I, LAWYER AND THE ETHICS OF ARTIFICIAL INTELLIGENCE

This article examines salient aspects of a recent UK decision on the use of a novel automated facial recognition technology by the police and considers how they may shape or reshape discussions on the ethics of artificial intelligence.

Introduction

In the 2004 Hollywood science fiction movie, *I, Robot*, set in 2035 A.D., social robots can be found in every household. Certified to be "Three Laws Safe", such robots coexist peacefully with humans and are programmed to save humans from any danger. But things start to go awry when a rogue robot is suspected to be behind the mysterious suicide of a prominent scientist. Inspired by science fiction writer Isaac Asimov's book of the same name, the movie explored a common theme in Asimov's stories: can artificially intelligent robots be constrained by The Three Laws of Robotics?

Today, in 2020 A.D., we are 15 years away from the fictional world of the movie *I, Robot* and are still very much in the "I, Lawyer" age. But the progress of artificial intelligence (**AI**) in recent years has accelerated concerns that we could soon find ourselves in an I, Robot epoch without adequate regulation of AI. As the use of AI continues to pervade our societies, our professions and our daily life, the ethics of AI (or AI ethics in short) has become increasingly important.

A recent UK decision on the use of a novel automated facial recognition technology (**AFR**) by the South Wales Police Force offers an interesting case study on how issues of AI ethics may need to be considered. Whilst AFR did not appear to be cutting-edge AI technology, its use of machine learning (an integral component of AI) raises similar issues. Drawing from the latest literature on AI ethics, this article examines salient aspects of the AFR case and considers how they may shape or reshape discussions on AI ethics.

Brief Facts of the AFR Case

In *R (on the application of Edward Bridges) v The Chief Constable of South Wales Police and others*,[1] the South Wales Police Force (**SWP**) had deployed the use of live AFR technology for about 50 large public events in a two-year pilot project called AFR Locate. First used in 2017 for the UEFA Champions League Final, AFR Locate involved "the deployment of surveillance cameras to capture digital images of members of the public, which are then processed and compared with digital images of persons on a watchlist compiled by SWP for the purposes of the deployment".[2] It was common ground that AFR Locate had been deployed overtly, and was not a form of covert surveillance.

Specifically, AFR assesses whether two facial images depict the same person, by taking and processing a digital photograph of a person's face to extract biometric data.

Subsequently, that data is then compared with facial images held on the watchlist. In the final "matching" stage of the AFR process, two outcomes are possible:

a. Where a possible match occurs, a human operator reviews the two images to confirm whether a positive match was in fact made. If no positive match is confirmed, no further action is taken. But if there is a positive match, a further assessment by other police officers stationed nearby is required before an intervention may be made (e.g. the person of interest may be stopped by a police officer for a conversation).

b. Where no match occurs, that person's data is automatically deleted, almost instantaneously, without any human observation at all.

A civil liberties campaigner by the name of Edward Bridges asserted that he was caught on camera through the use of AFR Locate on two occasions in December 2017 (at a busy shopping area) and March 2018 (at an exhibition). On both occasions, he stated that before seeing an AFR-equipped van, he was unaware of the use of AFR and had not been given any prior notice on its use. SWP accepted his evidence that he was present on both occasions and that on those occasions his image was recorded.

Mr Bridges brought a claim for judicial review on the basis that AFR was not compatible with the right to respect for private life under Article 8 of the European Convention on Human Rights (**the Convention**), data protection legislation and the Public Sector Equality Duty (**PSED**) under section 149 of the Equality Act 2010.

The High Court of England and Wales (**High Court**) dismissed Mr Bridge's claim for judicial review on all grounds. However, in a unanimous decision by the Court of Appeal of England and Wales (**Court of Appeal**), Mr Bridge's appeal succeeded on three out of the five grounds. This article focuses on two of these grounds for the purposes of the discussion on AI ethics:

a. The Court of Appeal held that the use of AFR Locate was not "in accordance with the law" under Article 8(2) of the Convention because it was not clear from the legal framework as to who could be placed on the watchlist and where AFR could be deployed. In this regard, it observed that too much discretion was given to individual police officers.

b. The Court of Appeal held that SWP was in breach of the PSED. In particular, it noted that the "human failsafe" component in the way in which AFR Locate was used was insufficient to discharge the PSED. More critically, it held that SWP had failed to take reasonable steps to make enquiries about whether the AFR Locate software had an unacceptable bias on grounds of race and/or sex.

