Data & Policy (2025), 7: e59
doi:10.1017/dap.2025.10019

RESEARCH ARTICLE

From ‘wild west’ to ‘responsible’ AI testing ‘in-the-wild’: lessons from live facial recognition testing by law enforcement authorities in Europe

Karen Yeung[1] and Wenlong Li[2]

[1]Interdisciplinary Professorial Fellow in Law, Ethics and Informatics, Birmingham Law School & School of Computer Science, University of Birmingham, Birmingham, UK
[2]Research Professor, Guanghua Law School, Zhejiang University, China
Corresponding author: Karen Yeung; Email: k.yeung@bham.ac.uk

Received: 28 October 2024; **Revised:** 30 April 2025; **Accepted:** 26 June 2025

Keywords: technology testing; Facial Recognition Technology (FRT); in-the-wild trial; law enforcement; AI testing

Abstract

Although ‘in-the-wild’ technology testing provides an important opportunity to collect evidence about the performance of new technologies in real world deployment environments, such tests may themselves cause harm and wrongfully interfere with the rights of others. This paper critically examines real-world AI testing, focusing on live facial recognition technology (FRT) trials by European law enforcement agencies (in London, Wales, Berlin, and Nice) undertaken between 2016 and 2020, which serve as a set of comparative case studies. We argue that there is an urgent need for a clear framework of principles to govern real-world AI testing, which is currently a largely ungoverned ‘wild west’ without adequate safeguards or oversight. We propose a principled framework to ensure that these tests are undertaken in an epistemically, ethically, and legally responsible manner, thereby helping to ensure that such tests generate sound, reliable evidence while safeguarding the human rights and other vital interests of others. Although the case studies of FRT testing were undertaken prior to the passage of the EU’s AI Act, we suggest that these three kinds of responsibility should provide the foundational anchor points to inform the design and conduct of real-world testing of high-risk AI systems pursuant to Article 60 of the AI Act.

Policy Significance Statement

This paper has direct policy implications concerning the testing of AI systems in real-world conditions, offering high-level principles which should inform how such tests should be undertaken to ensure that they are conducted in a legally, ethically, and epistemically responsible manner.

1. Introduction

Organisations everywhere are seeking to adopt AI tools and systems, inspired by their promised benefits. The expanding capabilities of machine learning algorithms to generate predictions by finding patterns within massive datasets underpin much of the ‘AI hype’ currently driving organisational take-up.

This paper is based on a Keynote Address by Karen Yeung at the Data & Policy Annual Conference, hosted on-line on 15 September 2021.

Although software developers proudly proclaim the accuracy of these predictions, the conditions under which accuracy is tested and evaluated are highly variable. Nor do accurate predictions necessarily translate into concrete benefits to the deploying organisation, let alone deliver wider public benefits. To help address uncertainty about their performance, reliability and robustness, the testing of AI systems occupies a vitally important role. Although software testing (including AI-enabled software) conventionally occurs 'in the lab' during development, it may also be undertaken 'in-the-wild', that is, in real-world settings. For example, in the digital media industry, A/B testing on live users is common, enabling developers to discover which online interventions better elicit the desired response, despite criticisms that these practices violate basic principles of research ethics (Benbunan-Fich, 2017; Jiang et al., 2019). In the automotive industry, the testing of driverless (or 'autonomous') vehicles on public roads without a human 'safety' driver is well underway (Roy, 2024). Yet these experiences, together with the longer history of new technology testing undertaken 'in the wild', remind us that testing processes may themselves be hazardous, even fatal. For example, GM's subsidiary Cruise is under investigation for several driverless vehicle accidents and near-misses, including an incident in San Francisco in which a pedestrian (who had already been knocked down by a human driven vehicle whose driver fled the scene) was hit by a Cruise car and dragged 20 feet, sparking public outcry and prompting the company to halt its operations nationally (Cano, 2024). Although the on-road testing of driverless vehicles is now subject to bespoke regulatory frameworks for the automotive industry in many jurisdictions, 'in-the-wild' AI testing remains a largely ungoverned 'wild west' despite the dangers they pose to others.

This is particularly troubling given that the dangers of real-world AI testing extend beyond conventional health and safety risks to potential fundamental rights interferences, which are often difficult to see, let alone systematically observe and measure (Purshouse and Campbell, 2019; 2022). To this end, we welcome provisions in the EU AI Act that govern the conduct of real-world testing for so-called 'Annex III high-risk' AI systems (consisting primarily of AI systems identified as posing a significant threat to fundamental rights and listed in Annex III of the Act) requiring prior approval by a market surveillance authority, including approval of a 'real world testing plan'. However, it is left to the European Commission to specify the detailed elements of such real-world testing plans pursuant to Art 60(1) of the Act. What principles should inform these elements and the Commission's decisions to authorise real-world testing of high-risk Annex III systems? In the absence of clear conventions or established best practices, we are in uncharted territory. It is precisely this gap that this paper seeks to fill, highlighting this serious governance deficit. We argue that there is an urgent need to ensure that these activities are undertaken 'responsibly' in a legal, ethical, and epistemic sense. Accordingly, they constitute critical requirements that should guide and govern the design, conduct, and evaluation of 'in-the-wild' testing of AI systems (particularly those which pose 'risks to fundamental rights') and should therefore inform the European Commission when specifying the elements that real-world testing plans must meet pursuant to Art 60(1).

Our argument proceeds in five parts. First, we begin by reflecting on the nature and purpose of tests generally, particularly in-the-wild technology testing, drawing on insight from STS and highlighting their potential benefits and dangers. Second, we argue that to safeguard against their dangers, the testing of new technology (including AI systems) in 'real world' conditions should be designed, conducted, governed, and evaluated with the aim of ensuring that testing is undertaken in an epistemically, legally, and ethically responsible manner. Third, we investigate live facial recognition technologies (FRT) trials undertaken by various law enforcement authorities ('LEAs') in Europe, which serve as rich comparative case studies. We argue that although these live FRT trials can be understood as welcome attempts by LEAs to acquire first-hand knowledge about the performance, capability, and usability of live FRT identification systems to carry out their law enforcement functions, they suffer from significant shortcomings. As such, they offer valuable lessons, confirming the need to ensure that such tests are conducted in a lawful and ethical manner and designed to produce test results that have epistemic integrity. Fourthly, we briefly highlight similarities and differences between the way in which AI systems are currently tested and evaluated with the testing of new drugs and vaccines before they can be made publicly available. A short conclusion follows, which touches upon how our analysis should help inform the interpretation and implementation

https://doi.org/10.1017/dap.2025.10019 Published online by Cambridge University Press

of relevant provisions of the AI Act, although an examination of the Act or specific provisions is beyond the scope of our paper.

2. Why test AI systems 'in-the-wild'?

Our examination begins by reflecting on the nature and purpose of testing generally, and new technology testing in particular, drawing on insights from Science and Technology Studies (STS), applied ethics, and the history of technology.

2.1. Why test anything?

Testing offers as a means for acquiring knowledge and understanding about the world. STS scholars have long regarded testing as crucial to making science and innovation 'special': It is through laboratory testing that science and engineering derive their exceptional power to 'render invisible or distant natures observable' (Latour, 1993; Mody and Lynch, 2010; Vertesi, 2015) and 'durably transform socio-technical arrangements in society' (Callon, 1986; Pinch, 1993). Sociologists have highlighted the purposive nature of any kind of test, in which expectations are built around a certain outcome, although those purposes might not be formally specified. For example, Trevor Pinch claims that it is usually clear to participants just what sort of test outcomes are to be expected, such as a rock band warming up by testing its sound system with the aim of producing demonstrable evidence that it operates as intended. In this respect, many tests are *performances* that can be witnessed by others, although they may have different interests in the outcome (Pinch, 1993, p. 26).

Data produced from technology testing are conventionally believed to provide access to the purely technical realm, revealing the technology's immanent logic through formal assessment of its performance (Pinch, 1993, 25). During the late 1980s, scholars of the sociology of science began to turn their attention to technology, viewing technology tests as analogous to science experiments (Marres and Stark, 2020). Like scholars of the sociology of scientific knowledge who preceded them, sociologists of testing demonstrated that there is an irreducibly social dimension to technology testing, entailing determinant social practices that 'makes up for the indeterminacy of scientific knowledge' (MacKenzie, 1989, 416-7; 1990; 1999). Thus, for MacKenzie, the task of the sociology of testing is to account for the legitimacy of technological testing and the knowledge it produces in *social* terms. To this end, sociologists of testing have highlighted the wide variety of norms and standards associated with why and how technology tests are conducted, emphasising the ways in which relevant communities of practice that engage in or are associated with these tests, understand their purpose.

Yet if technology tests are to achieve their intended purpose, they must be designed and conducted in a manner that can be expected to help achieve them. Compare, for example, academic examinations to test student achievement and the need for fair and appropriate 'test conditions' with consumers who 'test drive' different cars before making a purchasing decision, aimed at acquiring direct first-hand experience about whether a particular make and model meets their needs, objectives, and expectations. These familiar examples draw into sharper focus the primary purpose for which new technology is tested in 'live' settings: that is, to gather evidence to help assess whether, and to what extent, the technology can perform and produce expected (and desired) outcomes, rather than relying on claims of others. In short, in-the-wild (or 'live') testing offers learning and knowledge-gathering opportunities that controlled lab trials do not, and may be particularly important for AI technologies given that many are intended to operate adaptively in response to their environment. In this respect, in-the-wild testing may be particularly valuable and important because a technology's utility is derived from its real-world use, for which results from controlled laboratory testing cannot demonstrate. Yet whether in-the-wild testing in fact serves this purpose will depend on the test conditions, methods, and, above all, whether tests are designed, conducted, and evaluated with their intended purposes and context in mind. Hence, van de Poel argues that technology tests (which he calls 'experiments') must be responsible in an *epistemological* sense: to ensure that the experiment increases the possibilities of robust, reliable learning. Epistemological

https://doi.org/10.1017/dap.2025.10019 Published online by Cambridge University Press

responsibility entails setting up experiments in such a way that society can learn most from them and that the resulting knowledge is as 'reliable and relevant as possible' (van de Poel, 2016, 671). This, in turn, requires *methodological rigour* in test design, conduct, and evaluation.

2.2. The dangers of technology testing in-the-wild

Although in-the-wild technology testing can be valuable and important, its potential dangers are readily apparent. The history of technology is replete with examples of technology testing that caused serious harm to others and to the environment (Resnik and Hofweber, 2023), including nuclear weapons testing during the Cold War with devastating and on-going adverse biological, ecological, and social consequences, including the displacement of hundreds of people (Sheriff, 2023). When technology testing and experimentation is undertaken for the purposes of *scientific research* within academic settings, it is now governed by well-established principles of research ethics rooted in the principle of informed consent and administered in institutional systems now familiar to university researchers (called 'institutional review boards' in the United States, and 'research ethics committees' in the United Kingdom and Europe) (Buchanan and Ess, 2009). These ethical norms, which are primarily concerned with the protection of human subjects, were first introduced by the Declaration of Helsinki (World Medical Association, 2018), the first international set of ethical principles to govern research to prevent egregious research abuses undertaken on prisoners by Nazi scientists during World War II (Lederer, 2007). Unlike scientific research, however, technology testing conducted outside formal academic settings has hitherto escaped systematic scrutiny or regulatory oversight (Polonetsky et al., 2015; Tene and Polonetsky, 2016). Although the reasons are far from clear, it may be due, in part, because technology is defined by reference to its capacity to perform a useful function for which no recognisable professional community claims a social licence (*cf* Dixon-Woods and Ashcroft, 2008). Unlike scientific testing and experimentation, technology tests are not primarily aimed to produce 'verifiable truth claims' in the form of scientific knowledge. Rather, their aim is to generate evidence that the technology can indeed perform a useful function capable of delivering its intended benefits.

