs of the courts. By securing early, tangible successes, the CAIO can build trust and gain the authority needed to address broader court objectives.

### **Collaborative Ethics and Accountability**

76. While the CAIO will oversee the ethical aspects of AI acquisition, development, and deployment, it is essential that they work closely with existing officials who have overlapping responsibilities in ethics and accountability. Collaboration with dedicated Privacy and Civil Liberties engineers and those who have experience in managing the intersection of technology and fundamental rights, will be invaluable, as will oversight from senior judiciary, technologists and legal ethicists who can review systems regularly. AI is not standalone; thus, the CAIO will benefit from integrating their responsibilities with those of other officers, enhancing overall programmatic and technological standards.

### **Collaboration and Learning**

77. The CAIO should have the flexibility to delve into specific projects within the courts while also engaging in wider justice and government forums. These forums would facilitate regular meetings for sharing challenges, lessons learned, and best practices across different court systems. Such collaboration will promote information sharing and mission accountability, and support cross-department AI initiatives. While not aimed at creating regulatory standards, these forums could enhance alignment and consistency in desired outcomes across various judicial bodies.

78. By establishing a CAIO, the England and Wales criminal courts can seize the opportunities presented by AI while ensuring it is integrated responsibly and effectively, aligning technological advancements with judicial integrity and public trust.

## **6. Financial and Operational Considerations**

79. In preparing this submission we have been mindful of the financial and operational context in accordance with the scope of the review. As we have set out in section 3 above, any workflows should seek to integrate seamlessly into current operational processes with minimal disruption. We believe that any of the applications of AI set out in section 2 can be deployed in accordance with this principle.

80. Additionally, consistent with the “test and learn” approach set out in section 3, AI should be deployed through an iterative approach which continuously assesses efficacy and ROI of each use-case. This entails approaching AI not as a large-scale procurement exercise with multi-year horizons, but as a series of discrete use-cases, each with clearly articulated benefits, that can be delivered quickly against and with a positive return on investment.

## **7. Conclusion**

81. AI offers clear opportunities to increase efficiency and timeliness of the criminal court process. There are several clear and discrete opportunities for its deployment that can be enacted in a matter of weeks with minimal operational friction. The Ministry of Justice should urgently pursue these opportunities through well-structured and principled deployments, as they offer immediate opportunities to speed up justice for victims and defendants currently awaiting trial, reduce the backlog and improve public confidence in the justice system.

## ANNEX 2

### THE POLITICS OF AI REGULATION: AN OVERVIEW

Speech delivered on 28<sup>th</sup> January 2025 by RT HON SIR ROBERT BUCKLAND KBE KC, DAC Beachcroft Policy Unit.

2025 is shaping up to be a year of choices for the UK. Do we go for closer trade links with the EU or with the USA? As we search for economic growth, do we change tack on regulation and planning? And in the field of AI and innovation, do we follow the EU model of regulation, or do we align with the light touch approach now being taken by the new US Administration? Or are these binary choices, in fact, not the reality?

I have been considering these issues carefully over the past few years, with a particular focus on the impact that AI and machine learning will and is having on the administration of justice and the practice of law. This has been the subject of my Senior Fellowship at the Mossavar-Rahmani Center for Business and Government at Harvard Kennedy School. I should add that I chose this topic in early 2022, before the launch of ChatGPT and the explosion in use of LLMs!

Debates and conversations about this topic are often based upon misconceptions, so as all good lawyers do, let's start with definitions. I asked Chat-GPT the following question: *"What is the clearest definition of artificial intelligence (AI)": "A branch of computer science focused on creating systems or machines that can perform tasks typically requiring human intelligence. These tasks include learning from experience (machine learning), reasoning and problem-solving, understanding natural language, perceiving the environment, and making decisions.*

*This definition encompasses both the specific techniques used to create AI systems (like algorithms and neural networks) and the broader goal of enabling machines to emulate cognitive functions associated with human intelligence."* This is a pretty good summary, I think, having verified it against other sources!

