4

Transparency of Facial Recognition Technology and Trade Secrets

Rita Matulionyte

4.1 INTRODUCTION

Facial recognition technology (FRT) is being increasingly used by border authorities, law enforcement, and other government institutions around the world. Research shows that among the 100 most populated countries in the world, seven out of ten governments are using FRT on a large-scale basis.[1] One of the major challenges related to this technology is the lack of transparency and explainability surrounding it. Numerous reports have indicated that there is insufficient transparency and explainability around the use of artificial intelligence (AI), including FRT, in the government sector.[2] There are still no clear rules, guidelines, or frameworks as to the level and kind of transparency and explainability that should be expected from government institutions when using AI more generally, and FRT in particular.[3] The EU General Data Protection Regulation (GDPR) is among the first instruments to establish a right of explanation in relation to automated decisions,[4] but its scope is very limited.[5] The proposed EU

This chapter is a result of the project 'Government Use of Facial Recognition Technologies: Legal Challenges and Solutions' (FaceAI), funded by the Research Council of Lithuania (LMTLT), agreement number S-MIP-21-38.

[1] Paul Bischoff, 'Facial recognition technology (FRT): 100 countries analyzed' (8 June 2021), Comparitech, www.comparitech.com/blog/vpn-privacy/facial-recognition-statistics/.

[2] See, e.g., NSW Ombudsman, 'The new machinery of government: Using machine technology in administrative decision-making' (29 November 2021), State of New South Wales, www.ombo.nsw.gov.au/Find-a-publication/publications/reports/state-and-local-government/the-new-machinery-of-government-using-machine-technology-in-administrative-decision-making; European Ombudsman, 'Report on the meeting between European Ombudsman and European Commission representatives' (19 November 2021), www.ombudsman.europa.eu/en/doc/inspection-report/en/149338.

[3] See, e.g., Access Now, 'Europe's approach to artificial intelligence: How AI strategy is evolving' (December 2020), Report Snapshot, www.accessnow.org/cms/assets/uploads/2020/12/Report-Snapshot-Europes-approach-to-AI-How-AI-strategy-is-evolving-1.pdf, p. 3.

[4] Regulation 2016/679 of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L119/1, art 13.

[5] Sandra Wachter, Brent Mittelstadt, and Chris Russell, 'Counterfactual explanations without opening the black box: Automated decisions and the GDPR' (2018) 31 *Harvard Journal of Law & Technology* 841–847, at 842, 878, 879 ('a legally binding right to explanation does not exist in the GDPR').

Artificial Intelligence Act (Draft EU AI Act) sets minimum transparency standards to high-risk AI technologies that include FRT.[6] However, these transparency obligations are generic to all high-risk AI technologies and do not detail transparency requirements for FRT specifically.

Transparency and explainability are arguably essential to ensuring the accountability of government institutions using FRT; empowering supervisory authorities to detect, investigate, and punish breaches of laws or fundamental rights obligations; allowing individuals affected by an AI system's outcome to challenge the decision generated using AI systems;[7] and enabling AI developers to evaluate the quality of the AI system.[8] According to the proposed EU AI Act, 'transparency is particularly important to avoid adverse impacts, retain public trust and ensure accountability and effective redress'.[9]

At the same time, one should note that transparency and explainability of FRT alone would not help remedy essential problems associated with FRT use, and might further contribute to its negative impacts in some cases. For instance, if an individual learns about the government use of FRT in public spaces where public gatherings take place, this might discourage her from participating in such gatherings and thus have a 'chilling effect' on the exercise of her human rights, such as freedom of speech and freedom of association.[10] These considerations have to be kept in mind when determining the desirable levels of FRT transparency and explainability.

While there is extensive *technical* literature on transparency and explainability of AI in general,[11] and of FRT more specifically,[12] there is very limited *legal* academic discussion

[6] See European Commission, 'Proposal for a Regulation of the European Parliament and of the Council: Laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts' (21 April 2021) (hereafter draft EU AI Act), Com 206 Final, articles 13(1), 20, 60, 62.

[7] See, e.g., OECD, 'Transparency and explainability (Principle 1.3)' (2022), OECD AI Principles, https://oecd.ai/en/dashboards/ai-principles/P7.

[8] See, e.g., Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardozo, 'Machine learning interpretability: A survey on methods and metrics' (2019) 8(8) *Electronics* 832, 5–7; Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal, 'Explaining explanations: An overview of interpretability of machine learning' (3 February 2019), Working Paper, https://arxiv.org/abs/1806.00069.