2.3. Towards a principled framework for 'responsible' technology testing in-the-wild

It is against this backdrop of 'irresponsible' technology testing that contemporary AI testing in-the-wild arises for consideration. While the need to avoid causing harm to the health, safety and property of others may be self-evident in relation to the testing of weapons or machinery (including vehicles), the need for responsible testing is equally important when testing technologies that could threaten individual and collective goods, including potential fundamental rights violations (including privacy interferences) which may not give rise to tangible, readily visible harm. In the absence of any general institutional framework to govern, constrain, and monitor in-the-wild testing, AI developers and potential deployers cannot be expected and relied upon to conduct them in a responsible, rights-respecting manner, even as the power, sophistication, and penetration of AI systems continue to expand and deepen across multiple domains. Principled and effective frameworks to ensure that live technology testing, including the testing of AI systems, is undertaken responsibly are long overdue and urgently needed, as the case studies we examine in Section 2 highlight. But what does responsible technology testing 'in-the-wild' require? In the following section, we suggest that this entails three kinds of responsibility (epistemic, ethical, and legal) forming the foundation of three guiding principles that we provisionally suggest should guide and constrain how in-the-wild technology tests (including AI-enabled technologies) should be designed, conducted, and evaluated.

a. Epistemic integrity and responsibility

Epistemic integrity concerns the production of sound knowledge (Hammersley, 2020). The epistemic integrity of any test (whether 'in-the-wild' or otherwise) will depend upon its purpose, the conditions

https://doi.org/10.1017/dap.2025.10019 Published online by Cambridge University Press

under which it is conducted, and the manner and methods through which it is undertaken and evaluated. Epistemic validity denotes that the test results are accurate, reliable, and cognizant of the limitations of the test, including error rates. Epistemic validity lies at the foundation of the scientific method, demanding care and attention in research design, methods, and protocols, that is, the written procedures or guidelines that provide the blueprint for scientific study, including testing methods and practices, rooted in the need for reproducibility that is considered foundational to scientific integrity. (Persaud, 2010, 1132) The epistemic validity of scientific tests and experiments demands that the purpose of any test, in addition to the test thresholds that must be met to count as 'success', is specified in advance. Yet epistemic integrity does not appear to have been highlighted in the discussions of technology testing. One notable exception is the 'Volksvwagen (VW) dieselgate' scandal which erupted following the discovery that VW had for many years installed software 'cheat devices' into VW diesel vehicles which could automatically detect whether they were being tested in the lab for emissions and thus operated in 'low pollution/poor engine performance' mode, and then automatically switched to 'high pollution/high engine performance' mode when the vehicle was being driven on the road. (Hammersley, 2020) The Dieselgate scandal is a sobering reminder that epistemic integrity is as important in the design and conduct of testing technology for regulatory and other real-world uses, as it is for laboratory tests for scientific research (Van de Poel , 2017), primarily to provide sound, reliable, and accurate information, but also to avoid wasting time, effort, and expense associated with poorly designed tests and testing processes. Technology tests should therefore be designed and undertaken in an epistemically responsible manner, which requires that tests are designed and conducted with the purpose of the test in mind. In practice, this means that the test's purpose (or purposes) must be identified and expressly stated in advance to serve as the reference point for identifying and putting in place appropriate test conditions (whether in the lab or 'in-the-wild') and utilising appropriate methods, including identification in advance of the data that will serve as evidence of test performance and pre-specified criteria for evaluating its success.

In addition, as we will explain below, when testing AI and other digital technologies 'in-the-wild', it is also necessary to clarify their purposes, including whether the test seeks to evaluate

(i) **functional performance**: in terms of the technology's capacity to undertake specific tasks, including the generation of accuracy of AI-enabled outputs and predictions. For example, the accuracy of predictions in 'wild' settings may be harder to achieve than in highly controlled lab settings; and/or

(ii) **operational effectiveness**: Given that a technology is defined by reference to its capacity to serve a 'useful purpose', then an artefact capable of outstanding functional performance in carrying out a specific task may be of little practical use if it fails to facilitate the achievement of some specific socially useful purpose and thus deliver real-world benefits.

b. Ethically responsible

The history of harms and abuses arising from live technology testing and from scientific experimentation reminds us that, as van de Poel argues, technology experiments must be responsible in an *ethical* sense due to their uncertain impacts (Van de Poel, 2017, 83). Accordingly, because the conduct of a technology test or trial may itself generate dangers and uncertainties (not only to trial participants but also to affected stakeholders and other vital societal interests, including the environment), safeguards are needed to prevent harm and wrongdoing, and for which informed consent of trial participants is typically required.

c. Legally responsible

Since the Declaration of Helsinki, informed consent has been recognised as a necessary prerequisite to ensure that scientific research and experimentation are conducted in an ethically responsible manner. This

https://doi.org/10.1017/dap.2025.10019 Published online by Cambridge University Press

requires that research participants are informed of the potential risks of their involvement in research trials and their express consent to participate is obtained before proceeding (Miller and Wertheimer, 2010; Kirchhoffer and Richards, 2019). In contrast, contemporary technology testing is replete with examples in which those affected by them had not been consulted in advance, informed of their dangers, let alone given their informed consent. Numerous historical and contemporary examples illustrate the lack of prior consultation, informed consent, and transparency in testing. The Tuskegee Syphilis study conducted by the U.S. Public Health Service infamously withheld treatment from African American men to study the progression of syphilis, without their informed consent or knowledge (Reverby, 2013). During World War II, Nazi doctors conducted lethal experiments on concentration camp prisoners, violating various fundamental ethical principles, including autonomy (Vollmann and Winau, 1996). Moreover, civilians and soldiers were subjected to atomic testing without full awareness of the risks (Lemmens, 1999). More recently, digital technologies have been tested without user awareness and consent; the notorious Facebook's emotional contagion experiment manipulated users news feeds to assess emotional responses without their explicit knowledge (Jouhki et al., 2016; Boyd, 2016; Bird et al., 2016; Flick 2016). Yet the adverse impacts of technology testing conducted 'in the wild' include both tangible harms to health, safety, and the environment, and potential interferences with the legal and fundamental rights of others and important collective interests and societal values (Grimmelmann, 2015; Humphreys, 2015). In particular, a number of AI-enabled technologies have already demonstrably interfered with fundamental rights, including rights to privacy and data protection, particularly those entailing the collection and processing of personal data to profile individuals and/or which entail mass surveillance. Moreover, many contemporary digital technologies continue to be designed with men in mind while failing to serve the needs and interests of women, (Criado-Perez, 2020; Dastin, 2022) highlighting the need to ensure that technology testing should be undertaken in accordance with legal duties and obligations, including (but not limited to) legal duties to avoid discrimination on the grounds of race, gender, and other protected characteristics. In other words, care must be taken to ensure that technology tests undertaken in-the-wild are conducted lawfully in a manner that respects the legal and fundamental rights of others, and that any legal duties and obligations associated with their conduct are duly discharged.

3. A comparative case study: live testing of FRT 'in-the-wild' by law enforcement authorities in Europe

To demonstrate why adherence to principles of epistemic, ethical, and legal responsibility should guide and govern how in-the-wild AI testing is undertaken, we provide a comparative investigation of selected recent live facial recognition technology (FRT) trials undertaken by law enforcement authorities in Europe between 2016 and 2020. Our analysis shows that these tests were conducted in an ad hoc manner, often without adequate concern for epistemic validity and/or oversight, transparency, or public accountability, while posing considerable risks to the fundamental rights of individuals who came into contact with them. We begin with a brief account of how FRT systems function and why LEAs wish to test them, before describing four well-publicised live FRT trials undertaken in London, South Wales, Berlin, and Nice. We include these four case studies due to their publicity and level of public engagement, and these four provide an excellent contrast in critical aspects of trial design and development. There are other trials conducted but, given the limited accessibility and publicity of the trialling details and outcomes, we have excluded them from our analysis (Roth, 2021).

3.1. What are 'live' Facial Recognition Technology (FRT) systems?

To understand LEAs' interest in live FRT systems, and the challenges associated with testing them 'in the wild', a basic understanding of what FRT is and how it functions is needed. (Ada Lovelace Institute, 2019; 2022) FRT systems utilise digital image recognition software to capture and analyse the physical characteristics of human faces to identify and match a person's facial image against a stored facial image database, typically for identification or verification purposes (Gates, 2004; 2011). They rely on

https://doi.org/10.1017/dap.2025.10019 Published online by Cambridge University Press

computational algorithms that are configured to automatically scan digital images to detect human faces by searching for characteristics indicating the presence of certain facial features (Introna and Nissenbaum, 2010; Mordini and Tzovaras, 2012; Trigueros et al., 2018; Buolamwini et al., 2020), which detect and measure patterns in facial measurements to create a mathematical 'template' for each facial image identified (Scottish Parliament Justice Sub-Committee on Policing, 2020, 5). These algorithms are developed by parsing large volumes of training data, consisting of static digital facial images, through machine learning models that recognise distinctive facial features. FRT matching is probabilistic in that it evaluates the statistical similarity between two facial image templates and identifies a 'match' depending upon a pre-programmed statistical error threshold. FRT systems have been used retrospectively by various law enforcement authorities (LEAs) to match images from CCTV video footage of an incident involving an unidentified person against police image databases, utilising FRT's 1-1 identification function (Lochner, 2013; Scottish Parliament Justice Sub-Committee on Policing, 2020). As the field of machine vision continues to advance, its ability to accurately predict whether an individual in a digital image is correctly 'matched' to another image of that individual continues to improve, with more complex machine learning algorithms claiming to offer more granular and dynamic pattern recognition capabilities (Huang et al., 2008; Parkhi et al., 2015; Zafeiriou et al., 2015) designed to recognise particular parts of a face (e.g., colour of hair, texture of skin), to scan the iris, and to detect, capture, and analyse facial dynamics (expressions and micro-expressions) and other kinds of bodily movement, gait, and voice in real-time. (Mordini and Tzovaras, 2012). The attraction of live FRT lies in its promise of enabling the 'discovery' of wanted individuals (or to detect and identify 'suspicious behaviour' of any individuals whose identity may be unknown) within a flowing and uncontrolled crowd in *real-time*, so that they can be apprehended and taken into custody.

3.2. Lessons from the field: snapshots from selected live FRT trials by European LEAs

Armed with this understanding of live FRT systems, we now consider how FRT systems have been live tested by LEAs across several European states. We begin by describing each trial along with brief accompanying observations. A comparative critical analysis of the trials as a whole then follows, in light of the three guiding principles for 'responsible' in-the-wild technology testing provisionally identified above. Our sources were limited to publicly available, English-language sources, which varied considerably between sites, with the most extensive public reporting for the London trials. At each testing site, the LEA's aim was to test the performance of the technology's 'discovery' function and, in some cases, to test its contribution to the operational effectiveness in facilitating the successful apprehension of 'wanted' individuals in open public settings. As we shall see, testing for operational effectiveness requires the successful integration of several complex and challenging technical and social tasks, for which accurate matching of digital images of a person's face is but one.

3.2.1. London, United Kingdom

Although the London Metropolitan Police Service (MPS) first began experimenting with FRT in the late 1990s, (Fussey, 2012, 23) it has more recently partnered with private software firm NEC to trial its NeoFace software. Between 2016 and 2019, the MPS conducted ten 'operational' trials (to identify individuals who were in fact 'wanted' by the police rather than of volunteers) and one non-operational trial across London in various settings, ranging from highly crowded and unconstrained environments (including the London Notting Hill Carnival and Soho in central London), to open yet well-controlled environments with a narrow flow of people all facing towards the cameras (e.g., during the Remembrance Sunday parade in London's Parliament Square and at the Port of Hull as ferry passengers embarked and disembarked) and busy town centres including Stratford and Romford (see Figure 1, Fussey and Murray, 2019). Specific yet varied technological objectives for each trial were identified (National Physical Laboratory, 2020), each seeking to reflect and build upon the learning outcomes of earlier trials. The size of the watchlists varied considerably: Indeed, watchlist size was one specific component which the trials purposefully sought to test, with the NPL's evaluation report identifying a significant increase in watchlist

https://doi.org/10.1017/dap.2025.10019 Published online by Cambridge University Press

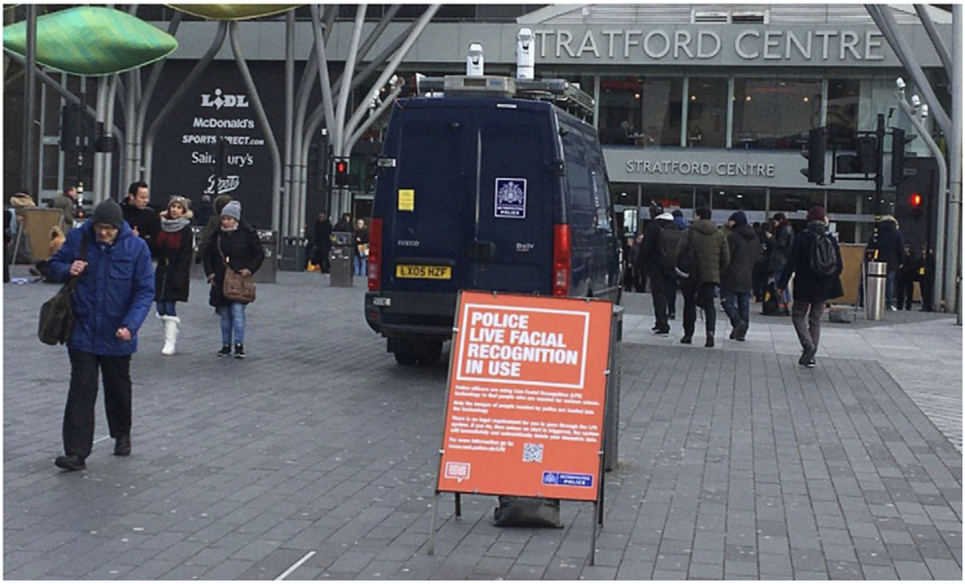

Figure 1. *Police control van and poster situated outside the Stratford Centre, where one of the London trials was conducted.*

size (from around 250 individuals to over 2000, including those wanted for violent offences) in the last few trials as a key contributor to the increased number of successful arrests (8 out of 9 arrests in total). (National Physical Laboratory, 2020, 3) The increase was particularly notable in high-traffic locations like central London, where the watchlist was not geographically filtered.

