5

Privacy's Loose Grip on Facial Recognition

Law and the Operational Image

Jake Goldenfein

5.1 INTRODUCTION

'Privacy' has long been central to understanding the impacts of facial recognition and related technologies. Privacy informs the intuitions, harms, and legal regimes that frame these technological systems. Privacy and data protection law already have a ready-at-hand toolkit for related practices such as closed-circuit television (CCTV) in public space, surreptitious photography, and biometric data processing. These regimes measure facial recognition applications against familiar privacy and data protection categories such as proportionality, necessity, and legality, as well as identifiability and consent. But as facial recognition becomes more widespread and diverse, and the tools, ecosystems, and supply chains for facial recognition become more visible and better understood, these privacy and data protection concepts are becoming more difficult to consistently apply.

For as long as privacy has been deployed to constrain facial recognition, analysts have been decrying its inadequacy. This research typically identifies some novel dimension of harm associated with facial recognition that evades existing regulatory strategies. This chapter proposes an alternate diagnosis for why privacy fails to deliver premised on the nature of facial recognition as a broader socio-technical system. The jurisprudence shows that privacy and data protection function as intended at the level of 'applications' such as one-to-one and one-to-many identification and identity verification systems. But emerging cases show how privacy concepts become awkward and even incoherent when addressing different dimensions of the facial recognition ecosystem – at the level of 'tools' and supply chains, such as biometric image search engines and the production of facial image datasets. Inconsistencies in how law connects to this part of the facial recognition ecosystem challenge the suitability of regulatory concepts like identifiability and consent, the nature of harm being addressed, and perhaps most fundamentally, how privacy conceptualises the nature of online images. New rules for facial recognition products and applications are being included in the in the risk-based regulatory regimes for artificial intelligence (AI) in development around the world. In the EU, these include prohibitions on *untargeted* scraping of facial images from

https://doi.org/10.1017/9781009321211.007 Published online by Cambridge University Press

the internet or CCTV footage to create facial recognition databases. But as described below, the industrial organisation of the facial dataset business will continue to thwart these regulatory efforts, and privacy and data protection will continue to be legal bases for litigation against companies using facial recognition today and in the future.

This chapter offers an account as to why privacy concepts lose traction in this arena. It argues that existing regulatory approaches reflect an understanding of images as primarily 'representational', whereas facial recognition demonstrates that online images are better understood as 'operational' or 'operative'. The operational image does not simply represent a referent but actively enables and participates in a sequence of automated operations. These operations take place at the level of facial recognition supply chains, where existing law struggles to find traction. Law's inability to come to terms with the operational image pushes existing legal categories to the limits of their utility.

5.2 FACIAL ANALYSIS AND IDENTIFICATION

Privacy law and emerging AI regulations have effectively addressed the 'watch list' type facial recognition applications that come up in human rights litigation. For instance, the 2020 *Bridges* v. *South Wales Police* decision found the South Wales Police (SWP) force's use of facial recognition in public to identify individuals on a watch list was a violation of Article 8 of the European Convention on Human Rights (ECHR).[1] SWP deployed their surveillance system at large public events, using CCTV towers that collected footage of individuals in public, and performed real-time facial recognition against a database of persons of interest. Despite legislation allowing for the creation of that watchlist, the exact parameters for inclusion were not clear. The practice violated the ECHR Article 8 because, while proportionate and strictly necessary for the law enforcement purpose for which it was deployed, it failed to be 'in accordance with the law' in certain respects. Specifically, the enabling legislation and applicable Codes of Conduct failed to adequately specify rules around who could be the subject of surveillance (i.e., who could be placed on a watch list in the first place), or where facial recognition systems could be deployed. The enabling law thus gave police too much discretion. These issues have also clearly informed the regulation of biometric identification by law enforcement in the EU AI Act.

But the Bridges case also highlighted some conceptual issues of interest to the argument made in this chapter. In particular, the court's conceptualisation of facial recognition as something different from both (1) police taking photographs of people in public and (2) the collection of biometric data such as fingerprints.[2] Facial recognition occupied a place somewhere between the two in terms of level of intrusion, generating some conceptual discomfort for privacy. And while this was ultimately

[1] *Bridges* v. *South Wales Police* [2019] EWHC 2341 (admin).