[9] Draft EU AI Act, para. 38.

[10] Interview participant 2, NGO representative.

[11] See, e.g., Upol Ehsan, Q. Vera Liao, Michael Muller, Mark O. Riedl, and Justin D. Weisz, 'Expanding explainability: Towards social transparency in AI systems' (May 2021), *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Article No. 82, pp. 1–19; Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera, 'Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI' (2020) 58 (June) *Information Fusion* 82–115.

[12] Jonathan R. Williford, Brandon B. May, and Jeffrey Byrne, 'Explainable face recognition', *Proceedings of Computer Vision – ECCV: 16th European Conference*, Glasgow, UK (23–28 August 2020), Part XI, pp. 248–263; Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller (eds.), *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (Springer International Publishing, 2019).

about the requisite extent of transparency and explainability of FRT technologies, and challenges in ensuring it, such as trade secrets. The goal of this chapter is to examine to what extent trade secrets create a barrier in ensuring transparent and explainable FRT and whether current trade secret laws provide any solutions to this problem.

This chapter first identifies the extent to which transparency and explainability is needed in relation to FRT among different stakeholders. Second, after briefly examining which types of information about AI could be potentially protected as trade secrets, it identifies situations in which trade secret protection may inhibit transparent and explainable FRT. It then analyses whether the current trade secret law, in particular the 'public interest' exception, is capable of addressing the conflict between the proprietary interests of trade secret owners and AI transparency needs of certain stakeholders. This chapter focusses on FRT in law enforcement, with a greater emphasis on real-time biometric identification technologies that are considered the highest risk.[13]

Apart from the critical literature analysis, this chapter relies on empirical data collected through thirty-two interviews with experts in AI technology. The interviews were conducted with representatives from five stakeholder groups: police officers, government representatives, non-governmental organisation (NGO) representatives, IT experts (in academia and private sector), and legal experts (in academia and private sector) from Europe, the United States, and Asia-Pacific (October 2021–March 2022, online). The data collected from these interviews is especially useful when identifying the transparency and explainability needs of different stakeholders (Section 4.2).

Keeping in mind the lack of consensus on the terms 'AI transparency' and 'AI explainability', for the purpose of this chapter we define the concepts as follows. First, we understand the 'AI transparency' principle as a requirement to provide information *about* the AI model, its algorithm, and its data. The AI transparency principle could require disclosing very general information, such as 'when AI is being used',[14] or more specific information about the AI module – for example, its algorithmic parameters, training, validation, and testing information. While this concept of transparency might require providing very different levels of information for different stakeholders, it does not include information about *how* AI decisions are being generated. The latter is covered by the principle of 'AI explainability', which we define in a narrow technical way; that is, as an explanation of *how* an AI module functions, and how it generates a particular output. Such explanations are normally provided using so called Explainable AI (XAI) techniques.[15] Generally speaking, XAI techniques might be 'global', explaining the features of the entire

[13] See, e.g., draft EU AI Act, arts 5, 21, 26.

[14] Such as in OECD, 'Transparency and explainability'; Australian Government, 'Australia's artificial intelligence ethics framework' (7 November 2019), Department of Industry, Science and Resources, www.industry.gov.au/data-and-publications/australias-artificial-intelligence-ethics-framework.

[15] Shane T. Mueller, Robert R. Hoffman, William Clancey, Abigail Emrey, Gary Klein, 'Explanation in human-AI systems: A literature meta-review synopsis of key ideas and publications and bibliography for explainable AI' (5 February 2019), DARPA XAI Literature Review, arXiv:1902.01876; Maja Brkan

module; or 'local', which explain how a specific output has been generated.[16] While this chapter largely focusses on FRT transparency and its possible conflict with trade secret protection, it also briefly reflects upon the need for FRT to be explainable.

In the following sections, we discuss the scope of explainability and transparency that different stakeholders need in relation to FRT in law enforcement (Section 4.2), in which situations trade secrets may conflict with these transparency and explainability needs (Section 4.3), and whether the 'public interest' defence under trade secrets law is capable of addressing this conflict (Section 4.4).

4.2 FRT TRANSPARENCY AND EXPLAINABILITY: WHO NEEDS IT AND HOW MUCH?

Before examining whether trade secrets conflict with FRT transparency and explainability principles, we need to clearly identify the level of transparency and explainability that different stakeholders require in relation to FRT. We demonstrate that different stakeholders need very different types of information, some of which is – and some is not – protected by trade secrets.

