NEW SOUTH WALES
BAR ASSOCIATION

Our ref: 23/582

19 April 2024

The Hon Jeremy Buckingham MLC
Chair
Portfolio Committee No. 1 – Premier and Finance
NSW Parliament

By email: PortfolioCommittee1@parliament.nsw.gov.au

Dear Chair,

Inquiry into Artificial intelligence (AI) in New South Wales – Questions on Notice

1. The NSW Bar **Association** refers to the evidence given on behalf of the Association to the Inquiry into Artificial intelligence (**AI**) in New South Wales by Dr Ben Kremer SC on 11 March 2024. The Association provides the following responses to the Questions on Notice directed to Dr Kremer during his evidence.

Question at Pages 27-28, Transcript of proceedings on 11 March 2024

A. Question

> **The Hon. Dr SARAH KAINE:** I have a question for you, Dr Kremer, about your submission as well. I am interested in the second case study about using AI to replace or supplement judicial decision-making. Obviously, that has come out of concerns, it seems to me, about existing use or places where it has been used. Could you give a bit more information about the examples that might have informed that particular case and recommendations?
>
> …
>
> **The Hon. Dr SARAH KAINE:** Thank you for taking that on notice—I would be really interested. This hasn't come out of any Australian jurisdiction or judicial body? It's more the concern that this type of use of AI in the judicial process has been seen elsewhere?

B. Answer

2. The Association is not aware of the use of an AI system in Australia to replace or supplement judicial decision-making. Indeed, the Association has concerns whether that would be legally permissible for at least three reasons:

   (a) first, it is not at all clear that the statutory language reposing the performance of tasks (including the making of decisions) in judges would extend to allowing those tasks to be performed by software in place of a judge. It is likely that if the software performs the task

in whole or possibly in part, then (in law) the task may be considered not to have been performed properly, or indeed performed at all. This question may be more difficult to answer if the task is performed by the judicial officer but software has influenced, perhaps to a high degree, their performance of that task;

(b) secondly, where a judge is exercising federal jurisdiction, additional constitutional considerations (relating to the nature of a Chapter III court and the function of a decision-maker in such a court) would intrude;

(c) thirdly, the use of an AI to perform would likely introduce potential consequences for the treatment of such decisions on appeal or via a challenge to their validity, including ascertaining what were the real reasons (if any) for the decision made by the AI.

3. The Association is aware that 'predictive risk assessment' tools, which include predictions of the likelihood of future criminality by a particular person, provide information that informs a number of criminal procedure decision points including bail, sentencing and parole. Tools such as 'Level of Service Inventory – Revised' (LSI-R), 'Static 99 Revised' (Static-99R) and STABLE-2007 are often used by external psychologists or Corrective Services NSW to consider risk of reoffending based on criminogenic needs, and risk of sexual reoffending.[1] Assessments 'as to the likelihood of persons with histories and characteristics similar to those of the offender committing a further serious offence' may also be used when considering post-sentence orders under the *Crimes (High-Risk Offenders) Act 2006* (NSW) for high-risk sex and violent offenders.[2] The 'Violent Extremism Risk Assessment Version 2–Revised' (VERA–2R) tool is often used when a court is considering whether to make an extended supervision order or a continuing detention order under the *Terrorism (High Risk Offenders) Act* 2017 (NSW).[3]

4. The Association does not understand these or similar tools to involve AI as understood in the Committee's Terms of Reference. Rather, the Association understands that they involve algorithmic calculations that can be performed manually or by the use of software into which the user inputs data, and where the software simply operates a pre-determined algorithm (e.g. giving weights to particular inputs to arrive at an overall score).