Given that these FRT deployments resulted in the apprehension and arrest of a number of people whose faces were matched to watchlists, their description as mere ‘trials’ by the London MPS is seriously questionable. Although described as ‘trials’ to publicly indicate that their use on these occasions did not necessarily reflect a decision to adopt and deploy FRT on a permanent basis, they were decidedly ‘real’ in the legal and social consequences for those whose faces triggered a match alert. Because it was not feasible to obtain informed consent of all members of the public who passed in front of the cameras given the unconstrained settings in which they were sited, the police purported to rely upon their common law powers as the legal basis for these deployments, which has been expressed as the power of a police officer to ‘take all steps which appear to [be]…necessary for keeping the peace, for preventing crime or for protecting property from criminal injury’ (*Rice v Connolly* [1966] 2 QB 414, 419 per Lord Parker CJ; MPS 2021). Accordingly, the London trials resemble an active police ‘operation’ conducted on an experimental basis rather than a more modest test of FRT’s ‘functional performance’ in real-world settings (discussed below).

Across all ten trials, the MPS engaged individuals in response to alerts on a total of 27 occasions, with 10 alerts confirmed as correct identification matches following field officer engagement, resulting in 9 arrests. Of these, only two arrests were made during the earlier trials until the watchlist size was substantially enlarged, and a total of approximately 18,260 faces were scanned across the ten operational trials. For the first few London trials, passers-by had no means to avoid passing in front of the surveillance cameras until this was insisted upon by the Information Commissioner and civil society organisations (Fussey and Murray, 2019; National Physical Laboratory, 2020). Although signs and leaflets were made available at the trial locations during the FRT deployments, the MPS did not seek to engage with and consult the public about its intention to conduct these trials, nor were affected local communities given an opportunity to influence the trial design or conduct (Fussey and Murray, 2019). The independent evaluation conducted by Essex University researchers criticised the trials for their lack of transparency, engagement with local communities, and potential privacy violations (Fussey and Murray, 2019; Fussey and Murray, 2020), whereas the National Physical Laboratory (NPL) report (2020), commissioned by the MPS, highlighted the effectiveness of the technology in improving arrest rates, concluding that live FRT was a valuable tool in law enforcement. After the trials conducted, the MPS subsequently adopted facial recognition technology for use in public spaces on a permanent basis in January 2020, including its controversial deployment during the coronation of King Charles.

3.2.2. Wales, United Kingdom

Between 2017 and 2020, 69 ‘operational’ trials were conducted by the South Wales Police (SWP). SWP also sought to rely on common law police powers rather than seeking informed consent from all affected individuals. It also used NEC’s NeoFace software, compiling bespoke watchlists populated with the facial images of various wanted individuals rather than those of volunteers. However, the Welsh trials differed from the London trials in several ways: the SWP also sought to test the FRT software’s one-to-one verification functions, as well as its ‘discovery’ function (Davies et al., 2018, 30 et seq.), and many trials occurred in settings involving large, uncontrolled crowds, including major football matches, boat races, music concerts, royal visits, and, controversially, a peaceful anti-arms protest (see Figure 2). Unlike the London trials, no specified technical or other purposes are identified in the independent evaluation report by Cardiff University researchers, other than vague claims that they were intended to ‘test overall utility’. Although the SWP stated that one of its objectives was to test the impact of live FRT on crime prevention, the Cardiff researchers indicated that they had been unable to quantify what, if any, impact the technology had on crime prevention (2018, 43). These trials were subject to a judicial review challenge by privacy campaigner Edward Bridges (supported by Liberty, a human rights advocacy group), which ultimately resulted in the Court of Appeal finding that the trials were unlawful due to several legal deficiencies relating to data protection law, equality law and human rights law (*R (Bridges) v Chief Constable of the South Wales Police* [2020] EWCA Civ 1058) some of which are discussed in Section 4.4 below. However, the Court refused to pronounce on the general legality of FRT deployment by the police, which has subsequently taken place. This has led to diverse interpretations of this landmark judgement among academics, government ministers, and practising lawyers. For instance, evidence gathered by the Ryder Review reveals significantly diverging views: Some believe that the Court of Appeal deemed the use of

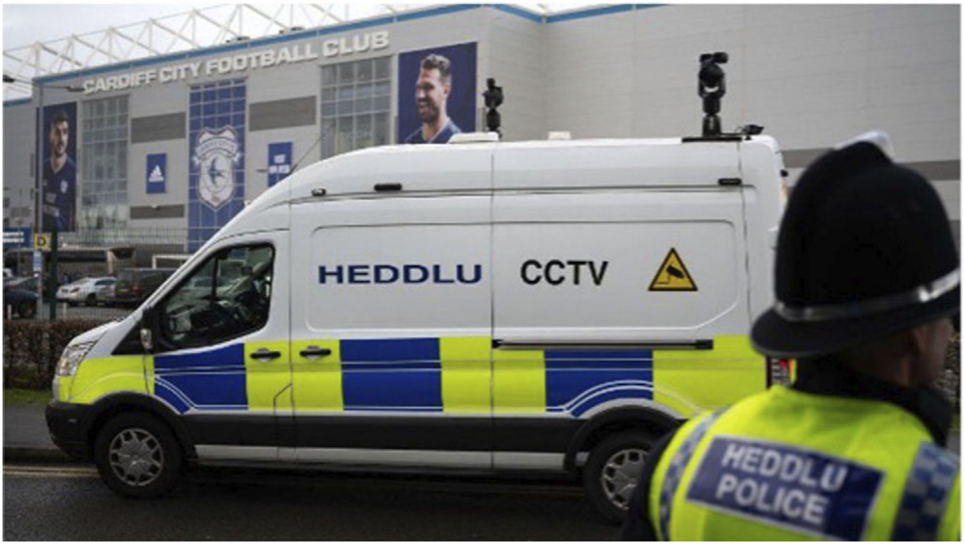

Figure 2. *Police van involved in one of the Welsh trials conducted outside Cardiff City Football Club.*

https://doi.org/10.1017/dap.2025.10019 Published online by Cambridge University Press

live FRT lawful, despite finding that the exercise of common-law power was not ‘in accordance with law’ due to an inadequate legal framework, while others argue that, in spite of the deficiencies of the legal framework coupled with insufficient impact assessment, the Court of Appeal did not conclude that SWP had ‘broken’ any laws, thereby suggesting the use of live facial recognition (live FRT) was lawful (Ryder, 2022, 53-54).

One striking feature of the Welsh trials concerned the extraordinarily wide, vague, and changing nature of the FRT watchlists and the casual approach taken to the type, quality, and source of facial images used to populate them. As a result, independent evaluators performed an additional controlled field trial to test the impact of image quality, which resulted in the police re-adjusting the composition of their watchlists from time to time (Davies et al., 2018, 36 et seq.). In addition to police custody images, the SWP watchlists included low-quality images scraped from social media, which is reported to have significantly affected overall system performance (Davies et al., 2018, 18, 38). However, because Edward Bridges was not included in any of the watchlists, longstanding questions concerning the legality of police retention and use of custody images (including images of individuals who were not subsequently charged or convicted of any crime) and images from other sources were not contested in court. Even the Cardiff researchers evaluating the trials (Davies et al., 2018, 43) stated that they were unable to identify precisely how and why the size and composition of the watchlist mattered, although ‘it does seem to make a difference’ while recommending that ‘decisions about watchlist size and composition’ be made available for public scrutiny. Thus, the independent evaluation report concluded with a call for ‘realistic’ and evidence-led approach to the evaluation of FRT in policing, highlighting the need to dispel myths and counter misinformation that has seeped into public and policy discussions, often fuelled by ‘influential, fictional portrayals of [FRTs] in films and television’ (Davies et al., 2018, 43).

3.2.3. Berlin, Germany

Live FRT trials were undertaken in Berlin between August 2017 and July 2018 in a partnership between the German national rail operator Deutsche Bahn AG (DB) and the German Federal Police (Bundespolizei or BPOL), backed by the German Federal Ministry of the Interior and supported by the German Ministry of Education and Research, which was then devoting serious investment to convert ‘traditional surveillance tools’ (such as fixed-site CCTV cameras) into policing tools (Eireiner, 2020). Berlin Sudkreuz train station was selected as the trial site, enabling the use of FRT software with existing video surveillance cameras installed at the station as part of over 6000 CCTV cameras already installed in 900 train stations across Germany. (Eireiner, 2020) The Berlin trials formed part of DB’s wider project of ‘Safety Station’ (German Federal Police, 2018, 9) to promote ‘rail safety’ by testing a wide range of ‘safety security technologies’, including live FRT. Unlike the British trials, the Berlin trials tested and compared FRT software produced by three different manufacturers—including Dell, ELBEX (a German security service provider), and L-1 Identity Solutions AG (a German software company). The trial was undertaken in two phases: the first aimed at evaluating the potential benefits of live FRT for policing purposes, and the second to test the software’s capacity to detect suspicious behaviours (e.g., acts of violence), stray objects, and dangerous situations (e.g., individuals in distress). Hundreds of volunteers (312 for phase 1 and 201 for phase 2) were recruited to participate either for a six-month or one-year trial period, and their biometric data were collected to build a database. All participants were given a 25-euro Amazon gift card, and a few of them who crossed through the most received Apple watches as an additional incentive.

Although the German Federal Commissioner for Data Protection (Andrea Vosshoff) approved the trial on condition that participants provided voluntary informed consent (Chase, 2017), controversies arose following reports that some were not being performed in a closed, isolated environment and volunteers were mixed with other ordinary passengers who may or may not have been aware of, nor intentionally participated, in the trial (Eireiner, 2020). To inform ordinary passengers that the trials were taking place, blue stickers and signs were placed in three different areas, indicating how they could opt out, i.e., by choosing to pass through the white stickered area rather than the blue (see Figure 3). Critics have therefore referred to the Berlin trials as ‘trials in hiding’ (Eireiner, 2020), alluding to the opacity of both the software

https://doi.org/10.1017/dap.2025.10019 Published online by Cambridge University Press

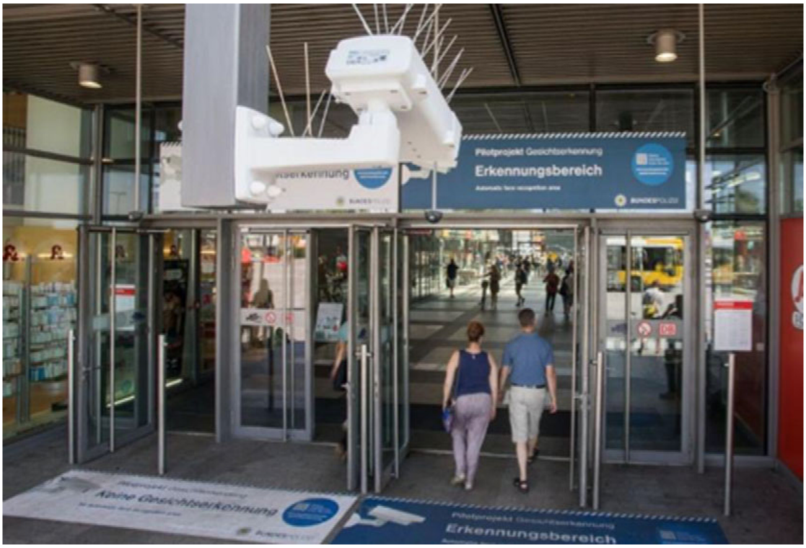

Figure 3. *The blue and white stickers outside Berlin Sudkreuz station notifying those entering of ongoing live FRT trials inside the station.*

and the conduct of the trials. Critics claim that passengers whose faces were parsed by the system were ‘research participants’ from whom informed consent should have been sought. Others questioned the adequacy of consent obtained from volunteers once it was later revealed that volunteers were not informed of the full range of personal data collected about them, including the speed and temperature recorded by the transmitters given to each volunteer to identify the accuracy of any alerts (or failure to generate an alert). In January 2020, the German Ministry of Interior decided not to install face recognition software immediately at German train stations and airports, deciding instead to increase the number of CCTV cameras at train stations and other public gathering spaces (Eireiner, 2020). The evaluation by the German Federal Police concludes that the facial recognition technology available on the market holds potential value but, prior to its full integration into police operations, further steps for implementation must be considered, including determining the exact areas of responsibility and defining the databases or files with which it should be integrated to ensure optimal usage (German Federal Police, 2018, 7).