The next two sections review the Court of Appeal's reasoning on these two grounds in greater detail.

"In accordance with the law"

Although both the High Court and the Court of Appeal agreed that AFR was a novel technology which went beyond "the taking of photographs or the use of CCTV cameras"[3] in public by the police, they differed on whether there was a clear and sufficient legal framework governing whether, when and how AFR Locate may be used.

It was undisputed that the three elements of the legal framework in question comprised the UK Data Protection Act 2018, the Surveillance Camera Code of Practice and SWP's local policies.

At first instance, the High Court found that each element of the legal framework provided "legally enforceable standards" and that the use of AFR Locate was "sufficiently foreseeable and accessible for the purpose of the 'in accordance with the law' standard" under Article 8(2) of the Convention.[4] The fact that AFR was a novel technology did not mean that it fell "outside the scope of existing regulation, or that it [was] always necessary to create a bespoke legal framework for it".[5]

The Court of Appeal, however, found that none of the three elements of the legal framework provided clear guidance on where AFR Locate could be used and who could be put on a watchlist. In particular, the Court of Appeal noted that SWP's Privacy Impact Assessment stated that besides "persons wanted on suspicion for an offence, wanted on warrant [and] vulnerable persons", any other person "whose intelligence is required" could be placed on the watchlist.[6] The Court of Appeal found that this final category was not objective as it could encompass "anyone who is of interest to the police".[7] As such, individual police officers had been given excessive discretion to decide who should be put on the watchlist.

As to where AFR Locate could be deployed, the Court of Appeal observed that SWP's Standard Operating Procedure and Data Protection Impact Assessment (DPIA) also did not provide any clear guidance. Although the DPIA stated that AFR Locate could be deployed for all types of events "ranging from high volume music and sporting events to indoor arenas", this was considered to be a "descriptive statement", rather than a "normative requirement", and the range of events stipulated was "very broad and without apparent limits".[8] The Court of Appeal pointed out that the DPIA could have, for example, indicated that the AFR Locate should be deployed at a location where it was reasonably envisaged that people on the watchlist would be present. Again, it found that too much discretion was given to individual police officers.

Breach of the PSED

The Court of Appeal also allowed Mr Bridge's appeal on the ground that SWP, as a public authority, had failed to discharge its positive duty under the PSED to ensure that AFR Locate did not have a potentially discriminatory impact. The Court of Appeal first considered the argument that the "human failsafe" component in the way in which AFR Locate was used was sufficient to discharge the PSED. This safety feature required the SWP not to take any step against any member of the public unless a positive match made by the automated system was confirmed by two human beings, which included at least one police officer.

Noting that the "human failsafe" went towards the substance of the decision-making, rather than the process that must be followed, the Court of Appeal held that the safety feature was not material to the PSED. Further, it referred to the well-established proposition that human beings could also make mistakes, especially in the context of identification.

The second, more critical, argument concerned divergent expert opinions on whether the training datasets utilised by AFR technology had an inbuilt bias. On the one hand, an independent expert witness testified that AFR systems could suffer from training "bias" due to "any imbalance in the demographic of subjects in the training datasets, resulting in the AFR system having a high false alarm rate or a high false reject rate for that particular demographic".[9] The witness, however, observed that SWP could not assess the discriminatory impact of AFR Locate specifically as it was not privy to the datasets used to train the AFR system (which was proprietary information).

On the other hand, the AFR manufacturer's witness stated that "great care, effort and cost" had been incurred to address issues relating to racial and gender bias.[10] Any impact of gender bias was minimised as the training dataset contained "roughly equal quantities of male and female faces".[11] With regard to racial bias, the training data included "a wide spectrum of different ethnicities" that "ha[d] been collected from sources in regions of the world to ensure a comprehensive and representative mix".[12]

The Court of Appeal recognised that it was not in a position to adjudicate on the differing expert opinions, but held that the fact remained that SWP failed to "satisfy themselves, either directly or by way of independent verification, that the software program in this case [did] not have an unacceptable bias on grounds of race or sex".[13] It also expressed the hope that "as AFR is a novel and controversial technology, all police forces that intend to use it in the future would wish to satisfy themselves that everything reasonable which could be done had been done in order to make sure that the software used does not have a racial or gender bias".[14]

The Ethical Issues Raised by the AFR Case

From an AI ethics perspective, the AFR case raises three interesting issues concerning accountability, safety and bias that merit further examination.