It is also vital that we understand *how* AI is being developed. Two months ago, I visited Stanford University to talk about their Annual AI Index with some of their academics. In the Index, it was noted amongst other things that in 2023, industry produced 51 notable machine learning models, while academia contributed only 15. There were also 21 notable models

resulting from industry-academia collaborations in 2023, a new high. According to AI Index estimates, the training costs of state-of-the-art AI models have reached unprecedented levels. For example, OpenAI's GPT-4 used an estimated \$78 million worth of compute to train, while Google's Gemini Ultra cost \$191 million for compute. In 2023, 61 notable AI models originated from U.S.-based institutions, far outpacing the European Union's 21 and China's 15. New research from the AI Index reveals a significant lack of standardization in responsible AI reporting. Leading developers, including OpenAI, Google, and Anthropic, primarily test their models against different responsible AI benchmarks. This practice complicates efforts to systematically compare the risks and limitations of top AI models. Despite a decline in overall AI private investment in 2023, funding for generative AI surged, increasing nearly eight-fold from 2022 to reach \$25.2 billion. Major players in the generative AI space, including OpenAI, Anthropic, Hugging Face, and Inflection, reported substantial fundraising rounds.

A survey from Ipsos shows that, over the last year, the proportion of those who think AI will dramatically affect their lives in the next three to five years has increased from 60% to 66%. Moreover, 52% express nervousness toward AI products and services, marking a 13-percentage point rise from 2022. In America, Pew data suggests that 52% of Americans report feeling more concerned than excited about AI, rising from 38% in 2022.

The sheer pace of change makes these figures already dated, but they provide an important backdrop: unlike development of the internet thirty years ago, which in many ways was the creation of governments, the AI revolution is being driven by industry and the private sector.

And only this month, we have news of the new Stargate AI infrastructure initiative from Open AI, SoftBank and Oracle, publicly endorsed by President Trump, with \$500b of investment and a removal of guardrails around technological development. There are some question marks about the extent to which this initiative is actually funded, and all of this is taking place against the continuing argument about the use of open source AI to develop models, which raises safety concerns but at the same time concerns about restrictive practice and the accrual of power and regulatory capture by the tech giants if a closed source approach is taken. In the last few days, the launch of China's DeepSeek start up LLM has cast doubt on

the assumed dominance of the US market in AI development and how this is proving to be a breakneck race for supremacy.

Amidst all of this sudden change, the regulatory challenge for national and supranational governments, is, therefore, huge. How to create effective guardrails that don't hamper innovation and investment and which will not quickly become left behind by fast-moving technological development. Should we let a hundred flowers bloom and have no regulatory framework? Is it too late in any event?

Nothing is new under the sun, and I think we can learn some lessons from failures to oversee previous technologies. Let's look at two UK examples.

From 1<sup>st</sup> April 2019 to 30<sup>th</sup> September 2023, 3,102,392 criminal cases were received into the SJP, including 609,164 via the digital service. The main criticism of the digital Single Justice Procedure system is one related not to its internal operation, but to external human factors, leading to unfairness when dealing with vulnerable defendants. This resulted in an MoJ review and further assurances about human input as the system is rolled out even further.

Last year, the UK Parliament took the unprecedented step of passing legislation that will have the effect of removing or quashing of criminal convictions for theft, fraud and false accounting recorded against hundreds of sub-postmasters who were responsible for local branches of the Post Office network. This measure was taken after the scale and extent of an underlying IT failure in the HORIZON accounting system used across the network became increasingly apparent after a long-running campaign by aggrieved postmasters and a number of Court of Appeal cases where convictions were overturned.

One of the key questions is why there was no forensic examination by suitable experts of the system and its outputs. The question of the ability to challenge IT evidence is one that has had a wider impact on public confidence both in machine processes and the ability of the system itself to be transparent and to acknowledge failure. The showing of a major TV drama based upon the scandal over Christmas 2023 attracted millions of UK viewers. This, therefore is not an issue of marginal public importance, and highlights the human context in which machines and technology old and new must operate.