[2] Ibid., at [85], citing *S and Marper* v. *UK* [2018] Eur Court HR 1581 and *Catt* v. *UK* (European Court of Human Rights, Application no. 43514/15, 24 January 2019).

https://doi.org/10.1017/9781009321211.007 Published online by Cambridge University Press

of little consequence to the court's decision, with facial recognition easily enough absorbed into a human rights proportionality analysis without having to delve deeper into facial recognition's 'in-between' character, the inability to analogise with existing police techniques for this in-betweenness was not merely a matter of novelty. This type of watch list surveillance and associated photography, including real-time (non-automated) identification, has been practised by police for decades. But facial recognition's in-between character reflected something more fundamental about the media system that automates the identification task – its operationalism.

The argument made here is that facial recognition and related techniques are a function of the operational image.[3] The central insight of operationalism is that the ontology of images has shifted from one of representation to that of an element in a sequence of operations that are typically machine executed. Mark Andrejevic and Zala Volcic, for instance, describe the 'operational enclosure' through which the operational image includes automated identification, social sorting, decision-making, and responses that enable the governance of space.[4] Their basic example is facial recognition in retail stores that, when identifying a person on a watch list, not only calls security, but also actively locks the doors. This example also exemplifies Trevor Paglen's emphasis that the audience for (operational) images is no longer humans but rather machines.[5]

The operational image reconfigures images as the communicative instruments of automated non-human visuality. Images consumed by humans are increasingly the output of machines staging what they 'see' as a derivative function. But the primary audience of an image is a complex network of machines, with human-legibility a trivial or arbitrary secondary process. As Andrejevic and Volcic note, 'In the case of facial recognition technology, there is, still, a camera with a lens, but for the purposes of recognition and response no image need be produced.'[6] The operational function of an image in the facial recognition context is, on the one hand, its capacity to communicate biometric information to other machines, which can then trigger various actions as described by Andrejevic and Volcic. On the other hand, facial images themselves have become operational through their absorption into an ecosystem and economy of image databases, search engines, and AI model training and benchmarking. In other words, online images are operationalised by the biometric supply chain. This additional operational character is revealed through the existence of companies and tools like Clearview AI, as well as the proliferating number of massive image datasets built from web-scraping and surreptitious public photography.[7]

[3] Harun Farocki, 'Phantom images' (2004) 29 *Public* 12–22; Trevor Paglen, 'Operational images' (2014) 59 E-Flux (online); Mark Andrejevic and Zala Volcic, 'Seeing like a border: Biometrics and the operational image' (2022) 7(2) *Digital Culture & Society* 139–158; Rebecca Uliasz, 'Seeing like an algorithm: Operative images and emergent subjects' (2021) 36 *AI & Society* 1233–1241.

[4] Mark Andrejevic and Zala Volcic, 'Smart cameras and the operational enclosure' (2021) 22(4) *Television & New Media* 343–359.

[5] Paglen, 'Operational images'.

[6] Andrejevic and Volcic, 'Smart cameras and the operational enclosure', p. 347.

[7] See, e.g., Adam Harvey and Jules LaPlace 'Exposing.AI' (2021), https://exposing.ai.

Privacy and data protection law struggle to accommodate this theorisation of images and this domain of economic activity. For instance, the operational image ontology suggests images are always already enrolled in a biometric recognition process. Privacy and data protection law, however, understand images as 'representations' of a referent, amenable to subsequent human interpretation and inference. Under the GDPR, for example, images are only considered biometric data after 'specific technical processing' that renders it comprehensible to a machine.[8] In other words, privacy and data protection law insist on the separation of images and any biometric information that can be derived from them.[9] This means images alone cannot be biometric data. Various authors have pointed out that this is contrary to technical understandings of biometrics,[10] which would conceptualise every image as also a biometric sample, and the beginning of a biometric 'operation'. And as discussed Section 5.4.1.1, companies such as Clearview AI are exposing that a degree of processing of images, even if simply for aggregation in datasets, is already the default status of images online.

The following sections describe the different treatment of image and biometric data in existing law, with a focus on how the operational character of images expresses itself as conceptual confusion in how privacy addresses the tools and supply chains that make up the facial recognition ecosystem.