For the purpose of this analysis, we identified six categories of stakeholders who have legitimate interests in certain levels of transparency and/or explainability around FRT technologies: (1) individuals exposed to FRT; (2) police officers who directly use the technology; (3) police authorities that acquire/procure the technology and need to ensure its quality; (4) court participants, especially court experts, who need access to technical information to assess whether the technology is of sufficient quality; (5) certification and auditing bodies examining whether the FRT meets the required standards; and finally (6) public interest organisations (NGOs and public research institutions) whose purpose is to ensure, in general terms, that the technology is high quality, ethical, legal, and is used for the overall public benefit.

As could be expected, our interviews with stakeholders have shown that different stakeholders have different explainability and transparency needs in relation to FRT.

4.2.1 *FRT Explainability*

In terms of the explainability of FRT, few stakeholders need it as a matter of necessity. Among the identified stakeholder groups, certification and auditing bodies that examine the quality of technology might potentially find XAI techniques useful – as

and Gregory Bonnet, 'Legal and technical feasibility of the GDPR's quest for explanation of algorithmic decisions: Of black boxes, white boxes and fata morganas' (2020) 11(1) *European Journal of Risk Regulation* 18–50, at 18–19.

[16] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi, 'A survey of methods for explaining black box models' (2019) 51(5) *ACM Computing Surveys* 1–42.

https://doi.org/10.1017/9781009321211.006 Published online by Cambridge University Press

these may help identify whether, for instance, a specific AI module is biased or contains errors.[17] For similar reasons, XAI techniques might be relied upon by public interest organisations, such as NGOs and research institutions, that have expertise in AI technologies and want to assess the quality of a specific FRT technology used by police. AI developers themselves have been using XAI techniques for a similar purpose; that is, to identify AI errors during the development process and eliminate them before deploying them in practice.[18] However, XAI techniques themselves do not currently have quality guarantees and often face issues as to quality and reliability.[19] It is thus questionable whether experts assessing the quality of AI, or FRT more specifically, would give much weigh to such explanations.

Other stakeholders – police authorities, police officers, and affected individuals – are unlikely to find explanations generated by XAI techniques useful, mainly because of the technical knowledge that is required to understand such explanations. Further, according to some interviewees, when FRT is used for identification purposes, users do not need an explanation at all as the match made by FRT could be easily double checked by a police officer.[20]

Importantly, explanations generated by XAI techniques are unlikely to interfere with trade secret protection as they do not disclose substantial amounts of confidential information. As discussed later, in order to be protected by trade secrets, information should be of independent commercial value and kept secret.[21] XAI techniques, if integrated in the FRT system, would provide explanations to the end users, which, by their nature, would not be secret. Thus, owing to its limited relevance for our debate on FRT and trade secrets, FRT explainability will not be analysed here any further.

4.2.2 *FRT Transparency Needs*

In contrast, transparency around FRT is required by all stakeholders, although to differing extents. Depending on the level of transparency/information needed, stakeholders could be divided into three groups: those with (1) relatively low transparency needs, (2) high transparency needs, and (3) varying/medium transparency needs.

[17] Interview participant 1, IT expert.

[18] Ibid.

[19] See, e.g., Zana Buçinca, Krzysztof Z. Gajos, Phoebe Lin, and Elena L. Glassman, 'Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems' (2020), Proceedings of the 25th international conference on intelligent user interfaces, https://dl-acm-org.simsrad.net.ocs.mq.edu.au/doi/abs/10.1145/3377325.3377498; Julius Adebayo et al., 'Sanity checks for saliency maps' (2018) 31 *Advances in Neural Information Processing Systems* 9505, arXiv:1810.03292v3; Jindong Gu and Volker Tresp, 'Saliency methods for explaining adversarial attacks' (October 2019), Human-Centric Machine Learning (NeurIPS Workshop), https://arxiv.org/abs/1908.08413; similar from Interview participant 5, IT expert ('it's not clear to me if we'll ever come up with a particularly good explanation of how the combination of neural networks and all the technologies that go into face recognition work. Whether we'll ever be able to explain them').

[20] Interview participant 13, NGO representative.