It is important, therefore, to draw distinctions as to matters that are intrinsic to AI systems and extrinsic factors such as human oversight. What can be said is that the speedy nature of automated procedures means there is less opportunity for there to be an enlightened intervention and more risk of error being compounded. The lessons for the future use of AI systems are clear; without careful planning as to precisely how they are to be integrated, they are less likely to deliver justice in the eyes of the public.

Looking at international moves to set out frameworks for AI regulation, we have the Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law was signed by all members plus a range of other countries including the USA in September 2024. In summary, its provisions are firstly that activities within the lifecycle of AI systems must comply with the following fundamental principles: human dignity and individual autonomy; equality and non-discrimination; respect for privacy and personal data protection; transparency and oversight; accountability and responsibility; reliability and safe innovation, with specific requirements on transparency of information regarding AI systems, that is sufficient to allow challenge to the system and its decisions, the ability to lodge a complaint and to provide procedural safeguards to affected persons, with notice to be given of interaction with an AI system rather than a human being.

Finally, there are provisions to deal with the ongoing monitoring of the impact of AI use, by carrying out risk and impact assessments in respect of actual and potential impacts on human rights, democracy and the rule of law, in an iterative manner, and then to establish sufficient prevention and mitigation measures because of the implementation of these assessments with the option for the authorities to introduce ban or moratoria on certain application of AI systems (“red lines”).

The speed of ratification and the extent to which this applies beyond the public sector are matters that have yet to be fully seen, plus the position of the USA, which I shall come to shortly.

The Bletchley Park AI Safety Summit in late 2023, hosted by the UK and involving 27 countries including the USA, China, India, Brazil and the EU, was a major attempt at an intergovernmental level to outline the opportunities and risks posed by developing AI. The principles of the protection of human rights, transparency and explainability, fairness,

accountability, privacy and data protection are all contained within the Final Declaration. Whilst parts of the Declaration focused on “Frontier” AI development, the most relevant statements are these:

*“We encourage all relevant actors to provide context-appropriate transparency and accountability on their plans to measure, monitor and mitigate potentially harmful capabilities and the associated effects that may emerge, in particular to prevent misuse and issues of control, and the amplification of other risks.”*

At the Seoul AI Safety summit in May 2024, it was agreed that risk thresholds for frontier AI development and deployment would be developed, agreeing when model capabilities could pose “severe risks” without appropriate mitigation, for example the danger of AI evading human oversight by manipulation and deception or by autonomous replication and adaptation or secondly the risk of frontier AI capability to assist non-state actors in advancing the development, production, acquisition or use of chemical or biological weapons. A third AI safety summit is set to open in Paris next week.

Running in parallel with this has been the “Hiroshima Process”, started by the G7 at the 2023 Summit, which is work being taken on by the OECD. This process relates to the development of Generative AI, but in their Declaration of April 2023, the G7 Digital and Technology Ministers undertook *“to convene future G7 discussions on generative AI which could include topics such as governance, how to safeguard intellectual property rights including copyright, promote transparency, address disinformation, including foreign information manipulation, and how to responsibly utilise these technologies.”*

Another thread is the United Nations AI Advisory Body which issued its final report in September 2024 with seven key recommendations, namely the creation of an international scientific panel on AI, a twice-yearly intergovernmental and multi-stakeholder dialogue on AI governance on the margins of existing UN meetings, an AI Standards Exchange, a Capacity development network, a global fund for AI, a global AI data framework and an AI office within the UN Secretariat. The question is to what extent this will influence international AI development, and frankly I am inclined to be somewhat sceptical.

We have some tangible regulation closer to home. The EU Artificial Intelligence Act 2024 is the world’s first comprehensive regulation on artificial intelligence. The AI Act is designed to ensure that AI developed and used in the EU is trustworthy, with safeguards to protect people’s

fundamental rights. The regulation aims to establish a harmonised internal market for AI in the EU, encouraging the uptake of this technology and creating a supportive environment for innovation and investment.

Its scope is broad and forward-looking, as the definition of AI includes “software that...for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations or decisions influencing the environments they interact with.” Systems like chatbots are covered, but reference to “predictions” and “recommendations” cover many software applications that aren’t conventionally regarded as AI.