5.3 WHAT KIND OF DATA IS THAT?

5.3.1 *Images*

The following section spells out some of the internal ambiguities and inconsistencies that make the application of privacy and data protection to facial recognition supply chains difficult. The ambiguities exist even at the most basic definitional level. Privacy law typically deals with images that are identified, in cases where publication might diminish seclusion or reputation. Data protection law also governs anonymous images because the definition of 'personal data', the threshold for data protection's application, only requires that data be reasonably identifiable rather than identified.[11] There is a general presumption that images including a face satisfy

[8] *EU General Data Protection Regulation (GDPR)*: Regulation (EU) 2016/679, Art. 4(14).

[9] Ibid., Recital 51: 'The processing of photographs should not systematically be considered to be processing of special categories of personal data as they are covered by the definition of biometric data only when processed through a specific technical means allowing the unique identification or authentication of a natural person.'

[10] See, e.g., Bilgesu Sumer, 'When do the images of biometric characteristics qualify as special categories of data under the GDPR: A systemic approach to biometric data processing', IEEE International Conference of the Biometrics Special Interest Group (14–16 September 2022), referencing ISO/IEC 2382-37: 2022 Information Technology Vocabulary Part 37.

[11] *EU General Data Protection Regulation (GDPR)*: Regulation (EU) 2016/679, Art 4(1); See also *Breyer* v. *Bundesrepublik Deutschland* ECLI:EU:C:2016:779.

https://doi.org/10.1017/9781009321211.007 Published online by Cambridge University Press

that definition, the processing of which then requires a 'lawful basis', the most relevant being consent or the legitimate interests of the data processor.

The presumption that images showing a person's face are always personal data is not entirely settled, however. Even European national data protection authorities give conflicting advice. For instance, the UK Information Commissioner's Office notes that an image taken in public containing recognisable faces may not be personal data if the image is not subsequently processed to learn or decide anything about any of the individuals that are imaged.[12] The German data protection authority, however, argues that all images of people contain personal data: 'photographs, whether analogue or digital, always contain personal data … if persons can be identified on it'.[13] Advice given by other institutions is even more confusing. For instance, Oxford University's staff guidance on data protection suggests images will be personal data if individuals are the 'focus' of an image, but *not* if those individuals or groups are not the focus of the image, whatever that means.

Identification and identifiability are not always central to facial recognition and analysis, however. Not all facial recognition or analysis tasks link images to natural persons. Some may identify the same person across multiple instances of a database or across multiple cameras recording physical space. In these cases, there is an argument that facial images used in the biometric process still constitute personal information on the principle of 'singling out'. This early interpretation of 'identified' proposed by the Article 29 Working Party captures systems that distinguish an individual from a group of people without the need to connect them to a natural person.[14] Although cited several times in the jurisprudence, this definition is not necessarily authoritative.

5.3.2 *Biometric Data*

Under the GDPR, biometric data is a sub-species of personal data defined as the output of specific technical processing with a view to unique identification of a natural person.[15] It qualifies as a 'special category of personal data', requiring higher levels of protection including explicit consent for processing. The definition of 'identified' in this context is narrower than for personal data, as it requires a clear connection to a natural person. As Bilgesu Sumer notes, 'Under the current system, the threshold for identifiability for biometric data can be invoked only if there is an already identified

[12] Information Commissioner's Office (ICO), 'What happens when different organisations process the same data for different purposes?' (n.d.), https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/what-is-personal-data/what-happens-when-different-organisations-process-the-same-data-for-different-purposes/.

[13] Landesbeauftragte für Datenschutz und Akteneinsicht, 'Verarbeitung personenbezogener Daten bei Fotografien' (June 2018), www.lda.brandenburg.de/sixcms/media.php/9/RechtlicheAnforderungenFotografie.pdf.

[14] Article 29 Working Party, 'Opinion 4/2007 on the concept of personal data (WP 136, 20 June 2007)'.

[15] *EU General Data Protection Regulation (GDPR)*: Regulation (EU) 2016/679, Art 9(1).

individual under the GDPR."[16] Some data privacy laws, such as Australia's, include 'biometric templates' as protected 'sensitive data'. But as mentioned earlier, not all biometric templates are identified or created for the sake of identification, meaning the Australian definition raises confusing questions of whether there can be 'sensitive data' that is not also 'personal data'.