5. Concerns have been expressed about the accuracy of these tools in their current form, including whether the methodology on which they are based encodes (and therefore reproduces) bias, such as by over-weighting or under-weighting particular factors, or including irrelevant factors or omitting relevant factors.[4] The Association's submission to the inquiry notes criticism of the VERA-2R, but as far as the Association is aware, there has not been any authoritative determination as to the validity of any of these tools. The existence of controversy in respect of 'deterministic' tools (i.e. ones which involve a straightforward algorithmic calculation, where the algorithm is known, the calculation can be repeated and whose steps leading up to the output can be followed to see the contribution to the final output) provides one reason to be cautious about

[1] See, for example, *State of New South Wales v Barrie (Final)* [2018] NSWSC 1005 at [69], [87], [112], [140]; *State of New South Wales v Graham James Kay* [2018] NSWSC 1235 at [38]-[46].
[2] Section 9(3)(d) (for an extended supervision order), and section 17(4)(d) (for a continuing detention order or extended supervision order).
[3] See sections 25(3)(c) and 39(3)(b) respectively.
[4] Hart, SD, Michie, C, and Cooke, DJ (2007). Precision of Actuarial Risk Assessment Instruments: Evaluating the "margins of error" of group v. individual predictions of violence. British Journal of Psychiatry. 190(49), 60-65.

using a 'non-deterministic' (i.e. AI) model where these steps may not be fixed, and where the same tool may produce different outputs when run at different times.

6. The Association is also aware of controversy about the use, in the United States, of computerised models where the algorithms used are kept secret (usually for reasons of commercial confidentiality), so that while the inputs and output are known to the user, the steps performed to arrive at the output are not known or able to be assessed or evaluated. One example is the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) software used in many United States courts to predict recidivism for the purpose of bail decisions, which has been criticised for exhibiting a bias that under-predicts future recidivism by white prisoners, and over-predicts recidivism by black prisoners.[5] Given that the Association understands that AI tools are likely to be based upon the operation of models whose operation is not fully understood, or indeed understood at all, and where there is no single algorithm but a complex interaction within the AI software that may change over time, the concerns about 'closed-source', 'black-box' algorithms are heightened for AI systems.

7. The Association is also aware of a number of other developments which give reason to believe that AI software will be developed with the aim of performing, or supplementing, adjudicative functions for potential use within New South Wales.

8. First, a number of foreign jurisdictions are trialing or using systems that incorporate AI in relation to adjudicative functions, to varying extents:

   (a) Brazil's Supreme Federal Tribunal uses software that automates the examination of appeals and provides recommendations on, inter alia, legal precedents and potential courses of action.[6] Over 40 other courts in Brazil use AI programs to categorise legal resources and identify applicable precedents.[7]

   (b) In March 2022, Saudi Arabia introduced 'virtual enforcement courts' which operate without human intervention.[8] The courts are intended to streamline a 12-step litigation process down to two steps.

   (c) In the United Arab Emirates, the Abu Dhabi Judicial Department introduced a smart court initiative in 2022 which uses AI to enhance and expedite adjudication processes.[9]

   (d) Estonia explored automating small contract disputes, but has since stated that it has no immediate plans to introduce automated courts.[10]

---

[5] Angwin, J et al 'Machine Bias', ProPublica (online, 23 May 2016) (available here).
[6] Eduardo Villa Coimbra Campos, 'Artificial Intelligence, the Brazilian Judiciary and Some Conundrums', SciencesPo (Blog Post, 3 March 2023) (available here).
[7] Brehm K et al, The Future of AI in Brazilian Judicial System: AI Mapping, Integration, and Governance (Report, 2020) 14 (available here).
[8] 'Justice minister inaugurates Virtual Enforcement Court in Saudi Arabia', Zawya (online, 28 March 2022) (available here).
[9] Rasheed, A, 'Abu Dhabi criminal cases now followed up by artificial intelligence', Gulf News (online, 8 August 2022) (available here).
[10] 'Estonia does not develop AI Judge', Republic of Estonia Ministry of Justice (Web Page, 16 February 2022) (available here).