The Act applies to non-EU based AI providers that serve EU citizens and to EU based providers that serve citizens in any jurisdiction. An American AI company serving EU users would be covered by the Act, as would a French company serving Canadian users. Free and open-source systems, save for those classified as high risk, are exempt from the Act. AI systems solely used for military purposes are also exempt.

The premise of the applicability of the Act’s provisions is based upon the level of risk. The higher the risk, the more stringent the restrictions. This ranges from no restrictions to an outright ban on systems deemed to have an unacceptable risk. The Act classifies AI systems into four buckets: prohibited, high risk, limited risk (including general purpose AI models like ChatGPT) and minimal or no risk. Justice activities fall into the high-risk category.

On 20<sup>th</sup> January 2025, on the day of his Inauguration as President, amongst a flurry of activity, Donald Trump signed an Executive Order rescinding the AI Executive Order issued by President Biden on 30<sup>th</sup> October 2023. At a Federal Level, therefore, the USA has no guidance to give on AI regulatory policy. It looks as if Elon Musk has got his way, but only last year, he was one of the more vocal proponents of a Bill in California that would, amongst other things, have mandated a kill switch for frontier generative AI developments before Governor Newsome vetoed it last September. This may have been nothing more than an attempt to stifle Open AI, with whom he has a well-known falling-out (literally), but it does illustrate the multi-sided nature of the argument.

In reality, it is at State level that AI regulation is on the move. More than 40 state AI bills were introduced in 2023, with Connecticut and Texas actually adopting statutes. Both of

those enacted statutes establish state working groups to assess state agencies' use of AI systems to ensure they do not result in unlawful discrimination. Colorado introduced the first comprehensive AI regulation measure dealing with high risk AI developers and users in May last year.

In China, a four new AI regulations was introduced in 2022 and 2023. Most interestingly, in Article 4 of the Generative AI i, there is an obligation that AI Generated Content reflects core socialist values amongst obligations to respect “legitimate” rights and interests of other individuals.

In Article 6 of the Algorithm Recommendation Regulation, there is a requirement for adherence to “mainstream values” and that services are designed “to actively spread positive energy”. In short, these regulations are designed to control and limit what goes into systems, rather than mitigate the effects of their dissemination.

These regulations impose obligations on service providers, technical supporters and users as well as other entities including online platforms, with significant penalties for non-compliance.

On many levels, the restrictive nature of the control of content is not at all consistent with the approach taken in free societies, but, stripping aside the politics, it can be said that in certain settings in rule of law democracies, for example with justice AI systems, there will be a need to rigorously control and monitor the datasets used to ensure that properly sourced and verified legal information is within datasets used by lawyers and judges, with such decisions being made by independent judges and lawyers, not governments.

What, then, for the UK? From the references in the Government's new AI Opportunities Action Plan to the UK's pro-innovation regulatory approach it does look as if the UK wants to place itself between the USA and the EU as a distinct entity, whilst seeking to strike the right balance between ethical safety and innovation. Here, we might see a real difference that is now available because of Brexit, with the caveat that too much regulatory divergence might be seen as an unwelcome extra cost by investors.

Rightly, the Plan has made it clear that they expect existing regulators to look at AI issues within their own sectors, which to my mind will be the most rapid way to deal with oversight issues, albeit with two key dangers: 1. The inability to provide sufficient scrutiny and 2. A lack of focus on promoting innovation and investment.

Alongside the Action Plan, the Government's published consultation on copyright law sees the familiar battlelines being drawn between content creators who want an opt-out system and existing publishers such as traditional news media who want it the other way round.

Of the many decisions that the Government has to make this year, the choices on AI are the most important and consequential. My final word on this is that our own sector cannot just sit back and let Government do the heavy lifting. We will have to work harder, as lawyers and those responsible for the administration of justice, to ensure that AI systems used in UK justice are ethical, reliable, explicable, challengeable and accountable. As an optimist, I believe that we can do this, but the time to act is now.