If and when biometric data constitutes personal data at all was a live question in policy discussions around the scope of data protection at the turn of the millennium. In 2003, the Article 29 Working Party suggested that biometric data is not personal data when templates are stored without images.[17] By 2012, however, that same group, without much elaboration, indicated that 'in most cases biometric data are personal data'.[18] Biometric data was not considered sensitive (or a special category of) data at that point though, because it did not reveal sensitive characteristics about the identified person. This position evolved again with the GDPR, as policymakers began describing certain intrinsically sensitive characteristics of biometric data, such as its persistence (non-changeability, non-deletability), its capacity to make bodies 'machine readable', its use in categorisation and segregation functions, and the way it could be used to track users across space without ever linking to their natural identity.[19] However, if the purpose of processing biometric data is 'categorisation' rather than unique identification of a natural person, it is still not considered processing of a special category of personal data.

Data protection (and privacy) law's relationship to biometrics – the requirement that a natural person be identified for biometric data to be considered a special category of personal data, and the related exclusion of unprocessed (or raw) images or videos from the definition of biometric data – are strongly informed by older biometric techniques. They imagine a database containing biometric information generated through enrolling an individual in a biometric system such as fingerprinting or DNA extraction. Privacy law identifies DNA and fingerprint information as especially sensitive types of identity information, necessitating rigorous protections and checks and balances.[20] But the law that developed around these techniques did not anticipate the reality that biometric 'enrolment' is no longer the only way to build a biometric system. It did not anticipate that any image contains within itself, easily coaxed out through readily available algorithmic methods, biometric data that might readily contribute to the construction of a facial image dataset or facial recognition search engine, or some other part of the biometric supply chain.

The realities of biometric supply chains and facial recognition ecosystems trouble these long held settlements undergirding existing regulatory strategies. The

[16] Sumer, 'When do the images of biometric characteristics qualify'.
[17] Article 29 Working Party, 'Working document on biometrics (WP 80, 1 August 2003)'.
[18] Article 29 Working Party, 'Opinion 3/2012 on developments in biometric technologies (WP 193, 27 April 2012)'.
[19] European Data Protection Board, 'Guidelines 3/2019 on processing of personal data through video devices (Version 2.0, 29 January 2020)'.
[20] *S and Marper* v. *UK* [2018] Eur Court HR 1581.

https://doi.org/10.1017/9781009321211.007 Published online by Cambridge University Press

separation between ordinary portraits and biometric samples embedded in the data protection law does not match the reality that all images are now already also 'biometric samples' – the first step in the biometric processing pipeline. Acknowledging the operational character of images would help make sense of juridical treatments of facial recognition, under privacy and data protection, that are becoming increasingly diverse, as well as assist in drawing adequate legal attention to the processes and supply chains that make up the broader facial recognition ecosystem and economy. This is the less-visible system of circulation involving a range of corporate, government and university actors, using a variety of techniques such as web scraping and surreptitious photography, to produce products for research and profit such as benchmarking datasets, training datasets, facial recognition models, and search tools.

5.4 REPRESENTATIONALISM VERSUS OPERATIONALISM IN THE CASE LAW

5.4.1 *Non-Identity Matching Cases*

While images are operationalised for facial recognition through supply chains, privacy and data protection's failure to attend to the operational image manifests at all levels of the facial recognition ecosystem. Facial recognition is not always used to match a biometric template with a natural person. Facial analysis sometimes involves consumer profiling (demographics, sentiment analysis, etc.) or location tracking (i.e., identifying a person as they move through a store/space). These instances highlight some confusion and inconsistency within privacy and data protection's conceptual apparatuses.

The Office of the Australian Information Commissioner (OAIC), for instance, evaluated a profiling system used by the 7–11 chain of convenience stores. Without clear notice, 7–11 deployed a facial recognition system for demographic (age and gender) analysis of individuals that engaged with a customer feedback tablet. The system also created a faceprint (i.e., biometric template) for the sake of quality control. To ensure the same person did not give multiple survey results within a twenty-four-hour period, faceprints were stored and compared, with multiple matches within a twenty-four-hour period flagged as potentially non-genuine feedback responses.