[21] See Section 4.3.

https://doi.org/10.1017/9781009321211.006 Published online by Cambridge University Press

4.2.2.1 Low Transparency Needs

Individuals exposed to FRT, and law enforcement officers directly using the technology, require relatively general non-technical information about FRT (thus 'low transparency'). Individuals have a legitimate interest in knowing where, when and for what purpose the technology is used; its accuracy levels and effectiveness; legal safeguards put around the use of this technology; and in which circumstances and how they can complain about inappropriate or illegal use of FRT.[22] After individuals have been exposed to the technology and if this has led to adverse effects (e.g., potential violation of their rights), they might require a more detailed *ex post* explanation as to why a specific decision (e.g., to stop and question the individual) was made and how FRT was used in this context. Still, they do not need any detailed technical explanations about how the technology was developed, trained, or how exactly it functions, as they do not have the technical knowledge required for the interpretation of this information.

As one of our interviewees explained (in the context of migration/border control):

> So, for example, if I am a citizen stakeholder [and] my application for a visa is denied and it's based on my looks [that suggests that I] have some criminal records, then, of course, it has impacted me and I'm not happy, and I will ask for answers. Even [if the] activities [were] rectified, still [I'll ask for] answers on how come did you make this mistake? Why did you take me wrong [as] another person and it cost me my travel to be cancelled? So, to have explainability at this level, potentially you don't need to explain all of the algorithms. It's a matter of explaining why this sort of decision was made. For example, there was this person with similar facial features and the same name; or whatever some high-level explanation of what happened in the process that explains why mistake happened, etc.[23]

Second, police officers who directly use the technology will want access to general information about how the system functions, what types of data were used to train the system, the accuracy rates in different settings, how it should be used, its limitations, and so on.[24]

In addition, these stakeholders would benefit from user-friendly explanations about, for instance, which pictures in the watch-list were found to be sufficiently similar to the probe picture and the accuracy rate with relation to that specific match.[25] This would allow police officers to assess the extent to which they could

[22] This type of information is being currently provided, for example, on the UK Metropolitan police website: www.met.police.uk/advice/advice-and-information/fr/facial-recognition.

[23] Interview participant 1, IT expert.

[24] Interview participant 19, law enforcement officer.

[25] Watch list is the list against which the taken image is compared. When FRT is used in law enforcement context, the watch list normally comprises images of persons who are suspected or convicted for crimes, missing persons, etc. In case of a live FRT, the probe picture is a picture taken from the passing individual.

rely on a specific FRT outcome before proceeding with an action (e.g., stopping an individual for questioning or arrest). Information needs might differ between real-time/live FRT and post FRT (i.e., when FRT is used to find a match for a picture taken some time ago), as the former is considered higher risk.[26]

4.2.2.2 High Transparency Needs

Stakeholder groups that are required to assess the quality of a FRT system – certification and auditing authorities, and court experts – have high transparency needs. In order to conduct an expert examination of FRT technology, certification and auditing bodies require access to detailed technical information about the system. This might include algorithmic parameters, training data, processes and methods, validation/verification data and processes, as well as testing procedures and outcomes.

As one of the interviewed IT experts explained:

> But if, for example, there is an audit happening. […] then of course, at that level explainability means something completely different. It's about explaining how the system was designed, how it was being used, what sort of algorithms, what sort of data was used for the training, what sort of design and build decisions were made, and so on.[27]

Similar highly technical information could be demanded in court proceedings by court experts who are invited to assess the quality of FRT used by law enforcement authorities during legal proceedings. Detailed technical information would be necessary to provide technically sound conclusions.

4.2.2.3 Medium/Varying Transparency Needs

The third group of stakeholders might have varied information needs depending on their level of knowledge about AI technologies. Namely, law enforcement authorities, when acquiring the FRT system, would need information that allows them to judge the quality and reliability of the FRT system in question. If they have only general knowledge about FRT, they will merely want to know whether the technology meets the industry standards and whether it was certified/validated by independent bodies;[28] how accurate it is; whether it has been trialled in real life settings, the trial results, and so on. If they have expert knowledge in AI/FRT (e.g., in their IT team), they might demand more technical information, for example, about datasets on which it was trained and validated, and validation and testing information.

[26] For example, the draft EU AI Act treats live FRT in the law enforcement context as extremely high risk and generally bans them, with a few exceptions: see draft EU AI Act, Annex 3.

[27] Interview participant 1, IT expert.

[28] The draft EU AI Act requires all high-risk AI technologies, including FRT, to undergo certification procedures. This requirement, however, has not yet been established in other jurisdictions.