3.2.4. Nice, France

Several FRT initiatives in France have attempted to test FRT for identity-checking in various settings, including schools, airports, train stations, and more recently, the Paris Olympics (Pascu, 2020, Laanstra-Corn & Sewell, 2024). At the time the primary data for this paper were collected, the Nice trial was the only French trial performed in an open public environment, operated as part of a broader PIAVE project (Industrial Project of the Future), and related to several EU-funded projects. (Estrosi, 2019, 4). It was selected as the trial site because Nice now has the largest number of public CCTV cameras in France following the deadly terrorist attack in 2016 (Connexion, 2019a). The trial, lasting for three days during the 135th Nice Carnival in 2019 (16, 19–20 February 2019), was moderate in scale, conducted at four adjacent security-check gates (pictured below in Figure 4) where six cameras were installed at the E4 entrance to the Carnival, in which crowd flow was partially constrained. Nice police partnered with the Monaco-based cybersecurity company *Confidentia*, which provided its ‘Any Vision’ software free of charge to enable its algorithms to be tested for their accuracy (Kayali, 2019). The stated purpose of the trial was to test, in a live and retrospective manner, the ‘relevance of such a tool at the service of citizens of their well-being and safety’(Estrosi, 2019, 4). More specific test objectives are expressed in the Nice report, including testing the ability of the software to detect an unauthorised person entering the area, one-to-one

https://doi.org/10.1017/dap.2025.10019 Published online by Cambridge University Press

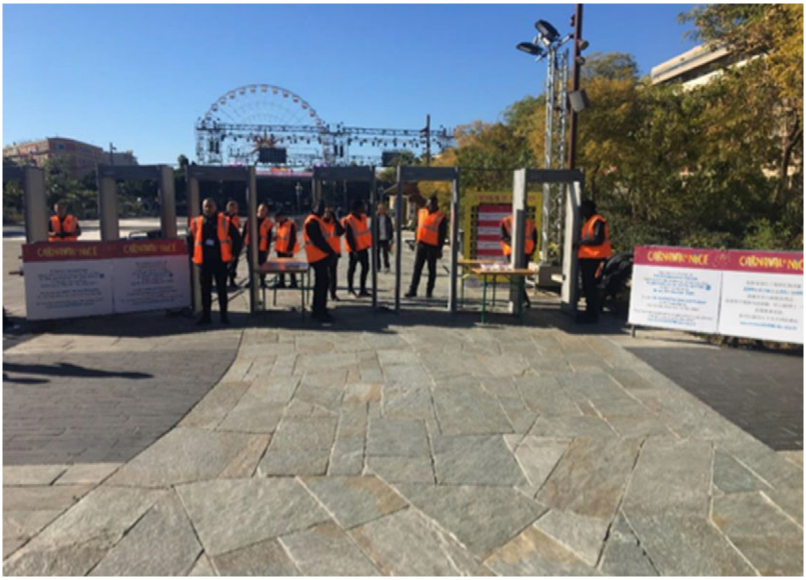

Figure 4. *The setting of the E4 entrance of the 135th Nice Carnival for the live FRT trials.*

identity verification, and one-to-many 'discovery' purposes. Specific scenarios were simulated using volunteers 'with no operational implications' (according to Commission nationale de l'informatique et des libertés (CNIL, 2019), including finding lost children and vulnerable persons, enabling more fluid crowd flow through access points, restricting access control, and arresting persons wanted by police, among others (Estrosi, 2019, 30).

The Nice trial was primarily based on volunteer participants who consented to the uploading of their facial image to the FRT watchlist. Other members of the public were also given the opportunity to participate by choosing to enter the area where live facial recognition technologies were being deployed, having been informed by public signage at the E4 entry into the Carnival arena that the trial was taking place. (European Union Agency for Fundamental Rights, 2019, 12). The size of the volunteer sample for one-to-one verification purposes consisted of only eight volunteers who submitted their photos for use in the watchlist, including a person wearing a disguise and another whose photo was 40 years old (Untersinger, 2019; Connexion, 2019a). Despite this small number of volunteers, the Nice evaluation report heralded the trial as a 'success'. Prior to the trial, a data protection impact analysis was performed and submitted to the French data protection authority CNIL, and a meeting with CNIL's 'Régalien et Conformité' department was convened (Estrosi, 2019, 5). In the absence of legislation or guidelines explicitly authorising the performance of live trials in public settings at scale, informed consent was relied upon as the legal and ethical basis for the trial (Estrosi, 2019, 4). Plans to test the FRT in operational contexts across France over a longer period of time (6 months to 1 year) were subsequently announced following the Nice trial (Connexion, 2019b). As the Mayor of Nice, Christian Estrosi, concludes in the report, the testing of FRT in a real-world setting was considered a success despite its limited scale, and emphasis is placed upon the need for a legislative framework to support future deployments and to balance security needs with the protection of individual rights (Estrosi, 2019, 30).

3.3. A comparative analysis of live FRT testing by LEAs

The live FRT trials by LEAs described above provide a rich set of real-world case studies that help to illuminate the benefits, challenges, and dangers associated with in-the-wild AI testing, while highlighting the current lack of clear governing norms or conventions to ensure that these activities are 'responsibly' undertaken. In the following discussion, we consider whether these trials were undertaken in an

https://doi.org/10.1017/dap.2025.10019 Published online by Cambridge University Press

epistemically, legally, and ethically responsible manner. We begin with some general observations concerning how these live FRT tests were designed and conducted and their animating purpose. First, these trials varied significantly across many dimensions, including their purpose, design, physical settings, size, scale, duration, methods, practices, interaction with individuals, and processes of human intervention and evaluation. Second, as 'trials' of new technologies, these tests were difficult to characterise along conventional lines. Each trial was undertaken voluntarily by the LEA on its own initiative. They were not undertaken as an academic or scientific research project (although it was a condition of the funding support for the London trials that they were subject to independent academic evaluation) (Merton, 1973). Although some AI-enabled robotic technologies are currently being extensively tested 'in the wild' for commercial development, including 'self-driving' vehicles (Stilgoe, 2018; Marres and Stark, 2020), these live FRT tests were not undertaken as research and development ('R&D') activities by software developers but by LEAs interested in testing whether commercially produced software might help them carry out their law enforcement functions (Rooksby et al. 2009). The LEAs sponsoring these tests had not yet purchased the technology, nor committed to deploying it on an ongoing basis, although in each case the sponsoring organisation relied upon commercially developed software trained on closed data-sets, for which neither the algorithms nor the training data used in their creation were made available. Accordingly, these trials appear to have been undertaken primarily by LEAs for the purposes of acquiring first-hand evidence concerning whether, and to what the extent, the technology could enhance their ability to carry out their law enforcement duties and functions in 'live' settings and, in turn, to help inform decisions about whether to permanently adopt the technology for policing purposes.

The following analysis proceeds on the basis that these live FRT trials were undertaken by the sponsoring LEA with this purpose in mind, seeking to identify the extent to which these trials were epistemically, legally, and ethically responsible.

3.3.1. Epistemic integrity and responsibility

As we have already explained (see Section 1), the epistemic integrity of any test will depend upon its intended purpose—particularly the question(s) which the test seeks to answer—which should inform how the test is designed, the methods and conditions under which it is conducted, the nature and quality of the evidence which the test is expected to produce, and the criteria for evaluating the test's 'success', all of which which should be specified in advance.

(a) Identifying the test object and objective?

Although we were not able to comprehensively identify publicly accessible technical documentation identifying the purpose, design, and/or specifications of each of the live FRT tests described above, it is possible to infer their purposes from the way in which they were carried out.

(a) Testing software for 'in domain' functional performance

One of the most striking differences in test design and testing methods turns on whether the sponsoring LEAs relied on volunteer participants, as per the Berlin and Nice trials, or were 'operational' in nature, and thus sought to identify and apprehend 'real' targets, that is, individuals who were in fact 'wanted' by law enforcement authorities, as per the London and Welsh trials. By relying upon volunteer participants, we may infer that the primary purpose of the Berlin and Nice trials was to test the 'functional performance' of the FRT software in real-world contexts, seeking to measure and evaluate software accuracy in identifying a match between the images of volunteers as they passed in front of an FRT-enabled video camera to the digital image uploaded to the FRT watchlist. As the Stanford Human Computing Laboratory has observed, rather than relying on claims about the accuracy of FRT accuracy obtained in stable, highly controlled lab settings, there is a vital need for 'in domain accuracy' metrics to address ongoing contestation about the relative benefits and dangers associated with using the technology. Hence, the

https://doi.org/10.1017/dap.2025.10019 Published online by Cambridge University Press

Berlin and Nice tests can be understood as a welcome attempt to elicit a more meaningful and realistic assessment of the functional performance of FRT matching in real-world contexts of use, particularly by LEAs in open public settings. By locating the trials at a busy urban train station, findings from the Berlin tests could usefully inform how accurately the technology may be expected to perform in other public train stations, appropriately reflecting one of the trial's motivating objectives in generating useful information to inform a decision whether to deploy live FRT across German train stations. In contrast, the Nice trials (Figure 4 above) were undertaken at a controlled gateway into and out of a public space in which individual one-to-one bag security checks were taking place, thus limiting the extent to which the accuracy findings from this trial could be appropriately applied to dissimilar physical environments.

Reliance upon volunteer participation, rather than 'live targets' was also necessary to generate epistemically reliable data about the software's accuracy in relation to different groups of participants, particularly for those with legally protected characteristics that may be (to some extent) physically observable, such as gender, age, racial origin, and ethnicity. For example, Berlin trial volunteers were issued with a transmitter that was used to locate them as they passed within and around the field site, enabling both true and false positive and negative alerts to be systematically and comprehensively monitored without the need for human intervention in the field. As a result, the design and conduct of the trial could, at least in theory, produce reliable and comprehensive data to calculate the software's 'in domain' matching accuracy. Despite their modest aims, the technical requirements for setting up and conducting these trials were considerable, entailing the successful and ongoing coordination of multiple items of hardware (including cameras to detect and capture images of human faces passing through the camera's field of view), monitoring equipment, a central control site (if used), computer processing capabilities connected to visual display monitors), reliable high-speed network connections, and communication devices configured to gather data and ensure real-time communications between the various technical components to allow continuous monitoring and collection of trial data in addition to FRT matching software, and a suitably compiled 'watchlist' or reference database of images against which facial images of those entering the field site are automatically captured and matched against. The use of volunteer participants meant that the compilation of FRT watchlists was reasonably straightforward, comprising facial images of all and only individual volunteers participating in the trial who had either supplied the image themselves or expressly consented to the capture, processing, and uploading of their facial image to the watchlist for the purposes of the trial.

(b) Testing socio-technical systems for 'operational effectiveness' to serve organisational goals

Unlike the Berlin and Nice trials, the London and Welsh FRT trials were described as 'operational', with the aim of enabling law enforcement officials to identify and apprehend real (or 'live') targets. Hence, the FRT watchlists were populated with facial images of individuals who were, in fact, wanted by law enforcement authorities, reflecting their underlying purpose in seeking to evaluate whether, to what extent, and under what conditions, the technology could bring about specific organisational outcomes which LEAs expected would flow from its use. For LEAs, the primary benefit typically assumed to flow from using live FRT is the enhanced capacity to *identify* and *apprehend* wanted individuals whose facial images are stored on FRT watchlists. Whether it *actually* generates these organisational benefits in real-world settings remains unknown and has not been systematically tested. Accordingly, these 'operational' trials had the potential to produce evidence that could help address this important knowledge gap.

This approach meant, however, that false negatives could not be identified and measured. In other words, the number of individuals whose facial images were included on the FRT watchlist and who passed in front of the camera without triggering a match alert (i.e., the 'ones that got away') could not be detected and measured (Fussey and Murray, 2019, 75). Hence, the results generated from the London and Welsh trials were *not* in fact indicators of software-matching accuracy, for they could only generate data concerning recorded true and false positives (cf National Physical Laboratory, 2020). Moreover, because field officers could not locate and stop every individual whose face had generated an automated alert and

https://doi.org/10.1017/dap.2025.10019 Published online by Cambridge University Press

warranted human investigation in crowded field settings (Davies et al., 2018), even the reported data on true and false positives were not complete and comprehensive. In addition, in the absence of ethnicity information supplied by volunteer participants, the police could not collect reliable ethnicity data, and instead attempted to guess the ethnicity of individuals who were the subject of automated match alerts, which were used to evaluate the system's potential racial bias (cf National Physical Laboratory, 2020, 1, 25). Hence, neither the London nor Cardiff trials were capable of producing epistemically reliable evidence of the 'in domain' accuracy of the facial matching software being trialled.