1. Can unguided discretion in the deployment of novel AI systems be justified in certain circumstances?

The lines of accountability in deploying novel AI systems require critical examination. At a fundamental level, accountability is concerned with "[w]ho or what answers to whom or to what?".[15] Accountability may be founded on different normative bases and its boundaries may therefore be delineated differently in different contexts.[16] For example, a law reform report on AI ethics published by the Singapore Academy of Law's Law Reform Committee ("**the LRC Report**") in July 2020 observed that the lines of accountability are more well-defined in the context of certain professional relationships (e.g. banker and customer).[17] The same may not, however, be true where "AI systems are deployed in a manner that exposes them to the wider public".[18] In the latter case, how should the lines of accountability be drawn?

The AFR case suggests that unguided or unchecked discretion in deploying novel AI systems with a wide public impact is not desirable. Even though the Court of Appeal accepted that AFR was not an intrusive technique, it observed that AFR was a novel technology and "involve[d] the capturing of the images and processing of digital information of *a large number of the members of the public*, in which *it is accepted that the vast majority of them will be of no interest whatsoever to the police*" [emphasis added].[19] The Court of Appeal also noted that the captured facial images constituted "sensitive" personal data, which was "processed in an autoemated way".[20]

On the other hand, because the operational details of the deployment of AI systems may be better left to public authorities to assess due to "changing circumstances",[21] it may not be viable for comprehensive normative guidance to be set out, even in internal policies. For instance, overly prescriptive guided discretion may have the unintended effect of limiting the deployment of AI systems in unexpected situations where serious consequences (e.g. harm to the public) may ensue.

Ultimately, we need to re-examine the meaning of "accountability" and consider the extent to which consequentialist arguments should hold sway in situations similar to the AFR case, where "the protection of private rights" has to be balanced against "the public interest in harnessing new technologies to aid the detection and prevention of crime". [22]

2. **How safe is safe enough?**

In AI ethics, the issue of safety has featured most prominently in automated driving. As an academic commentator has observed: "Given that automated driving will not wholly eliminate roadway deaths and injuries – how safe is safe enough – and how should this safety be demonstrated?"[23]

The same questions may be asked with reference to the AFR case. If the "human failsafe" component is not considered safe enough because of the risk of human error, what will meet the requisite standard? Presumably, the "human failsafe" was implemented as a safeguard against machine error. If neither human nor machine authority[24] can minimise the risk of unjustified police interventions, what values should the ethics of safety take into account?

One possible answer may to be adopt a utilitarian calculus, which may suggest that the costs of implementing a "human failsafe" component should not be excessive, if the risks of human error are small. On this view, absolute safety is never possible as it is always a matter of trade-offs. Such a position may, however, not be satisfactory if no laws or ethical principles govern how such trade-offs should be made, and who should make them.

The question of safety may also be complicated by broader policy concerns. For instance, there may well be circumstances where unjustified police interventions (i.e. wrongly stopping to question some members of the public based on faulty identification) are outweighed by the potentially disastrous consequences resulting from a failure to detect a subject of interest with malicious intent in time. Would this trigger the well-known "trolley problem"[25]or even the "ticking-time-bomb" scenario used in debating whether torture is ever justified? Would a utilitarian calculus be the best approach to resolving such issues?

The LRC Report recognised a similar potential ethical dilemma arising where "an AI system would have to perform an act which is unlawful in order to avert causing injury to human beings", for example, where "an autonomous vehicle may have to drive onto an empty pedestrian walkway to avoid colliding with a person".[26] The LRC Report suggested that policymakers ought to "consider whether norms should be prescribed regarding how such scenarios should be resolved, and whether these should reflect any applicable international standards or practices or instead be culture-specific".[27]

3. **Is bias in AI systems avoidable at all?**

It would appear incontrovertible that bias in AI systems should be absolutely outlawed. As noted in the LRC Report, "[a]n AI system should be rational, fair, and not contain biases that are intentionally or unintentionally built into their system which may harm a community of people or an individual".[28] As an illustration, the LRC Report suggested that where a government agency intended to deploy an AI system to assess a citizen's risk of committing certain types of offences, it should "evaluate potential impact on fairness, justice, bias and negative perceptions across affected communities, especially minorities".[29]