(e) China has adopted the most significant measures in aiming to develop 'smart courts' which integrate AI in dispute resolution by 2025, and internet courts are already operational.[11] The smart courts would allow court users to commence actions, serve documents, present evidence and resolve disputes online, including disputes involving intellectual property, e-commerce, financial disputes and product liability.[12] The courts have adopted a number of digital advancements to court procedures including automatic generation of pleadings, litigation risk assessment and document writing assistance.[13]

9. Secondly, the Association understands that the Judicial Commission of NSW has been developing 'Bail Assistant' software to guide decision makers through the complex process set out in the *Bail Act 2013* (NSW). According to a 2021 speech by the former Chief Justice, the Bail Assistant will be 'intended as a tool to support the judicial officer from start to finish to assess bail concerns efficiently, make an informed bail decision, and record the decision accurately', but is also 'designed to be a supervised machine learning system, which could use data from past bail decisions to predict probable outcomes and to bring up relevant precedent', although only as a guide rather than as a decision maker.[14]

Question at Page 30, Transcript of proceedings on 11 March 2024

A. Question

> **Ms ABIGAIL BOYD:** I want to pick up on facial recognition technology by police, which is mentioned in the Bar Association's submission in some detail, which is great. I've asked this question previously of some other witnesses as well. I asked about this in estimates in the weeks just gone, around the hundreds of leads for investigation that were generated using facial recognition technology. Apparently the systems we're using are based on some Cognitech technology which was part of the study in the US that found that faces of people of colour were 10 to 100 times more likely to be falsely identified than Caucasian faces. When I raised this in estimates and asked, "Are you concerned about the bias?", I was told there was no testing for that bias but also it wasn't a problem because it wasn't the only evidence used to charge someone. Does it concern you that people are being contacted on the basis of leads from facial recognition technology that could be biased even if that doesn't lead to actual claims? Can you talk about that? We'll start with you, Mr Kremer.
>
> …
>
> **Ms ABIGAIL BOYD:** The argument from New South Wales police appears to be that it's okay because a human is involved before they get to the point of making an arrest. From a human rights perspective, or from concerns around over-policing, is that sufficient if we've got systems that have bias, but then there's an individual that gets involved before action is taken?

[11] *White Paper on Trials of Beijing Internet Court* (White Paper, 14 October 2019) 5 (available here).
[12] Ibid.
[13] Ibid.
[14] Chief Justice T F Bathurst, *Sir Maurice Byers Lecture 2021: Modern and Future Judging* (3 November 2021), [48].

B. Answer

10. A threshold issue is whether the facial recognition technology (**FRT**) used by the New South Wales Police Force does, or may, exhibit any bias, and in particular some form of bias between different races. The Association is not aware of the precise technology or technologies used by police. While it is aware of significant controversy over the accuracy and use of commercial facial recognition software overseas, the Association is not aware of any studies which conclusively prove (or disprove) the existence of any form of particular systemic flaw or inaccuracy across racial lines.

11. As a result, the questions on notice can only be answered on a hypothetical basis. There are a number of key issues. The first, and fundamental, issue is what is meant by 'bias'. The Association understands that the term is usually used to mean that the FRT exhibits different degrees of accuracy between different races.[15] That is, when performing the task of identifying which images(s) contained within a database of photographs match a particular photograph (e.g. a picture taken of a suspect at or near a crime scene), the software may (for example) be 99% accurate when matching a white male, 98% accurate matching a white female, 94% accurate when matching a black male, and 79% accurate when matching a black female.[16] The result of such an error may mean that a matching image within the database is not detected (i.e. a false negative), or that the image of a different person within the database is falsely reported as matching the relevant photograph (i.e. a false positive).

12. The second issue is what task the FRT is being used for, and hence what the result of the error (i.e. false positive or false negative) may mean:

    (a) a false negative when used during an investigation might mean that an offender who should have been identified by the FRT is not identified, and may, for example, not become a person of interest in a police investigation. A false positive in such a process may mean that an innocent person is wrongly identified as a person of interest, and thus is investigated further. This may not involve any further action involving the person if there is some reason ruling the person out (e.g. the person is the wrong sex or race, or was provably not at the relevant place at the relevant time), or it could perhaps result in them being surveilled or approached for questioning.

    (b) if the relevant process involves police using FRT to control access to a public place, or an event open to members of the public, then a false negative might lead to a person not being stopped for further screening while entering, but a false positive might result in an innocent person being impeded (and, for example, subject to secondary screening or searching) while going about activities of daily life.