As a final stakeholder group, public interest organisations (researchers and NGOs) have a legitimate interest in accessing information about government FRT use as 'they are the ones that are most likely to initiate […] strategic litigation and other initiatives',[29] and ensure that government is accountable for the use of this technology.[30] Similarly to law enforcement, their transparency needs will differ depending on their expertise and purpose. Those without expert knowledge in AI might be interested in general information as to which situations and purposes, and to what extent, law enforcement is using FRT; the accuracy levels and effectiveness of the technology in achieving the intended aims (e.g., whether the use of FRT led to the arrest of suspected persons or preventing a crime); and whether there have been human rights impact assessments conducted at the procurement level and their results.[31] Those with technical expertise in AI might want access to algorithmic parameters and weights, training and validation/verification data, or similar technical information, allowing them to assess the accuracy and possible bias of the technology (similar to the high level transparency discussed earlier).[32]

These three levels of transparency are relevant when determining the situations in which trade secret protection might become a barrier to ensuring the transparency demanded by stakeholders.

4.3 IN WHICH SITUATIONS MIGHT TRADE SECRETS INHIBIT TRANSPARENCY OF FRT?

There are a number of challenges in ensuring transparency around FRT.[33] One of them is trade secrets, which can arguably create barriers to ensuring transparency of AI technologies in general and FRT technologies in particular. The example often used is the *State* v. *Loomis* case decided by a US court, in which the defendant was denied access to the parameters of the risk assessment algorithm COMPAS owing to trade secrets.[34] In this section, we demonstrate that the answer is more nuanced: while trade secrets might create barriers to transparent FRT in some situations ('actual conflict' situations), they are unlikely to interfere with transparency needs in other situations ('no conflict' and 'nominal conflict' situations).

[29] Interview participant 21, legal expert.

[30] Interview participant 5, IT expert ('Particularly, I mean, transparency is a very useful means of regulating governments abusing their position'); similar from interview participant 2, NGO representative.

[31] Interview participant 13, NGO representative.

[32] Interview participant 2, NGO representative ('for us in civil society, knowing the parameters that were set around accuracy and the impact that might have on people of colour, might be a useful thing to know, contest the use case').

[33] Another possible challenge is government secrets (the government may not want to disclose certain information for public security reasons, for example). The challenge in ensuring FRT explainability is technical (technical ability to provide explanations of how a specific AI functions).

[34] *State* v. *Loomis* 881 N.W.2d 749, 755, 756, fn.18 (Wis. 2016), cert. denied, 137 *S. Ct.* 2290 (2017).

4.3.1 *The Scope of Trade Secret Protection*

In order to understand the situations in which trade secrets interfere with transparency needs around FRT, it is first necessary to clarify which information about FRT could be potentially protected by trade secrets.

Trade secrets are of special importance in protecting intellectual property (IP) rights underlying AI modules, including FRT. In contrast to other IP rights (patents, copyright), trade secrets could be used to protect any elements of AI modules as long as they provide independent commercial value and are kept secret.[35] Trade secret protection requires neither investment in the registration process nor public disclosure of the innovation.[36] While trade secret protection has its limitations, such as a possibility to reverse engineer technology protected by trade secrets,[37] and a lack of protection against third-party disclosure,[38] the software industry has so far successfully used trade secrets to protect its commercial interests.[39]

As far as trade secrets and AI are concerned, courts have already indicated that at least certain parts of AI modules can be protected as trade secrets, such as source code, algorithms, and the way a business utilises AI to implement a particular solution.[40] Keeping in mind the requirements for trade secret protection – secret nature and commercial value – a range of information about AI (including FRT) could be possibly protected by trade secrets: the architecture of the algorithm, its parameters and weights; source code in which the algorithm is coded; information about the training, validation and verification of the algorithm, including training and validation/verification data, methods and processes; real life testing information (in which settings it was tested, and the methods and outcomes of testing), and so on. All this

[35] See Tanya Aplin, Lionel Bently, Phillip Johnson, and Simon Malynicz, *Gurry on Breach of Confidence: The Protection of Confidential Information* (2nd ed., Oxford University Press, 2012).

[36] See, e.g., Clark D. Asay, 'Artificial stupidity' (2020) 61(5) *William and Mary Law Review* 1187–1257, 1243, Notably, significant financial costs might be incurred to ensure that information maintains secret.