Nevertheless, other commentators have suggested that the ethical inquiry may not be so straightforward as "[t]here may be a trade-off between effectiveness of the algorithm and the countering of bias".[30] In addition, even "if certain characteristics like race are ignored or removed, machine learning systems may identify so-called proxies for such characteristics, which also leads to bias".[31]

The AFR case has highlighted unresolved questions on whether any possibility of racial and gender bias could have been avoided by using a more diverse training dataset[32] (notwithstanding the precautions taken by the AFR manufacturer) or through better algorithm design.[33] Moreover, given that an individual may be identified in multiple ways,[34] would race and gender be conclusive of whether bias exists in an AI system? In the context of automated facial analysis tools, it has been suggested that AI systems may need to be tested "intersectionally" to include other individual traits such as skin colour.[35] In the final analysis, avoiding bias in AI systems at all costs may not be practical or feasible.

Conclusion

As the High Court in the AFR case astutely observed: "The algorithms of the law must keep pace with new and emerging technologies".[36] The novelty of the AFR technology deployed by the SWP was an important factor in the Court of Appeal's decision that its use was unlawful, even within an existing legal framework. The analysis in this article underscores that the law may not supply all the answers to difficult questions of AI ethics relating to accountability, safety and bias. As we move ever closer to a *I, Robot* era, lawyers will need to, in this still existing "I, Lawyer" world, embrace AI ethics in their quest to devise algorithms to resolve the legal conundrums posed by AI.

First published in the September 2020 issue of the Singapore Law Gazette

References

1. (2020) EWCA Civ 1058, on appeal from (2019) EWHC 2341 (Admin).
2. (2020) EWCA Civ 1058 at (1).
3. (2020) EWCA Civ 1058 at (85); (2019) EWHC 2341 (Admin) at (25) and (54).
4. (2019) EWHC 2341 (Admin) at (84).
5. *Ibid*.
6. (2020) EWCA Civ 1058 at (123).
7. (2020) EWCA Civ 1058 at (124).
8. (2020) EWCA Civ 1058 at (130).
9. (2020) EWCA Civ 1058 at (193).
10. (2020) EWCA Civ 1058 at (196).
11. (2020) EWCA Civ 1058 at (196).
12. (2020) EWCA Civ 1058 at (196).
13. (2020) EWCA Civ 1058 at (199).
14. (2020) EWCA Civ 1058 at (201).
15. Joshua A. Kroll, 'Accountability in Computer Systems' in Markus D. Dubber, Frank Pasquale and Sunit Das (eds.), *The Oxford Handbook of Ethics of AI* (United States: Oxford University Press, 2020), 181 at 183.
16. *Ibid*.
17. Singapore Academy of Law, Law Reform Committee, "Applying Ethical Principles for Artificial Intelligence in Regulatory Reform" (July 2020) at para. 2.53<https://www.sal.org.sg/sites/default/files/SAL-LawReform-Pdf/2020-09/2020%20Applying%20Ethical%20Principles%20for%20AI%20in%20Regulatory%20Reform_ebook.pdf> (accessed 20 July 2020).
18. *Ibid*.
19. (2020) EWCA Civ 1058 at (87).
20. (2020) EWCA Civ 1058 at (88)-(89).
21. *Supra*, n 15 at 186.
22. (2019) EWHC 2341 (Admin) at (4).
23. Bryant Walker Smith, 'Ethics of Artificial Intelligence in Transport' in Markus D. Dubber, Frank Pasquale and Sunit Das (eds.), *The Oxford Handbook of Ethics of AI* (United States: Oxford University Press, 2020), 669 at 675.
24. *Id*, at 677-79.
25. *Supra*, n 17 at para. 2.10.
26. *Ibid*.
27. *Ibid*.
28. *Id*, at para. 2.13.
29. *Ibid*.
30. Mark Coeckelbergh, *AI Ethics* (United States: The Massachusetts Institute of Technology, 2020), at p. 131.
31. *Ibid*.
32. *Id*, at pp. 135-36.
33. See e.g. Michael Kearns and Aaron Roth, *The Ethical Algorithm: The Science of Socially Aware Algorithm Design* (United States: Oxford University Press, 2020).
34. Timnit Gebru, 'Race and Gender' in Markus D. Dubber, Frank Pasquale and Sunit Das (eds.), *The Oxford Handbook of Ethics of AI* (United States: Oxford University Press, 2020), 253 at 258-59.
35. *Id*, at 258.
36. (2019) EWHC 2341 (Admin) at (1).