It is also evident that designing and conducting an 'operational' trial of the larger socio-technical system in which the FRT software was embedded was far more complex, demanding, and resource-intensive than the more limited trials of the 'functional performance' of the software's matching accuracy in live settings that were undertaken in Berlin and Nice. To deploy live FRT to apprehend specific individuals during a real-world policing operation, the generation of a match alert by an FRT system must then be translated into successful on-the-ground action through which field officers can promptly, lawfully, and successfully evaluate whether the matched individual is in fact the person that the authority wishes to apprehend. Not only does this entail the effective operation and coordination of all the technical components required for a more modest 'functional performance' trial, but that technical system must also operate in a coordinated and collaborative fashion with multiple human actors and organisational and social policies, norms and practices capable of operating effectively as a complex socio-technical system to help ensure that any decisions and actions of police officers taken in response to a match alert were made in accordance with law, police codes of ethical practice and organisational policies.

3.3.2. The legal and ethical dimensions of testing new technologies 'in the wild'

The epistemic integrity of in-the-wild technology testing is perhaps the primary normative benchmark for characterising its design and conduct as 'responsible', given that the primary purpose of any test is to produce valid knowledge about the test object. However, if a technology is tested outside highly controlled lab environments, the process and impacts of the test on affected others and the environment may raise legal and ethical concerns. The nature and significance of these concerns will be contingent on the nature and functionality of the test object, the design of the test, the chosen testing site, the number of people potentially affected directly by the test, and the nature and extent of those effects and the test processes themselves. When AI-enabled technologies are tested in-the-wild, those tests are likely to raise a much wider range of legal and ethical concerns if the purpose is to test a socio-technical system for its operational effectiveness in a live setting, in comparison to more modest tests of the software's 'in domain' functional performance. This difference is vividly illustrated when comparing the live trials aimed at evaluating software performance in Berlin and Nice, with the trials of the socio-technical systems for their operational effectiveness in securing the apprehension of live targets in London and Wales. In the latter tests, the generation of a 'match alert' by the FRT system then needed to be translated into on-the-ground action in order to apprehend the individual in question. Yet these trials showed that this task was anything but straightforward while generating significant concerns about whether police officers' engagements with members were lawfully conducted. In particular, utilising live FRT with 'real' targets in open public settings poses a number of substantial legal dangers because it requires sponsoring LEAs to engage in several important yet sensitive and legally fraught activities, five of which are highlighted in the following analysis: (a) the collecting and processing of a person's facial image without their knowledge and/or consent in public spaces, (b) the composition of watchlists, (c) field officer 'adjudication' decisions in response to automated match alerts about whether to stop and confront that individual, (d) the operation of automated identification match alerts with dangers of systematic bias and discrimination, and (e) the provision of public information and consultation with affected stakeholders, including the provision of practical alternatives to allow passers-by to avoid being involved in the trial. Each of these activities and the attendant legal duties and dangers that they pose to the legal rights and interests of affected others is examined more fully below.

https://doi.org/10.1017/dap.2025.10019 Published online by Cambridge University Press

(a) The collection and processing of facial image data without consent in public spaces.

The use of live FRT entails the collection and processing of the facial images of individuals as they pass in front of an FRT-enabled camera, thereby interfering with the privacy and personal data protection rights of all those who had their facial images scanned, collected, and processed without their express consent (EDPB, 2022). Deploying live FRT in open public settings also implicates several other fundamental and legal rights, including rights to freedom of expression and freedom of assembly (Fussey and Murray, 2019, 36–46; Liberty, 2019). Yet the failure of the London and Welsh trial sponsors to properly assess the nature of those rights interferences, and their arguable failure to take effective measures to prevent unlawful interference with those rights, is evident in at least three operational aspects of these trials.

First, whether and to what extent passers-by were informed that a live deployment of FRT was in process and had practicable opportunities to avoid having their facial images captured and processed by the FRT system. In both the Berlin and London trials, the practical opportunities for members of the public to avoid being caught up in the trials were criticised as inadequate (Fussey and Murray, 2019, 102). For example, during the initial London trials, passers-by had no means to avoid passing in front of the surveillance cameras until this was insisted upon by the Information Commissioner and civil society organisations. (Fussey and Murray, 2019; National Physical Laboratory, 2020).

Second, camera avoidance behaviour by members of the public was interpreted by field officers in the London trials as evidence that the technology was working as a 'crime disruption strategy' rather than being understood as legitimate and lawful action by those individuals who did not wish to have their facial image taken and processed. In addition, rather than recognising that individuals were legally entitled to preserve their rights of privacy by refusing voluntarily to offer their biometric information for capture during the London trial deployments, camera avoidance behaviour was also interpreted as grounds for suspicion and hence possible intervention by the police under ordinary police powers (Fussey and Murray, 2019, 101–102; LPEP, 2019, 31–2, 34). However, such interpretations may have been contrary to official policy—e.g., during the Stratford deployments, officers were informed during pre-trial briefings that an individual turning around and refusing to walk past cameras was 'not an indicator of suspicion in itself' as part of an individual's right to privacy (Fussey and Murray, 2019, 101; Fusset et al., 2021; Kotsoglou and Oswald, 2020). This interpretation reflects a basic failure by the LEAs to recognise that individual members of the public are entitled, *as a matter of legal right*, to refuse to allow their faces to be captured by live FRT as part of an operational trial deployment by a law enforcement authority. The right to privacy demands that no adverse inferences can be drawn by public authorities merely on the basis of a desire to avoid having one's facial image captured and matched against a database whose contents are not publicly known, let alone fined for disorderly behaviour (Robinson, 2020).

Third, the London MPS's failure to engage in public consultation or provide the public with adequate information in advance of the trials reflects a failure to ensure proper respect for the legal and democratic rights of individuals and members of the affected local community. The duty to provide adequate prior notice and opportunities to avoid the rights interferences entailed by the FRT deployment are safeguards reinforced by constitutional obligations of transparency and accountability imposed on public authorities generally, and law enforcement authorities in particular, demanding that the public should, at minimum, be notified and informed of the basic information about the proposed trials are met, such as date, location, duration, and purpose, and basic obligations of transparency in public reporting on conduct and outcomes. Given that these trials have all been conducted in an overt manner, thereby avoiding burdensome legal obligations for covert surveillance by LEAs under the Regulation of Investigatory Powers Act of 2000, public information should be widely circulated to affected communities in advance with meaningful opportunities to participate in any decision to engage in such a trial, including decisions about their design, siting, conduct and operation. Particular care should be given to the needs of vulnerable or particularly affected groups. Yet in the London trials, for example, public information was only made available in July 2018 on the MPS's website, even though five trials had taken place before that time and apparently in breach of the Surveillance Camera Commissioner's Code of Practice (Surveillance Camera Commissioner, 2019; 2020; Fussey and Murray, 2019, 63). Although various 'community impact assessments' were conducted prior to the London Stratford test, these did not involve direct engagement with the community, nor did they gather specific views on LFR technology (Fussey and Murray, 2019, 66). They were comprised instead of police-held statistics and general assessments of the area, hardly indicative of local community support for the trials. This failure to consult the affected local community appears directly contrary to the long-established British model of policing known as 'policing by consent', which is widely regarded as an important foundation for establishing and maintaining the political legitimacy of policing practices.

(b) The legality of FRT watchlists

It is clear from judicial rulings by the European Court of Human Rights that visual monitoring of public spaces may interfere with the right to private life protected by Art 8 of the ECHR (recently affirmed in *Glukhin v Russia* ECHR 206 (2023)) and which is legally protected in the United Kingdom under the Human Rights Act 1998 (HRA). In *Bridges*, both the High Court and the Court of Appeal affirmed that the use of live FRT in public settings constitutes an interference with an individual's right to privacy under Article 8(1) and entails the processing of 'sensitive' data under the EU Law Enforcement Directive ([2019] EWHC 2341; [2020] EWCA Civ 1058). Accordingly, if a person's facial image is uploaded to an FRT watchlist by LEAs to identify and apprehend wanted individuals in public spaces, this will invariably interfere with their Article 8 rights. The processing of facial images in this way will only be lawful if (a) the conditions specified under data protection legislation are duly complied with, and (b) any interference with the right to private life under Article 8(1) is either waived by the individual right-holder through the provision of express, informed consent to that interference, or if the interference satisfies the requirements of Article 8(2). For live FRT tests that seek only to identify volunteer participants involved in the trial, the composition and creation of FRT watchlists for the purposes of live FRT testing presents relatively few legal difficulties provided that volunteers are properly informed of the nature and conduct of the trial and expressly consent to the collection, storing and processing of their personal data (including the collection and processing of their facial image data).

In contrast, difficult questions arise when LEAs wish to use FRT technology to identify live targets in public spaces. The *Bridges* case left unresolved questions about the lawful composition of FRT watchlists for use by law enforcement authorities in public, including the significance and seriousness of the underlying 'offence' or the legitimacy of the reason why that person is 'wanted' by the police in order to justify the inclusion of their facial image on a live FRT watchlist used to facilitate their apprehension in public. Under the ECHR, legal interferences with the Art 8(1) right to privacy are permissible, but only if the conditions of Article 8(2) are met, requiring that any interference is (a) 'in accordance with law', (b) pursues a 'legitimate aim', which for Article 8(2) purposes includes measures to protect the 'interests of national security, public safety or the economic well-being of the country, the prevention of disorder or crime, the protection of health or morals, or the protection of the rights and freedoms of others', and (c) is 'necessary in a democratic society'. This third component has been interpreted to require that the proposed interference will be legally permissible only if it satisfies the requirements of *necessity and proportionality.* This is a strict and demanding test, which the composition of the watchlists used in the London trials is unlikely to have met. Fussey and Murray's independent evaluation of the London FRT trials highlight two important considerations that bear upon necessity and proportion: first, the type and seriousness of the offences deemed suitable for watchlist selection and second, whether the individual is wanted on warrant or merely 'wanted by the police' even when based on a low level of intelligence (Fussey and Murray, 2019, 58), particularly given the risk of stigmatisation associated with their use in live FRT deployments.

Although the watchlists used in the London 'operational' trials were populated by individuals wanted on outstanding arrest warrants, they also included images of a much broader, amorphous category of persons including 'individuals not allowed to attend the Notting Hill Carnival', 'individuals whose attendance would pose a risk to the security and safety of the event', 'wanted missing' individuals and children, and even individuals who 'present a risk of harm to themselves and to others' as well as individuals who 'may be at risk or vulnerable' (Davies et al., 2018, 42; Fussey and Murray, 2019, 56;

https://doi.org/10.1017/dap.2025.10019 Published online by Cambridge University Press

National Physical Laboratory, 2020; MPS, 2021). No mention is made in the relevant data protection impact assessments prepared by the sponsoring LEAs concerning the level of seriousness of the underlying offence or other risk posed by the individual (Fussey and Murray, 2019, 57; Bridges, 2020).

The two Welsh deployments challenged in *Bridges* reveal that watchlists included not only those wanted for serious crimes or on arrest warrants but also individuals wanted for summary offence, 'every person suspected of committing a crime in the South Wales area' (High Court judgement, paras 11, 14), 'individuals who may be vulnerable', and 'other persons where intelligence is required' (Court of Appeal, para 123). The vagueness of this final class was found by the Court of Appeal to leave excessive discretion in the hands of the police, a finding that resonates with Fussey and Murray's criticisms of the London trials' underlying presumption that LFR technology was non-intrusive and thus treated as benign, failing to take account of its increased capability, particularly when compared with other forms of overt surveillance such as open street surveillance cameras, with the result that the more intrusive features of LFR were not given sufficient scrutiny and thus raising concerns about human rights compliance (Fussey and Murray, 2019, 60; London Policing Ethics Panel, 2018). Equally troubling are reports that in some of the Welsh trials, images from privately sourced databases were used to populate FRT watchlists, including images scraped from the internet or social media, and thus likely to have contravened data protection law. Note that the potential illegality arises from the use of custody images, particularly those who have been arrested but not charged or convicted. Accordingly, it is more than likely, in our view, that the FRT trial watchlists were compiled in an unlawful manner, although this question is yet to be authoritatively ruled upon in court.

(c) Were the trials legally authorised and conducted in accordance with law?