[37] For more, see Tanya Aplin, 'Reverse engineering and commercial secrets' (2013) 66(1) *Current Legal Problems* 341–377.

[38] Katarina Foss-Solbrekk, 'Three routes to protecting AI systems and their algorithms under IP law: The good, the bad and the ugly' (2021) 16(3) *Journal of Intellectual Property Law & Practice* 247–258; Ana Nordberg, 'Trade secrets, big data and artificial intelligence innovation: A legal oxymoron?' in Jens Schovsbo, Timo Minssen, and Thomas Riis (eds.), *The Harmonization and Protection of Trade Secrets in the EU: An Appraisal of the EU Directive*(Edward Elgar Publishing Limited, 2020), pp. 194–220, at p. 212.

[39] See, e.g., Sylvia Lu, 'Algorithmic opacity, private accountability, and corporate social disclosure in the age of artificial intelligence' (2020) 23(99) *Vanderbilt Journal of Entertainment & Technology Law* 116–117 (contending that software industry has relied on trade secret law to protect algorithms for decades and AI algorithms are no exception).

[40] See, e.g., *LivePerson, Inc. v. 24/7 Customer, Inc.*, 83 F. Supp. 3d 501, 514 (SDNY, 2015) (finding algorithms based on artificial intelligence eligible for trade secret protection).

information is often seen by AI developers as of commercial value and kept secret,[41] and thus could be potentially protected as trade secrets.[42]

4.3.2 *When is the Conflict between Trade Secrets and the AI Transparency Principle Likely to Arise?*

Keeping in mind the broad range of information about the FRT that could be protected as trade secrets and the transparency needs of stakeholders (identified earlier), three types of situations could be distinguished.

4.3.2.1 No Conflict Situations

First, in some situations, there would be no conflict between stakeholder's transparency needs and trade secret protection as the information requested by the stakeholder is generally not protected by trade secrets. For instance, individuals subject to FRT would only want general information about the fact that FRT is used by a government authority, where and for what purposes it is used, and so on.[43] Similarly, police officers using the technology would only need a general understanding of how the technology functions, in which situations it could be used, its accuracy rates, and so on.[44] Owing to its generally public nature and lack of independent economic value, this information would normally not be protected as trade secrets.

4.3.2.2 Nominal Conflicts

In some other instances, 'nominal' conflict situations are likely to arise. First, certification and auditing organisations that are examining the quality of FRT technologies might require access to extensive technical information related to FRT that has commercial value and could be protected by trade secrets, such as algorithmic parameters, training, validation and verification information, and all information related to real-life trials.[45] Similar information might be requested in court proceedings by court experts who are invited to assess the reliability of the FRT system in question.[46] As discussed earlier, these types of technical information are likely to be

[41] Interview participant 1, IT expert.

[42] Note that even if all of this information could be 'factual' trade secrets, not all of it would qualify as 'legal' trade secrets. For a distinction between the two see Sharon K. Sandeen and Tanya Aplin, 'Trade secrecy, factual secrecy and the hype surrounding AI' in Ryan Abott (ed.), *Research Handbook on Intellectual Property and Artificial Intelligence* (Edward Elgar, 2022), pp. 442–450; see also Camilla A. Hrdy and Mark A. Lemley, 'Abandoning trade secrets' (2021) 73(1) *Stanford Law Review* 1–66.

[43] See Section 4.2.2.1. While this information could be protected as government secrets, it would not be protected as a trade secret as it does not have independent commercial value.

[44] See Section 4.2.2.1.

[45] See Section 4.2.2.2.

[46] See Section 4.2.2.2.

https://doi.org/10.1017/9781009321211.006 Published online by Cambridge University Press

protected as trade secrets: AI developers consider them commercially valuable and tend to keep them secret.[47]

However, we refer to these types of situations as 'nominal' conflicts since they could be managed under existing confidentiality/trade secret rules that form part of certification/auditing processes or court procedures. Certification and auditing organisations are normally subject to confidentiality and use the confidential information provided by AI developers for assessment purposes only. Similarly, in court investigations, procedural rules determine how trade secrets disclosed during the court proceedings are protected from disclosure to third parties or to the public.[48] Since these situations are already addressed under current regulatory or governance frameworks, we will not examine them further.

4.3.2.3 Actual Conflicts

The third type of situations – related to transparency needs of law enforcement authorities and public interest organisations – are of most concern, and we refer to them as 'actual conflicts'.