7–11 argued that neither the images collected nor faceprints extracted were personal information because they were not collected or processed for the sake of identifying a natural person. The images were also automatically blurred when viewed by human staff. The OAIC determined, however, that the twenty-four-hour matching system 'singled out' individuals by comparing each person's faceprint against all other faceprints held in the system, which required giving them a unique identifier. Here, the images and faceprints were linked by a 'purpose', which was the pseudo-identification. Contrary to other similar legal regimes (i.e., the US State of Illinois

Biometric Information Privacy Act (BIPA),[21] and the GDPR), the OAIC even found that the *raw* images collected were biometric information, and thus sensitive information, because they were collected for the purpose of biometric identification. The recombination of image and biometric data in this case that is so explicitly rejected elsewhere would reflect some acknowledgement of the operational character of the image, but it is better understood as an outlier, representing conceptual confusion more than a considered position. It has not been replicated in subsequent OAIC determinations considering facial recognition.[22]

Other legal regimes, such as BIPA, more explicitly avoid the issue of how to conceptualise images in a biometric context. Rather than recognise images as potentially also 'biometric samples', BIPA simply excludes photographs from its definition of biometric identifiers. The *creation* of biometric information alone invokes the Act, eliding the issue of biometric data and identifiability.[23] To that end, TikTok's collection of facial landmarks used in demographic profiling for advertising and augmented reality 'filters' and 'stickers' was illegal under BIPA. Despite TikTok's arguments that all biometric data collected was anonymous, it ultimately settled the case for $92 million as questions of identifiability and anonymity (i.e., the relations of biometric information to images) are not relevant to the BIPA regime that applies as soon as biometric data has been generated.

The diversity of legal treatments and the problems associated with maintaining the separation between images and biometric data only intensifies as we move further along the facial recognition supply chain.

5.4.1.1 Clearview AI Cases

Clearview AI collects as many images of people available online as possible (approximately 1.5 billion images collected per month), storing them in a database linked to their source URLs. Clearview AI extracts biometric information from every face in every image and uses that biometric data to create a unique mathematical hash for each face. Those hashes make the image database searchable via a 'probe image' that is itself hashed and compared against the database. Any matches between the probe image and the image database are then provided to the user along with image URLs. Litigation so far has assumed the availability of the system only to law enforcement (and related entities), although Clearview AI now also provides biometric products to the private market.

[21] *Biometric Information Privacy Act* (740 ILCS 14/).

[22] See, e.g., Megan Richardson, Mark Andrejevic, and Jake Goldenfein, 'Clearview AI facial recognition case highlights need for clarity on law' (22 June 2022), CHOICE, www.choice.com.au/consumers-and-data/protecting-your-data/data-laws-and-regulation/articles/clearview-ai-and-privacy-law.

[23] See, e.g., *Patel* v. *Facebook* No. 18-15982 (9th Cir. 2019) – 'the development of a face template using facial-recognition technology without consent' is an invasion of a privacy interest.

https://doi.org/10.1017/9781009321211.007 Published online by Cambridge University Press

Judicial treatment of Clearview AI has consistently found that the company processes personal and sensitive data, and therefore requires consent from the individuals in the images it collects. Clearview AI persistently argues that the data it processes is neither personal nor sensitive, but fails on this claim. The French data protection authority, CNIL, similarly (although somewhat circularly) stipulated in its finding against Clearview AI that images are personal data as soon as an individual can be recognised, and that Clearview AI's capacity to compare an image with another makes those images identifiable.[24] Because Clearview AI does not perform a specific processing operation for the unique identification of a natural person however, the images it collects and biometric data it extracts are not special categories of personal data. Because Clearview AI only processes personal data and not special categories of personal data, that processing *could* be lawful even without consent under GDPR Article 6, for instance if in the legitimate interests of the company. However, the court dismissed the possibility of any legitimate interest because individuals who placed their images online would not have 'reasonably expected' those images to participate in a biometric search engine that might be used for law enforcement purposes.[25] But this finding around reasonable expectations is a flimsy hook on which to hang Clearview AI's privacy violations, and explicitly rejects the operational character of images. Do individuals still expect that images published online are not used to train AI models or produce image datasets? Do individuals still believe that the function of an online image is its presentation to other humans? How long can such expectations persist?