    (c) if the process involves the police using FRT as evidence to charge or prosecute an individual for a particular offence, then a false positive could result in a miscarriage of justice.

[15] Furl, N, Phillips, J and O'Toole, A J, "Face recognition algorithms and the other-race effect: computational mechanisms for a developmental contact hypothesis", Cognitive Science, Volume 26, Issue 6, 2002, pp 797-815.
[16] E.g. Buolamwini, J and Gebru, T, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification", Proceedings of Machine Learning Research 81:1–15, 2018.

13. Different policy issues arise in all of the above cases.

14. The third issue - related to the second - is what role the FRT has in the relevant process in which it is being used. For example, is it being used as a sole decision-maker in an automated process, or is it at most a guide, perhaps one of many, available to a human? Alternatively, is it being used to enliven powers already granted to a person, such as to entitle a police officer to 'suspect on reasonable grounds' the matters needed to exercise powers to 'search persons and seize and detain things without warrant' (s 21), to 'search vehicles and seize things without warrant' (s 35) and to 'arrest without warrant' (s 99) under the *Law Enforcement (Powers and Responsibilities) Act 2002* (NSW)?

15. A fourth issue is the source of the relevant error(s), and whether or not they can be addressed and ameliorated, and if so over what timeframe. For example, it has been suggested that one cause of differential accuracy of an FRT is when it has been trained upon more images of some members of society than others, so that the errors are a result of a deficit in training data. A 2014 study suggested that one dataset, 'LFW', which is composed of celebrity faces, contained images that were 77.5% male and 83.5% white.[17] A particular FRT that can be trained to equivalent accuracy across all relevant members of society will be treated very differently from one whose inaccuracies cannot be corrected.

16. Although there is thus a very wide area in which errors of FRT could be material, the Association would be concerned if any government agency or body, particularly a police force, were using a technological tool that exhibited differential performance or effectiveness between people based upon their race, sex, or age, particularly if such use involved any potential detriment to a person due to inaccuracy. The depth of that concern would increase very greatly if the effect of any error in FRT led to differential or unjustified use of coercive or investigatory powers or processes. The effect of having an 'individual that gets involved before action is taken' would depend upon who that person is, what role they have, what the relevant action is, and any potential avenues of regulation or redress in relation to the process.

17. So far as AI is concerned, the uncertainties noted above in the use of FRT, when balanced against the importance of human rights, underlie the Association's third suggested prohibited practice (at [41](d) and [53] of its submission). That is, the Association has taken the approach of erring on the side of caution.

18. The Association understands that the use of FRT in general is being examined as part of the review of the federal *Privacy Act 1988* (Cth). The Association also supports the Australian Human Rights Commission's recommendation that there be a moratorium on the use of biometric technologies, including facial recognition, in high-risk areas of decision making until appropriate safeguards can be introduced.[18] That will address uses of FRT outside the scope of the questions on notice, such as collection and aggregation of images captured from persons using public spaces, or to 'track and control specific demographics', both of which pose difficulties with

[17] Han, H and Jain, Anil K. Age, gender and race estimation from unconstrained face images. Dept. Comput. Sci. Eng., Michigan State Univ., East Lansing, MI, USA, MSU Tech. Rep.(MSU-CSE-14-5), 2014.
[18] Australian Human Rights Commission, *Human Rights' and Technology Final Report 2021,* Recommendation 20.

respect to many human rights, including the rights to privacy, freedom of peaceful assembly and association, freedom of expression and freedom of movement.[19]

III. Transcript Correction

19. Page 28, at 7 lines up from the bottom of the page, "accord" should be "a court".

Conclusion

20. The Association thanks the Committee for the opportunity to respond to the questions on notice. Should you have any questions about the above, please contact in the first instance Lara Parmenter, Policy Lawyer at

Yours sincerely,

Ruth Higgins SC
President

[19] UN Committee on the Elimination of Racial Discrimination, General Recommendation No 36: *Preventing and Combating Racial Profiling by Law Enforcement Officials*, CERD/C/GC/36 (24 November 2020) [35].