Law enforcement authorities might need access to certain technical information about the FRT (e.g., training, validation and testing information) in order to evaluate its reliability before procuring it.[49] Public interest organisations, such as NGOs and research organisations, might need access to even more detailed technical information (algorithms, training and validation data, testing data) in order to provide an independent evaluation of the effectiveness of the FRT system used by law enforcement.[50] As mentioned earlier, technical information is generally considered by AI developers as commercially valuable and is likely to be kept confidential.

It is worth noting that law enforcement authorities are able to obtain certain information through contract negotiation.[51] However, it is questionable whether this solution is suitable in all cases. Owing to a lack of adequate legal advice, bargaining power, or simply the novel nature of AI technologies, law enforcement authorities might fail to negotiate for appropriate access to all essential information that will be needed during the entire life cycle of the FRT system. Government authorities using AI tools acquired from third parties have already encountered the problem

[47] See Section 4.3.1.

[48] For example, Court Suppression and Non-Publication Orders Act 2010 (NSW) s 9 allows a court to make a suppression or non-publication order if it is necessary to prevent prejudice to the proper administration of justice.

[49] See Section 4.2.2.3.

[50] See Section 4.2.2.3.

[51] Similar has been suggested for AI acquisition process for government institutions: see Jake Goldenfein, 'Algorithmic transparency and decision-making accountability: Thoughts for buying machine learning algorithms' in *Closer to the Machine: Technical, Social, and Legal Aspects of AI* (Office of the Victorian Information Commissioner, 2019), https://ovic.vic.gov.au/wp-content/uploads/2019/08/closer-to-the-machine-web.pdf.

of subsequently getting access to certain confidential information about the AI module.[52]

Similarly, while public interest organisations might acquire certain information about FRT used by government through freedom of information requests,[53] this solution is limited as the legislation generally protects trade secrets from public disclosure.[54] Therefore, we see both of these situations as an *actual conflict* between trade secret rights of AI developers and the AI transparency needs of two major groups of stakeholders (law enforcement authorities and public interest organisations).

4.4 DOES TRADE SECRET LAW PROVIDE ADEQUATE SOLUTIONS?

Trade secret law provides certain limitations that are meant to serve the interests of the public. Namely, in common law jurisdictions, when a breach of confidentiality is claimed, the defendant could raise a so-called public interest defence. In short, it allows defendants to avoid liability for disclosing a trade secret if they can prove the disclosure was in the public interest.[55] As explained by the House of Lords, protection of confidential information is based on the public interest in maintaining confidences, but the public interest sometimes favours disclosure rather than secrecy.[56] However, this public interest defence is of limited, if any, use in addressing the conflict between trade secrets and the legitimate transparency needs of identified stakeholders in an FRT scenario.

First, the scope of this defence is unclear.[57] Some judicial sources suggest the existence of a broad public interest defence, which is based upon freedom of the press and the public's right to know the truth.[58] Other court judgments suggest that the defence should encompass no more than an application of the general equitable defence of clean hands, namely the information that exposes a serious wrongdoing of the plaintiff should not be classified as confidential in any case (iniquity rule).[59]

[52] Interview participant 26, government representative.

[53] See, e.g., *Freedom of Information Act 1982* (Cth).

[54] See Elizabeth A. Rowe, 'Striking a balance: When should trade-secret law shield disclosure to the government?' 96 *Iowa Law Review* 791–835, at 804–808.

[55] For an overview of the public interest defence, see Aplin et al., *Gurry on Breach of Confidence*.

[56] *Attorney-General* v. *Guardian Newspapers Ltd* [1990] AC 109, 282 (Spycatcher case) ('although the basis of the law's protection of confidence is a public interest that confidences should be preserved by law, nevertheless that public interest may be outweighed by some other countervailing public interest which favours disclosure' (Lord Goff)) Similarly, in *Campbell v. Frisbee*, the UK Court of Appeal held that the confider's right 'must give way where it is in the public interest that the confidential information should be made public'. See *Campbell* v. *Frisbee* [2002] EWCA Civ 1374, [23].

[57] See Karen Koomen, 'Breach of confidence and the public interest defence: Is it in the public interest? A review of the English public interest defence and the options for Australia' (1994) 10 *Queensland University of Technology Law Journal* 56–88.