Beyond the legality of the collection and processing of facial images of passers-by, important legal questions also arise concerning whether the LEAs who engaged in live FRT tests had legal authority to do so, and whether all applicable legal duties had been discharged. In France, for example, the lack of any clear and explicit legal basis precluded the French authorities from undertaking live and large-scale trials, prompting the French government to develop legal and regulatory frameworks to support future new technology trials (Theodore and Trotry, 2020). The informed consent of volunteer participants was also relied upon to establish the legal basis for the Berlin trials. However, critics have challenged the adequacy of the consent process given that rail passengers who had not actively consented to participation in the trial were mixed with volunteers as they passed through the camera's field of vision, and volunteers were not informed of all the types of biometric data and information that were collected from them during the trial. Although the legality of the trials themselves was not challenged in *Bridges*, the Court of Appeal indicated that the common law powers of the police in England and Wales to 'prevent and enforce crime' provided a sufficient legal basis to support the Welsh FRT trials for the purposes of Article 8 of the ECHR. However, it also found that SWP policy concerning the locations in which FRT deployments could take place was 'very broad and without apparent limits', thereby effectively leaving the question of location to the discretion of individual police officers (albeit of rank Silver Commander or above) and therefore failed to meet the Art 8(2) requirement that the interference with the Art 8(2) rights to privacy were 'prescribed by law' (Court of Appeal, para 130).

(d) The Data Protection Impact Assessment

In none of the trials we examined was the explicit purpose of the trial to identify and evaluate the *adverse impacts* associated with deploying live FRT in real-world settings. This was somewhat surprising given the ongoing controversy surrounding its use, particularly by LEAs. However, it is possible that these trials were partly motivated by a desire to demonstrate publicly the value and utility of the technologies to stakeholders and the general public to help establish both the legitimacy of their motives as well as that of the technology itself. (Ryder, 2022, 160) One of the most important and yet poorly understood characteristics of networked digital technologies arises from the threats they pose to individual and

https://doi.org/10.1017/dap.2025.10019 Published online by Cambridge University Press

collective goods and values that are often *intangible* and do not lend themselves to precise, quantifiable measurement or observation, unlike threats to health and safety posed by, for example, new drugs, chemicals, or industrial processes (Yeung, 2019). Accordingly, undertaking material practices of the kind through which new technologies are tested, which pose risks to life, health, and safety, is unlikely to properly illuminate the nature, magnitude, and source of these more intangible dangers. Although some were partially surfaced in the form of 'flashpoints' that occurred during the conduct of these 'operational' trials, it is not surprising that a number of human rights interferences arising from the conduct of the trial only came to light following an assessment by a legal academic in partnership with a sociologist who independently observed and reported on the trials. (Fussey and Murray, 2019, 4) As a result, it is important that (1) attention to the specific *context* in which facial recognition and other biometric technologies generate dangers is required to properly assess whether the proposed use meets the legal principles of necessity and proportion involving careful identification of the affected legal rights, duties, and interests, (Ada Lovelace Institute, 2022) and that (2) this contextual legal assessment must occur *prior* to deployment so that necessary alterations to the design and conduct of the proposed deployment can be made, to help prevent unlawful interference with legal rights and to support LEAs in discharging their legal duties. It is precisely this role which data protection assessments are intended to perform, and which are mandated by European and UK data protection law for any proposed processing of personal data which poses a 'high risk to the rights and freedoms of individuals', and which must be undertaken *prior* to any such processing. Article 35(1) of the Data Protection Act 2018 provides that this assessment must take into account 'the nature, scope, context and purposes of the processing' and include (a) a general description of the envisaged processing operations; (b) an assessment of the risks to the rights and freedoms of data subjects; (c) the measures envisaged to address those risks; (d) safeguards, security measures, and mechanisms to ensure the protection of personal data and to demonstrate compliance.'

As one of us has argued elsewhere, data protection impact assessments (DPIAs) are intended to serve as 'early warning systems' to alert those who propose to collect and process personal data to legal dangers that require mitigation (Yeung and Bygrave, 2022). Although a DPIA was undertaken by SWP prior to the Welsh trials, one of the most important rulings of the Court of Appeal's judgement in *Bridges* was its evaluation of the DPIA as inadequate, and hence the trials themselves were unlawful. In particular, the Court of Appeal concluded that the SWP's assessment that the deployments did *not* interfere with the right to privacy protected under Art 8(1) ECHR was erroneous. The Court also found that, because the deployments gave excessive discretion to individual police officers concerning the composition of watchlists and location of the FRT deployments, the way the live FRT trials were conducted constituted an unlawful violation of the Art 8(1) right to privacy of all those whose facial images were captured and processed during the trial and thus the trials had not been carried out in a lawful manner. Our own analysis of the DPIA conducted by the London MPS prior to conducting live FRT tests indicates that its assessment was poorly done, reflecting legitimate fears that DPIAs may in practice be little more than a perfunctory, box-ticking exercise, exhibiting what Michael Power refers to as 'rituals of verification' (Power, 1997). Hence the Court of Appeal's ruling sets a welcome and valuable precedent, demonstrating that the data protection impact assessment, mandated under contemporary European and UK data protection laws for proposed 'high risk' processing of personal data, can play an important and indispensable role by serving as the institutional vehicle through which the kind of careful, context-specific evaluation of legal rights affected by a proposed deployment of intrusive biometric technologies must—as a matter of legal obligation and not merely good ethical practice—be undertaken *before* it is used, but this potential is yet to be realised.

(e) Automated match alerts and the legality of field officer interventions

Additional legal and ethical concerns arise in relation to 'adjudication' decisions and any ensuing interventions taken by police officers upon receipt of an automated FRT match alert. Automated match alerts are based on probabilistic estimates and thus are necessarily prone to error because 100% match accuracy cannot be guaranteed. Accordingly, any attempt by police officers to stop an individual whose

https://doi.org/10.1017/dap.2025.10019 Published online by Cambridge University Press

face has generated a match alert cannot be certain that the individual is in fact the person whose image is stored in the watchlist. Both the London and Welsh trials resulted in several hostile or otherwise unwelcome encounters between field officers and members of the public, particularly when police interventions followed erroneous alerts or the individuals were unaware of, or unhappy about, the use of live FRT itself or police intervention as they were going about their lawful business in public. For example, in London, this included an incident highlighted by Liberty in which five plain-clothed police officers descended on a 14-year-old black student in school uniform as he was walking down the street with his mother, who led him by the wrists to a side street, whereupon the alert was identified as erroneous. The incident was intensely distressing for him and incited an angry response from his mother (Fussey and Murray, 2019, 124). On another occasion, a male pedestrian who sought to cover his face to avoid being filmed by the cameras during the Romford trial was issued with a criminal fine for disorderly behaviour, an incident captured on video by a passerby (Robinson, 2020). Intimidating encounters between police officers and members of the public of this kind are likely to damage public trust in the police, including the deployment of live FRT in open public settings (Bradford et al., 2020).

These encounters underline the importance of careful attention to the configuration of alert thresholds for FRT matching software. The lower the threshold for generating a match alert, the higher the number of false positives: this means that when in operation, this enlarges the number of individuals whose faces will generate an automated alert indicating a match to the FRT watchlist. This, in turn, increases the number of in-the-field adjudications that must be made before then seeking to apprehend a wanted individual, while widening the net of individuals whom field officers may seek to stop as they go about their daily activities within the field site. Although the matter was not raised in *Bridges*, the setting of these thresholds necessarily implicates the necessity and proportion of the police use of FRT and therefore its legality, particularly its impacts on rights to privacy, to freedom of expression, and freedom of assembly. As a matter of British law, in the absence of evidence of reasonable suspicion, police officers are not legally entitled to stop and detain individuals. In our view, the mere generation of an FRT match 'alert', in and of itself, does not constitute reasonable evidence of suspicion, which is a precondition for the power of police to lawfully stop and detain an individual for questioning about a suspected crime. Hence, a police officer in receipt of an automated FRT alert would not, on that basis alone, be legally authorised to stop and detain individuals who are otherwise going about their lawful business, particularly given that the match may be erroneous. Although police officers in England and Wales are entitled to stop individuals and ask them questions about who they are and what they are doing, individuals are not obliged to answer these questions in the absence of reasonable suspicion that they have been involved in the commission of a crime. Accordingly, any initial attempt by police officers to stop and question an individual whose face is matched to the watchlist must be undertaken on the basis that the individual is not legally obliged to cooperate for that reason alone.

The probabilistic nature of FRT match alerts and the need for human intervention by a police officer to evaluate the validity of a match alert also highlights the need for clear organisational policy, operational protocols, and proper officer training to help ensure that any police intervention with an individual is undertaken in a lawful and respectful manner. In particular, because such interventions are likely to be regarded by individuals as unwelcome, alarming, and potentially stigmatising, it is critically important that these policies and protocols make it clear that the legal basis for coercive police intervention is *not* established on the basis of an FRT match alert alone. Accordingly, respect and sensitivity by field officers must be demonstrated when approaching individuals to confirm the accuracy of any alert, recognising that individuals are not obliged to cooperate and must be treated accordingly. In other words, participating police officers must bear firmly in mind the limits of their legal powers and the imperfect nature of the technology. In contrast, independent evaluators of the London trials expressed concern that field officers often appeared to proceed on the basis of a 'presumption of intervention' in response to automated alerts, while emphasising the importance of human adjudication, with 26 LFR matches during one London trials deemed sufficiently credible to intervene with a person of interest but only 8 of these were classed as correct once a human identity check had occurred (Fussey and Murray, 2019, 74).

(f) Compliance with equality laws and the danger of unlawful discrimination

Additional legal concerns arose in relation to potential bias in the operation of the FRT software employed (Buolamwini and Gebru, 2018; Castelvecchi, 2020; Leslie, 2020). In each of the trials we examined, FRT matching software produced by commercial developers was used, preventing the sponsoring LEAs from evaluating the underlying datasets upon which the software had been trained, and severely restricted their capacity to alter its operative parameters. However, because software's performance, including its accuracy, is a product of its underlying training data, having access to the training data is necessary for the organisation to evaluate the quality and rigour of the training process. In *Bridges*, the Court of Appeal ruled that the South Wales police could not rely on unsubstantiated assertions by the developer that the software had been trained on datasets that were sufficiently diverse without supporting evidence, resulting in SWP's failure to discharge its public sector equality duty arising under s.149 of the Equality Act 2010. In other words, the inability of the South Wales Police to access the training data to satisfy itself that the software did not unacceptably discriminate between persons of different gender, age, and ethnicity *directly* resulted in SWP's inability to meet its statutory obligations.

3.4. Trial monitoring, performance evaluation, and governance

The preceding discussion underlines the importance of developing and putting in place policies and protocols when designing a live technology trial, particularly when seeking to test legally and ethically sensitive technologies that entail the collection and processing of biometric data. Trial protocols, policies, and training for those involved in the conduct of the trial are essential, not only to prevent and mitigate each of the legal dangers referred to above but also to ensure the proper governance and oversight of the trial itself. Without proper governance to ensure that trial protocols are adhered to in the conduct of the trial, the trial sponsor cannot demonstrate that the evidence produced by the test is epistemically valid, nor that the trial was undertaken in a legally and ethically responsible manner. Yet our analysis of these live FRT trials indicated that the trials may have proceeded without clear governance and oversight, heightening the probability of unlawful and unethical practices and posing risks to the epistemic integrity of the trials themselves. We highlight two matters: (a) the need for coherence between the trial purpose and its design and conduct, and (b) pre-specification of 'success' thresholds, trial metrics, inter-trial comparison, and trial evaluation.

Firstly, except for the Berlin trial, little sustained attention was given to ensuring that the trial setting, design, and conduct corresponded to the purpose of the trial and could therefore be expected to generate reliable evidence that would address the questions which the trial sought to answer. For example, Fussey and Murray's independent review noted that one FRT test deployment took place in Westfield Shopping Centre, yet local intelligence identified high concentrations of offences at a location some distance away, entirely distinct from the shopping mall. Although senior local officers indicated that a significant proportion of criminality of concern was 'driven by homeless people', very few homeless people frequented the Shopping Centre where LFR was deployed, given that it is a largely private space and subject to stringent place management administration (Fussey and Murray, 2019, 90). On this basis, it seems unlikely that the siting and design of the trial could generate reliable evidence of its contribution to the trial's stated aim of reducing crime, which officers believed was 'driven by homeless people', throwing doubt on the epistemic validity of the trial findings.