There was a similar moment in an Australian finding against Clearview AI. The Australian regulator determined that the images collected by Clearview AI were personal data because Clearview AI's purpose is to facilitate identification.[26] And the biometric data was sensitive because the Australian definition includes biometric templates even if not used for the specific identification of a natural person. When contemplating whether Clearview AI satisfied any exceptions for processing sensitive information without consent, the OAIC indicated that the individuals whose personal and sensitive information was being collected by Clearview AI would not have been aware or had any reasonable expectation that their images would be scraped and held in a database. Further, no law enforcement exceptions applied because 'only a very small fraction of individuals included in the database would ever have any interaction with law enforcement'.[27]

[24] Decision 2021-134 of 1 November 2021 issuing an order to comply to the company Clearview AI (No. MDMM211166).

[25] *EU General Data Protection Regulation (GDPR)*: Regulation (EU) 2016/679, Art 6(1)(f) specifies that even if data is publicly available it still requires a legal basis for processing and is not automatically available for re-use. When processing publicly available data on the basis of a legitimate interests, the European Data Protection Board suggests users need to reasonably expect that further processing.

[26] Commissioner initiated investigation into Clearview AI, Inc. (Privacy) [2021] ALCmr 54 (14 October 2021).

[27] Ibid., at [172].

https://doi.org/10.1017/9781009321211.007 Published online by Cambridge University Press

These discussions of 'reasonable expectation' expose something about data privacy law's relationship to operationalism. On one hand, the breach of reasonable expectations about images law enforcement databases makes sense – there is a liberal privacy harm associated with being enrolled in a police database when a person is not deserving of suspicion. That has served as a normative boundary in privacy jurisprudence for some time. But on the other hand, this is not really enrolment in a police database: Clearview AI's database is an index of all the images on the internet that is, at the moment, primarily available only to police, but increasingly to private parties. Determining whether Clearview AI breached data privacy law with a normative standard associated with delimiting the state's policing powers,[28] does not seem adequate if we understand Clearview AI as just one of the large and growing number of image databases and biometric services that operationalise facial images by scraping the internet. What Clearview AI explicitly demonstrates is that there is no longer a police database; the internet is already an image database that is operationalised through a biometric supply chain.

Online images are sometimes viewed by humans or police, but they are primarily viewed by other machines such as web-scraping software and facial recognition algorithms for the sake of assembling the facial image datasets and searchable biometric databases that power a broader biometrics economy and ecosystem. Regulating these systems by consent (as required when defining the biometric data involved as sensitive – or a special category of personal – data) only makes sense when we imagine the internet as a media system browsed by humans,[29] where image consumption and processing is neither automatic nor at scale. Clearview AI is a jarring demonstration of the reality that humans do not browse the internet; the internet browses us.

5.4.1.2 Scraping and Dataset Cases

Clearview AI has exposed how legal settlements informed by rhetorics of 'open internet' that, for instance, stabilised the legality of web-scraping, indexing, and enabled search engines to evolve, are now straining in the context of massive data aggregation for training large machine learning models.[30] Facial recognition has its own scraping dynamics that produce not only search engines, but also facial image datasets that, while frequently produced by research teams in non-commercial contexts, have massive economic value and include a huge number of individuals. The market for datasets

[28] Jake Goldenfein, *Monitoring Laws* (Cambridge University Press, 2019).

[29] Chloe Xiang, 'AI is probably using your images and it's not easy to opt out' (26 September 2022), Vice: Motherboard, www.vice.com/en/article/3ad58k/ai-is-probably-using-your-images-and-its-not-easy-to-opt-out.