[58] See, e.g., Spycatcher case, 269 (Lord Griffiths); *Fraser* v. *Evans* [1969] 1 QB 349; *Hubbard* v. *Vosper* [1972] 2 QB 84; discussed in Trent Glover, 'The scope of the public interest defence in actions for breach of confidence' (1999) 6 *James Cook University Law Review* 109–137, at 115–116, 118.

[59] See discussion in Glover, 'The scope of the public interest defence'; *Corrs Pavey Whiting & Byrne* v. *Collector of Customs (Vic)* (1987) 14 FCR 434, 454 (Gummow J).

For instance, Australian courts have confirmed that disclosure in the public interest should be construed narrowly; it should limited to information affecting national security, concerning breach of law, fraud, or otherwise destructive to the public, and must be more than simply the public's interest in the truth being told.[60]

Most importantly, the defence does not provide interested stakeholders with an active right to request information about the FRT technology and its parameters. It is merely a passive defence that could be invoked by a defendant only after they have disclosed the information (or where there is an imminent threat of such a disclosure). In order to disclose the information, the defendant should already have access to the information, which is not the situation of law enforcement authorities or public interest organisations seeking information about the FRT.

The public interest defence could be possibly useful in some exceptional situations. For instance, the employee/contractor of an FRT developer might disclose certain confidential technical information about the FRT system with the public or a specific stakeholder (public authority, NGO, etc.) in order to demonstrate that the AI developer did not comply with legal requirements when developing the FRT system and/or misled the public and/or the government authority as to the accuracy of the FRT technology, for example. If breach of confidence is claimed against this person, they could argue that the disclosure served the public interest: the use of an FRT system that is of low quality or biased may lead to incorrect identification of individuals, especially ethnic or gender minorities, which may further result in the arrest of innocent people and violation of their human rights. The defendant could argue that the disclosure of technical information about such an FRT system would thus help prevent harm from occurring.

Even then, the ability of a defendant to rely on the public interest defence is questionable. For instance, the court might accept the defence if the information is disclosed to government authorities responsible for prosecuting breaches of law or fraud, as 'proper authorities' for public disclosure purposes,[61] but not to public interest organisations or the public generally.[62] While the law enforcement authority (which is also the user of FRT in this case) might qualify as a 'proper authority', a public interest organisation is unlikely to meet this criterion.

Furthermore, if a narrow interpretation of the public interest defence is applied, the defendant would have to prove that the disclosed information relates to 'misdeeds of a serious nature and importance to the country'.[63] It is questionable

[60] *Castrol Australia Pty Limited* v. *Emtech Associates Pty Ltd* (1980) 51 FLR 184, 513 (Rath J, quoting with approval Ungoed-Tomas J in *Beloff* v. *Pressdram* [1973] 1 All ER 241, 260); for a criticism of a narrow interpretation see Koomen, Breach of confidence and the public interest defence.

[61] See discussion in Jason Pizer, 'The public interest exception to the breach of confidence action: are the lights about to change?' (1994) 20(1) *Monash University Law Review* 67–109, at 80–81.

[62] See, e.g., *Francome* v. *Mirror Group Newspapers Ltd* [1984] 2 All ER 408.

[63] *Beloff* v. *Pressdram*, 260; see similar limitation in *Corrs Pavey Whiting & Byrne* v. *Collector of Customs*, 456 (Gummow J).

whether a low quality or biased FRT, or the AI developer hiding information about this, would qualify as a misdeed of such serious nature. More problematically, the defendant might not know whether the FRT does not meet certain industry or legal standards *until* the technical information is disclosed and an independent examination is carried out.

4.5 CONCLUSIONS

It is without doubt that transparency is needed around the development, functioning, and use of FRT in the law enforcement sector. The analysis here has shown that in some cases trade secrets do not impede the transparency around FRT needed by some stakeholders (e.g., affected individuals or direct users of FRT) and some possible conflicts could be resolved through existing arrangements and laws (e.g., with relation to the transparency needs of certification and auditing organisations, and court participants). However, trade secrets might conflict with the transparency needs of some stakeholders, especially law enforcement authorities (after acquiring the technology) and public interest organisations that might want access to confidential technical information to assess the quality of the FRT system. Unfortunately, trade secret law, with its unclear and limited public interest exception, is unable to address this conflict. Further research is needed as to how the balance between the proprietary interests of AI developers and transparency needs of other stakeholders (law enforcement authorities and public interest organisations) could be established.

https://doi.org/10.1017/9781009321211.006 Published online by Cambridge University Press