Our analysis also indicates that their sponsoring LEAs failed to specify in advance the metrics and thresholds that would be adopted to evaluate the success or otherwise of the trial, and a lack of clear standardised metrics and methods to evaluate FRT's functional performance or operational effectiveness. For example, during the London trials, various incremental changes and revisions were made to the trial parameters, particularly changes in the size and composition of the watchlist (ranging from 42 individuals through to 2401) (National Physical Laboratory, 2020, 3), and police practices in relation to watchlist criteria (Fussey and Murray, 73-85) reducing the ability to make meaningful comparisons between trials, and how the evidence generated by the trials as a whole should be interpreted and by whom. A comparison of the European and UK trial evaluations indicates that different parameters have been employed, some of which have not been published. While attempts were made to count the number of false positives across all the trials, there is substantial disagreement concerning how exactly these outcomes are defined and formulated, and in the case of the London and Welsh trials, no false negative data were available, so that the accuracy of automated alerts could not be assessed. In a newer report, the National Physical Laboratory (2023) concluded with the best configuration (setting the ‘face-match threshold’ to 0.6) of the Neoface v4 facial recognition software that could achieve ‘equitable’ outcomes across gender and ethnicity and reduce the likelihood of false positives (one in 6,000 being falsely matched when using a watchlist of 10,000 facial images, and one in 60,000 when using a watchlist of 1,000 images). Again, the rate of false negatives was not part of the evaluation. Although the Berlin tests were largely concerned with testing the functional performance of FRT matching software, the German Ministry of Interior did not specify what constituted ‘efficiency’ (Reuter, 2017) subsequently announcing that the pilot was a success with an accuracy of 80%, but in fact, the success rate ranges between 12% and 65.8% (Eireiner, 2020).

Difficulties in specifying meaningful performance metrics for evaluation are particularly pronounced in relation to the non-technical dimensions of the trial, including both its purposes and benchmarks for evaluating success or failure. For example, Fussey and Murray’s evaluation of the London trials noted that the detection of wanted individuals was always stated as the central purpose of these trials, yet this was always supplemented by additional claims of ancillary benefits, including the disruption of crime, crime displacement, and deterrence effects (Fussey and Murray, 2019, 88). Yet these purposes are quite different, raising distinct questions about how claims of effectiveness would be evidenced. In this respect, Fussey and Murray noted that the evaluations of test deployments have focused on the technological performance of LFR systems, yet the same standard of analysis is not applied to other areas of claimed effectiveness (Fussey and Murray, 2019, 88). For example, they note that deterrence effects are often difficult to substantiate, even when apparent, and although LFR was justified on the basis of offering ‘reassurance to the community by providing a visible symbol that crime was being tackled’, no evidence was provided to support such claims that LFR had a positive effect in allaying the fear of crime (Fussey and Murray, 2019, 88). This bold assertion is seriously problematic, particularly given the long and notoriously complex debates about the accurate measurement of public fear of crime in the criminological literature. In this respect, we concur with Fussey and Murray’s claim (2019, 89) that ‘when accounting for these complexities as a whole, it is extremely difficult to claim public benefit and support without a high standard of evidence’.

The absence of pre-specified determination of ‘success’ thresholds significantly undermined the epistemic basis for evaluating the trials as ‘successful’ or otherwise. For example, although the use of operational live FRT trials may be intended to test the system’s ability to facilitate the effective identification and apprehension of real persons of interest to the police, both the London and Welsh trials relied upon a large cadre of participating field officers tasked with locating and stopping the identified individuals to assess whether the FRT match alert was accurate, and to undertake any further intervention for correctly identified persons where legal grounds for arrest were established. Yet in the ordinary course of day-to-day policing duties, there would not typically be large numbers of police in the immediate vicinity who could be immediately called upon to intervene and apprehend the person of interest. (Davies et al., 2018, 21) Nor, as previously indicated, was there any way of identifying and measuring the number of false negatives (i.e., the ‘ones that got away’). Accordingly, it is questionable whether even operational trials provide reliable evidence of the extent to which live FRT can enhance LEAs’ ability to effectively and efficiently identify and apprehend wanted individuals in everyday policing contexts.

Perhaps the most vivid example of a lack of rigour in trial evaluation is evident in the claims made in an MPS-commissioned evaluation report undertaken by the National Physical Laboratory (NPL)—the UK’s national measurement standards laboratory (National Physical Laboratory, 2020). For example, it compares police use of FRT to conventional police ‘manhunts’ and counts the number of arrests made following the use of FRT across the London MPS trials to conclude that it offers a ‘favourable comparison’

https://doi.org/10.1017/dap.2025.10019 Published online by Cambridge University Press

in terms of resource use. This is done without *any* attempt to quantify the cost of the infrastructure, hardware, and other technological components of an FRT system (which are very minor in relation to a conventional manhunt) and technological components of an FRT system, even as it overlooks the fact that ‘manhunts’ are intended to identify and apprehend a single suspected very dangerous offender rather than a collection of unrelated individuals who are of interest to police for a variety of reasons, many of whom had no connection whatsoever to suspected involvement in serious crime (Trigg and Milner, 2023). The NPL report also makes the bold statement that ‘the trials indicate that LFR will help the MPS stop dangerous people and make London safer’ without offering any concrete evidence concerning how this conclusion is arrived at (National Physical Laboratory, 2019, 4). Even more troubling is the NPL’s tables of numerical data intended to display the technical performance of the system in each of the trials, yet it provides details about false and true positives without any mention of false negatives, presumably because this data could not be collected given that the tests were designed to identify and apprehend ‘live’ targets. Yet despite the absence of this crucial data, nor a clear statement of the criteria used to identify success or failure, the NPL report concludes that ‘the trial indicates that LFR will be an effective policing tool that helps the MPS stop dangerous offenders and make London a safer place’ (National Physical Laboratory, 2020, 28). The basis upon which this conclusion was reached, therefore, remains methodologically indefensible and unjustified by the publicly available evidence.

4. Discussion: Responsible technology trials ‘in the wild’

Before concluding, we reflect on our findings in the context of broader discussions about emerging technology governance, briefly and selectively comparing and contrasting the testing and evaluation of new drugs before they can be publicly released with the current approach to AI testing and evaluation including s 60 of the EU AI Act concerning real-world testing for ‘high-risk’ Annex III AI systems. Whilst applied ethics scholars have argued that new technologies should be introduced both in an *epistemologically responsible* sense, requiring experiments and trials to be set up and undertaken in manner likely to produce knowledge that is reliable, robust, and relevant (van de Poel, 2016, 671), and secondly an *ethical* sense, due to the adverse potential impacts of such experiments on society (Van de Poel, 2017, 83), our analysis highlights the need to attend to the need to ensure that technology testing is undertaken *in accordance with law*, and for effective *institutional frameworks* to ensure accountability and oversight of the trials themselves, particularly when undertaken ‘in the wild’.

The live FRT technology trials analysed above were not legally mandated. Instead, they were undertaken voluntarily by LEAs to acquire first-hand evidence concerning whether the technology would provide the expected level of functional performance (Berlin, Nice) or substantially enhance their ability to find and apprehend wanted individuals in open public settings (London, South Wales) in the service of organisational objectives. To the extent that these LEAs were motivated by scepticism about the validity of marketing claims made about the performance and benefits of AI systems made by software vendors, this is a welcome development. As distinguished statistician David Spiegelhalter argues, when confronted by an algorithm, we should expect trustworthy claims about the system—what the developers say it can do, and how it has been evaluated. Despite the proliferation of various ‘checklists’ oriented towards ‘ethical AI’, we share his concern that contemporary discussions about the ethics of AI have taken for granted that algorithms will be beneficial when implemented (Spiegelhalter, 2020, 5): One of the motivating reasons underpinning our investigations was a concern about the lack of reliable evidence to demonstrate that the impact of networked digital technologies (particularly AI) for public sector use has been significantly positive. (Misuraca and van Noordt, 2020)

To this end, Spiegelhalter’s comparative analysis of the process of structured evaluation (see the first two columns of Table 1), which has been familiar to statisticians for decades (and mandated by law for new pharmaceuticals), and the way in which algorithms are currently evaluated, is particularly illuminating. He notes that in nearly all the published literature on medical and policing algorithms, these evaluations have focused almost exclusively on Phase I—claimed accuracy on digital datasets, with few a small number of Phase II tests (involving comparisons between the algorithmic judgement and expert

https://doi.org/10.1017/dap.2025.10019 Published online by Cambridge University Press

Table 1. *Accepted phased evaluation structure for pharmaceuticals (columns 1 and 2) and a proposed analogous structure for evaluation of algorithmic tools (columns 3 and 4)*

Phase	Pharmaceuticals	FRT	Generalisability to other AI systems
Phase I	Safety: initial testing on human subjects	Technical accuracy: initial testing in controlled laboratory settings on curated datasets, assessing false positives/negatives and error rates	Technical performance: evaluating accuracy, robustness, cybersecurity, and bias in simulated conditions within controlled laboratory settings
Phase II	Proof-of-concept: estimating efficacy and optimal use on selected subjects	Real-world testing for FRT accuracy on volunteer participants in controlled field settings to evaluate bias, demographic performance variations, and operational constraints	In-domain testing: evaluating an algorithm's fitness for its intended deployment purpose and context, accuracy, bias and robustness, and possible unintended adverse effects in controlled field environments
Phase III	Randomized Controlled Trials: comparison against existing treatment in a clinical setting	Real-world trials: deployment in controlled field settings (including public spaces) with monitoring for functional performance and operational effectiveness (false positives, human-machine interaction testing, and monitoring for unintended (adverse) effects	Controlled field trials: testing AI in closely supervised real-world environments (including public spaces) to monitor performance across a range of variables (accuracy, bias, cybersecurity, robustness) as well as user interaction, effectiveness of risk controls, and adverse impact testing and monitoring
Phase IV	Post-marketing surveillance, for long-term side effects	Post-deployment monitoring and evaluation, including regular evaluation of system drift, regulatory compliance audits, and review and updating to evaluate and address emerging and new risks	Post-market monitoring throughout the lifecycle to ensure traceability, accountability, responsibility, compliance with evolving laws, incident reporting, including periodic review and updating

Source: Adapted by authors based on Spiegelhalter, 2020.

human judgement) and 'very few' Phase III evaluations to check whether a system in practice does more good than harm (Spiegelhalter, 2020, 6). Yet without explicit evidence to demonstrate that algorithms shown to have reasonable accuracy in the laboratory actually generate real-world benefits in practice, the trustworthiness of claims made about the benefits and value of algorithmic technologies are difficult to sustain. More extensive comparative evaluation between the testing and evaluation of new drugs and that of novel AI technologies before they can be made publicly available is beyond the scope of this paper but constitutes a fertile site for future research and investigation.

https://doi.org/10.1017/dap.2025.10019 Published online by Cambridge University Press

Spiegelhalter's four-phase structure for new technology evaluation also provides a helpful framework for reflecting upon the nature and significance of the way in which live FRT tests examined in this paper were designed, conducted, and evaluated. These trials took the form of controlled field tests intended to assess software performance (in the case of the Berlin and Nice trials) and their operational effectiveness (in the case of the Welsh and London trials) in real-world contexts. So understood, controlled field trials of software-driven technology have the potential to play an important role in addressing uncertainty about their 'in domain' performance and whether they can deliver tangible benefits to society in real-world use contexts, for which knowledge and evidence have been persistently lacking. To this end, we agree with the call by Stanford Human Computing Laboratory for meaningful 'in domain accuracy' performance metrics (which Phase III trials should generate), and the need to attend to specific domains of application to assess live FRT accuracy to enable genuine progress in addressing on-going controversy surrounding these technologies, particularly in public settings to support law enforcement authorities (Ho et al., 2020).

But, at least for highly rights-intrusive technologies, particularly biometric surveillance systems that can be deployed remotely, such as live FRT, we must insist upon evidence of real-world benefits such that their adverse impacts on fundamental rights and democratic freedom can be justified in accordance with legal tests of necessity and proportion. In other words, establishing the trustworthiness of live FRT requires more than evidence of the software's matching capabilities in the field: It also requires testing whether the software can be integrated successfully into a complex organisational, human, and socio-technical system to generate the expected benefits to the deploying organisation that can be plausibly regarded as socially beneficial, in a manner that is consistent with respect for fundamental rights (Castelluccia and Le Métayer, 2020, 19-20). As Pinch (1993) observes, because many technologies depend for their operation on the skilled, concerted actions of a user (the technology operator) for their successful operation, in these cases, technology tests are as much about testing the user as the machine. But, as we have seen, because this entails significant legal dangers (given the conditions under which police officers may lawfully stop and detain individuals), 'operational' testing of highly intrusive technologies should not occur unless and until a careful legal evaluation of those dangers has been conducted, and appropriate safeguards put in place to ensure that the conduct of the trial will not result in unlawful rights violations.

Our analysis of suggests that in-the-wild technology trials can, in principle, create opportunities to generate reliable evidence through which the capabilities of these technologies to serve and promote legitimate and lawful operational objectives can be investigated, thereby contributing to informed public deliberation about whether such technologies are likely to serve the public interest, including questions about their value for money. But whether they do in practice, particularly in the absence of clear guidance and governance frameworks to ensure that these trials in an epistemically, legally, and ethically responsible manner, we worry that they such tests will be little more than 'show trials'—public performances used to legitimate the use of powerful and invasive digital technologies in support of controversial political agendas for which public debate and deliberation is lacking, while deepening governmental reliance on commercially developed technologies which fall far short of the legal and constitutional standards which public authorities are required to uphold. Just as sociologists of testing have highlighted how tests are used as vehicles for justification, we should be particularly attentive to the way in which technology trials are undertaken to evaluate whether claims based on those test results can be defended and justified.