[30] See, e.g., Benjamin L. W. Sobel, 'A new common law of web scraping' (2021–2022) 25 *Lewis and Clark Law Review* 147–207; Vladan Joler and Matteo Pasquinelli, 'Nooscope' (2020) https://nooscope.ai/.

https://doi.org/10.1017/9781009321211.007 Published online by Cambridge University Press

was estimated to be $9 billion in 2022.[31] There are a number of giant image datasets containing images of any person for whom there are a multitude of images available online – be they celebrities, political figures, or activists.[32] For instance, the 'Have I been Trained' tool can identify whether individuals are included in the notorious LAION 5B and LAION 400M datasets, used to train a substantial number of AI tools, and since refined into a large number of other industrially valuable image datasets.[33] To some extent, the new rules in the EU AI Act will prohibit this type of indiscriminate scraping by Clearview AI. But because the rules only address 'untargeted' scraping for the creation of 'facial recognition databases' it will hardly disturb the facial image dataset industry. As discussed below, apart from Clearview AI, the majority of the industry is vertically dis-integrated, meaning entities doing scraping are producing facial image datasets not biometrically identified facial recognition databases like Clearview AI.

Scraping and image datasets are often produced by companies or research institutions not themselves involved in biometric analysis or facial recognition applications, but who still perform a critical task in the facial recognition supply chain. Companies producing image datasets typically argue that images without names do not constitute personal information. Alternatively, they may claim to only index image URLs not the images themselves (i.e., making images available for other parties to download) so as to not process image data at all. If they are processing images, that processing is claimed to be legal because it is in the legitimate interests of the entity,[34] Many image datasets made available without any associated biometric information, with subsequent users performing biometric analysis to link particular individuals across multiple images. Sometimes they are simply used to test and benchmark algorithmic models, enabling a demonstration of an algorithm's efficacy.[35] These companies mostly evade privacy scrutiny and will likely avoid regulation by the AI Act. Clearview AI managed to attract legal attention for its supply chain activities because its vertical integration (i.e., because it scraped the images, ran the biometric analysis, and sold the identification service) linked those supply chains to the product / application level where privacy and data protection more comfortably apply.

Image datasets are also created without web-scraping – typically through surreptitious photography. Facial recognition in public space has different demands

[31] Madhumita Murgia, 'Who's using your face? The ugly truth about facial recognition' (19 April 2019), *Financial Times*, www.ft.com/content/cf19b956-60a2-11e9-b285-3acd5d43599e.

[32] Ibid.

[33] https://haveibeentrained.com/

[34] See, e.g., https://laion.ai/faq/.

[35] The diversity of actors in the facial recognition supply chain also enables problematic 'data laundering' practices. Datasets are legally constructed by research institutions using non-commercial research exceptions to copyright law, but then made available to commercial entities that use them for profit: see Andy Baio, 'AI data laundering: How academic and nonprofit researchers shield tech companies from accountability' (30 September 2022), Waxy, https://waxy.org/2022/09/ai-data-laundering-how-academic-and-nonprofit-researchers-shield-tech-companies-from-accountability/. reporting on a Meta owned generative text-video tool trained on the WebVid-10M dataset that was initially scraped from Shutterstock, as well as the XPretrain dataset released by Microsoft of millions of videos scraped from YouTube with text descriptions.

https://doi.org/10.1017/9781009321211.007 Published online by Cambridge University Press

to identify verification systems that use portraits for biometric enrolment. Images scraped from the web are frequently too posed and flat-angled to produce biometric models able to identify individuals from images and video captured from more common surveillance vantage points. Facial recognition in the wild needs images of people walking around, looking at their phones, being unknowingly recorded. This is why, for instance datasets such as Brainwash, produced with a webcam in a café, capturing images of returning customers waiting to order coffee, as well as the Duke-Multi-Target, Multi-Camera Unconstrained College Student Dataset, produced with synchronised surveillance cameras taking pictures of students walking between classes from a university office window, are so valuable.[36] Data scientists are increasingly seeking access to CCTV footage for building novel datasets.[37] Although surveillance for dataset construction does not raise the same risk of real-time mass surveillance that animates privacy thinking, in the world of operationalism, those images still participate in the facial recognition ecosystem and economy, raising new critical questions that few existing legal concepts, let alone privacy, are able to answer.

A comprehensive analysis is beyond the scope of this chapter, but no legal regime clearly imposes meaningful limitations in this domain. The *HiQ* v. *LinkedIn* case seemingly upheld the legality of scraping under the US Computer Fraud and Abuse Act, even if contrary to platform terms of service.[38] Scraping does not interfere with personal property interests because there are no property rights in data. Exploitation of Creative Commons non-commercial licensed images is permissible because of the data laundering (commercial/non-commercial) techniques described in footnote 37 as well as the general copyright exemptions for research purposes.[39] Some argue that scraping images to build datasets or train algorithms does not involve market substitution or replication of any 'expressive' dimension of images, meaning it may not violate copyright anyway.[40] There are already fair use (or equivalent) exceptions for search engines in many jurisdictions.[41] The US privacy-adjacent right of publicity is unlikely to apply when a scraped image has no commercial value prior to its appropriation and exploitation and does not result in subsequent publication.[42]

[36] See, e.g., Harvey and LaPlace, 'Exposing.AI'.

[37] See, e.g., UC Riverside Video Computing Group. 'Datasets' (n.d.), https://vcg.ece.ucr.edu/datasets.

[38] *HiQ Labs* v. *LinkedIn Corp*., 938 F.3d 985 (9th Cir. 2019).

[39] See, e.g., Ryan Merkley, 'Use and fair use: Statement on shared images in facial recognition AI' (13 March 2019), Creative Commons, https://creativecommons.org/2019/03/13/statement-on-shared-images-in-facial-recognition-ai/.

[40] Sobel, 'A new common law of web scraping'.

[41] See, e.g., Jonathan Band, 'Google and fair use' (2008) 3 *Journal of Business & Technology Law* 1–28.

[42] See, e.g., including for contrasting views, Wendy Xu, 'Recognizing property rights in biometric data under the right to publicity' (2020–2021) 98 *University of Detroit Mercy Law Review* 143–166; Lisa Raimondi, 'Biometric data regulation and the right to publicity: A path to regaining autonomy over our commodified identity' (2021) 16(1) *University of Massachusetts Law Review* 200–230; A. J. McClurg, 'In the face of danger: Facial recognition and the limits of privacy law' (2007) 120 *Harvard Law Review* 1870–1891.

It will be interesting to see the outcome of the pending *Vance* v. *IBM* litigation concerning IBM's refining of Flickr's YFCC100M dataset into the Diversity in Faces dataset.[43] But this case also deals only with governance of biometric information and not the images from which that biometric data is derived, meaning it will not enjoin dataset creation more generally. At the same time, industry- and research-aligned actors have started pushing in the other direction, arguing for freedoms to use and reuse datasets,[44] rights 'to process data' without consent,[45] with clear exceptions for copyright or usufructuary rights over property interests to maximise capacities to build and train machine learning models.[46]

5.5 CONCLUSION

The way privacy and data protection are configured may make sense if online images are representations of individuals, browsed by humans, at risk of certain autonomy effects; but it makes much less sense if images are already part of a socio-technical ecosystem, viewed primarily by machines, used to train and benchmark facial recognition algorithms in order to produce economic value. Privacy and data protection's representationalism struggles to grasp the mobilisation of images as supply chain components in a dynamic biometric ecology. This chapter has argued that the issues in this 'back end' of the facial recognition ecosystem are very different from those that have been typically raised in privacy discussions. Here, regulatory questions intersect with what has become a new frontier of value creation in the digital economy – facial recognition model training. The concern is no longer exclusively losing anonymity in public, but also information being captured from public spaces, not for the sake of identifying you, but for the sake of generating an archive of images of you in the wild in order to train facial recognition models and extract economic value. Once we pay attention to how facial recognition systems are built and function, privacy and data protection start to lose their grip.

[43] *Vance* v. *IBM* Case: 1:20-cv-00577.

[44] PIJIP, 'Joint comment to WIPO on copyright and artificial intelligence' (17 February 2020), Infojustice, https://infojustice.org/archives/42009.

[45] See, e.g., Mauritz Kop, 'The right to process data for machine learning purposes in the EU' (2021) 34 *Harvard Journal of Law & Technology – Spring Digest* 1–23.

[46] See, e.g., Inioluwa Deborah Raji, Timnit Gebru, Margaret Mitchel, Joy Buolamwini, Joonseok Lee, and Emily Denton, 'Saving face: Investigating the ethical concerns of facial recognition auditing' (2020), AAAI/ACM AI Ethics and Society Conference 2020; Vinay Uday Prabhu and Abeba Birhane, 'Large image datasets: A Pyrrhic win for computer vision?' (2020), arXiv:2006.16923.