In this respect, the experience of live trials of FRT by London MPS is particularly troubling; not only is the data and logic underpinning the NPL's evaluation report proclaiming the 'success' of these trials woefully inadequate, even more alarmingly, the London MPS announced in January 2020 its intention to use FRT in London on a permanent basis, without making any reference to findings from its testing programme, including claims by the MPS that the NeoFace software was 'extremely accurate' (Robinson, 2020). This announcement suggests that the London MPS FRT deployments were described as 'trials' merely to assuage public fears that the MPS was planning to adopt the technology on a more permanent basis, while providing opportunities to 'normalise' these technologies in the public mind without any genuine interest in collecting reliable evidence to inform decisions about whether adoption was likely to serve lawful and legitimate operational objectives.

Seen in this light, various provisions in the EU AI Act concerning testing (including those concerned with testing within an AI regulatory sandbox: ss 57–59) are a welcome step in the right direction. For example, the Act refers to 'testing' of various kinds that must be conducted in relation to 'high risk' AI systems prior to being put into service or placed on the market, including:

- Art 10, which requires that testing data used during the process of training AI models for high-risk AI systems meet the 'quality criteria' set out in Art 10(2)–(5), including bias detection measures;
- Art 9(5), which requires that the most appropriate and targeted risk management measures are identified and tested to ensure that they perform consistently for their intended purpose, and to eliminate or reduce risks (including risks to fundamental rights) 'as far as technically feasible through adequate design and development'; and
- Art 13, which requires the provision of instructions for use to accompany the AI system, which include, among other things, information about the levels of accuracy (including metrics), robustness, and cybersecurity against which the system has been tested and validated, as required by Art 15.

Although examination of these provisions is beyond the scope of our inquiry, our analysis (particularly our comparative case study examination of FRT testing) is directly relevant to Art 60, which permits providers and potential providers of high-risk AI systems listed in Annex III to undertake real-world condition testing, provided that the conditions set out in Art 60(4) have been met. This includes, among other things, the submission and approval of a real-world testing plan by a market surveillance authority, which meets the 'detailed elements' that the Commission has specified pursuant to Art 60(1). Based on the preceding analysis, we suggest that these detailed elements (which the Commission is yet to specify) should collectively ensure that approved real-world tests are conducted in an epistemically, legally, and ethically responsible manner. When combined with the duty of high-risk AI providers to eliminate or reduce risks 'as far as technically feasible', real-world testing could serve as an important safeguard against the risks of interferences to fundamental rights that might arise from high-risk AI systems. In relation to the use of FRT in public settings, it offers a vehicle for preventing the kind of casual approach to 'watchlist' composition adopted in the UK FRT trials by LEAs which included images of thousands of 'wanted' individuals, based on broad and amorphous categories that would not, in our view, satisfy the tests of necessity and proportion needed to justify the privacy and data protection interferences thereby entailed. Yet it is conceivable that the capacity of FRT software to store facial images could be technologically configured to restrict the number of images it may hold at any one time, and that Article 9(5) could be interpreted to mandate 'privacy and data protection by design' safeguards of this kind to reduce the likelihood and severity of FRT 'drift'. If so, then these and other 'human rights by design' safeguards could be important in helping prevent the well-known dangers associated with the expansion of databases by stealth. Castelluccia and Le Métayer point out that the French DNA national database was originally created to centralise the fingerprints of persons convicted of extremely serious offences (i.e., the murder of a minor person preceded or accompanied by rape, torture, or barbaric acts) but by 2018 had been successively extended to include nearly 3 million DNA profiles. They show how similar expansion can be readily anticipated in relation to FRT systems (2020, 18):

> Some opponents even argue that …a deliberate strategy of facial recognition promoters…consist [s] in using it first in contexts where the purpose seems legitimate and then gradually to extend their use. The 'slippery slope' argument must therefore be considered seriously, but it must be analysed precisely and in concrete terms to avoid sophisms. With regard to the example of access control to station platforms, it could be argued that there is a significant risk of generalisation to all modes of transport (metro, tram, bus etc) that would lead to a total loss of freedom to move anonymously. Beyond transport, we could also imagine a generalisation to all closed places where people gather (cinemas, theatres, shopping centres etc.): if these systems are considered effective, why would the

https://doi.org/10.1017/dap.2025.10019 Published online by Cambridge University Press

> protection that they offer be limited to public transport? If this generalisation is not acceptable, where should the red line be set and what safeguards should be provided to ensure that this limit will be enforced? In other cases, it is the databases themselves that could later be extended or cross-referenced with other databases.

The dangers of 'drift and shift' in the deployment of FRT highlight both the critical need for clear, bright-line, legal safeguards in the face of ever-expanding mass yet fine-grained surveillance capabilities of AI-enabled systems, and the role of real-world testing to produce demonstrable evidence that they can effectively deliver their promised benefits without generating unacceptable risks.

5. Conclusion

We have highlighted the urgent need for regulatory norms and institutional frameworks to govern the design, conduct, and evaluation of in-the-wild technology testing, particularly for AI systems that may threaten the rights and interests of others. Technology tests are currently a largely ungoverned 'wild west' without legal guarantees to ensure that they are being conducted in an epistemically, ethically, and legally responsible manner, as our analysis of live FRT trials by LEAs across four European sites between 2016 and 2020 demonstrates. Under the EU AI Act, this is set to change, albeit only in relation to high-risk systems that fall within its purview. Although 'in-the-wild' testing has the potential to act as an important opportunity to collect evidence about how new technologies perform and operate in real-world deployment environments, such tests may themselves cause harm and wrongfully interfere with the rights of others. Our case studies also highlight an important difference between a technical system's 'functional performance' and its 'operational effectiveness': The mere ability of a technology to perform specific functions (such as automated image matching) does not necessarily translate into concrete benefits for the organisation deploying it. As the UK trials demonstrate, the task of translating an automated FRT match alert into a lawful apprehension of a wanted individual by police is far from simple for it requires the successful integration of multiple technical, organisational, environmental, and human components, while generating a host of significant yet thorny legal and ethical dangers, which have not yet been adequately or properly grasped by LEAs wishing to embrace them. Furthermore, none of these trials generated clear evidence that the technology actually delivered real-world benefits in the form of improved *operational efficiency* in locating, identifying, and apprehending criminal suspects by law enforcement authorities, nor by how much.

Although much of the policy and academic debate has hitherto focused on the social and ethical acceptability of AI-enabled technologies, we have highlighted the need to ensure compliance with legal rights and duties as a prerequisite for responsible in-the-wild technology testing. Leaving aside the EU's AI Act, a considerable body of existing law already applies to the conduct of live FRT trials by LEAs, including data protection legislation, human rights legislation, equality laws, and statutory and common laws that condition and constrain the activities and decisions of law enforcement authorities. Yet awareness and understanding about how existing laws should guide and constrain these trials appear to be seriously lacking. As the experiences on live FRT trials in the United Kingdom demonstrate, even assuming well-meaning intentions on the part of the LEAs seeking to trial these technologies, there are multiple difficult normative judgements and trade-offs that must be made when designing and conducting FRT trials, to ensure that legal rights are appropriately respected. Accordingly, several oversight bodies in the United Kingdom that have some connection with biometric technologies utilising machine vision techniques have repeatedly called upon the British government to enact new laws to clarify and restrain the indiscriminate use of these technologies but continue to fall on deaf ears.

Digital technologies, including AI, are ultimately tools in the hands of those who deploy them. While data-enabled technologies can be harnessed in ways that might provide valuable new insight, digital data have unique properties that enable them to be instantly copied at virtually no cost and readily transmitted and circulated across the world at scale, yet in ways that leave no technological trace of how it has been handled. Combined with the opacity of databases, digital data, particularly facial data, has extraordinary

https://doi.org/10.1017/dap.2025.10019 Published online by Cambridge University Press

potential to be misused and even weaponised in ways that may readily escape our awareness, let alone systematic, independent, and meaningful scrutiny. To this end, it is vital that due account is taken of the highly powerful, intrusive, and scalable properties of live FRT that explain the controversy surrounding it, particularly given its capacity for misuse and over-reach in ways that contravene basic principles of liberal democratic societies while interfering with rights to privacy, to freedom of expression and freedom of assembly, and the basic rights to go about one's lawful activity in public without unjustified interference by the state. Accordingly, in our view, evidence of their effectiveness in producing the desired benefit must pass an exceptionally high threshold if these tests are to demonstrate that the contribution of these technologies to legitimate policing purposes is of such great value that they might justify such serious rights interferences at scale. The capacity for the state to instantly and continuously identify and track (at a distance) every individual is the stuff of dystopian science fiction (American Civil Liberties Union, 2021). It is this feature alone that magnifies and accentuates the need for great care in the design and conduct of these trials, let alone their permanent deployment. Although we welcome provisions of the EU's AI Act intended to safeguard against fundamental rights interferences, including the conduct of real-world testing for 'high-risk' AI systems, they do not appear to require testing to establish the proof of 'real-world efficacy' which is legally mandated before new drugs can be released, and which, in our view, is a significant omission.

In arguing for the importance of 'responsible' AI testing, and in focusing our case study examination on the testing of live FRT by LEAs, we are not signalling our support for the use of these technologies. On the contrary, our argument is founded on the need for reliable evidence to verify unproven claims by LEAs that FRT will enable them to 'find terrorists' and 'locate missing children'. For some, the dangers of live FRT are so great that the technologies should be treated as analogous to nuclear weapons and therefore banned outright (Restrepo, 2023). We have considerable sympathy with this view, given that these technologies are novel, highly intrusive, prone to misuse, and difficult to contain, and hence many civil society organisations have criticised the AI Act's restrictions on the use of live FRT and other biometric technologies as too weak. In the United Kingdom, the political climate appears tipped firmly in favour of boosting rather than imposing restrictions on the use of AI-enabled surveillance technologies, with little political appetite to significantly restrict the use of FRT by law enforcement authorities, let alone an outright legal ban. The best we can hope for in the United Kingdom is a regulatory framework to govern in-the-wild trials of rights-intrusive technologies to ensure that they are responsibly undertaken in an epistemically, legally, and ethically responsible manner. Our arguments are intended to apply to all biometric monitoring technologies, and are not confined to live FRT systems and which public authorities may be keen to adopt and subject to in-the-field testing.

The long history of failure to attend to the unintended effects of new pharmaceuticals and the need for rigorous, ethically conducted, legally mandated, and independently evaluated clinical trials should be a sobering wake-up call for those otherwise in thrall to the spell of the 'Digital Enchantment' (Yeung, 2023; Junod, 2008). If we wish to nurture and sustain democracy and individual freedom in a digital age, we must insist upon independent scrutiny and vigilance at the 'evaluation' stage prior to the public release of highly intrusive digital technologies, including live FRT. Although Luke Stark has referred to live FRT as the 'plutonium of AI', perhaps it is better understood as the Thalidomide of AI: a new technology enthusiastically embraced across the globe based on untested promises about its benefits without rigorous trials to generate evidence establishing its capacity to deliver on its promises and without sufficient attention to readily foreseeable collateral damage. Yet, unlike Thalidomide, the damage wreaked by the widespread use of live FRT may be much harder to see in the absence of thousands of dead or deformed babies, instead taking the form of the incremental and insidious removal of the conditions that make individual freedom and authentic self-creation in public spaces possible. If we are to take seriously our basic freedom to go about our lawful business in public spaces without state interference, and the opportunity for self-creation and development which this freedom affords, then we must be vigilant. This includes a need to test these technologies in a responsible manner to demonstrate that they do in fact generate valuable social benefits, given their economic and other costs, and to undertake such tests responsibly.

https://doi.org/10.1017/dap.2025.10019 Published online by Cambridge University Press

Data availability statement. Not applicable. This is a conceptual paper based on publicly available information and did not create any.

Author contribution. The authors jointly conceived this project. Wenlong Li undertook the detailed case study investigations. Karen Yeung devised the analytical framework, developed the arguments based on the case studies, and produced the original draft text which both authors iteratively revised and refined to produce a final draft.

Funding statement. The authors did not receive any external funding to support this research

Competing interests. Karen Yeung is currently the recipient of an award from the Engineering and Physical Sciences Research Council Grant number EP/Y009800/1, through funding from Responsible Ai UK (Project title: AI Equality by Design, Deliberation and Oversight), which provides funding to enable her to represent and support Equinet (the European Network of Equality Bodies) in the on-going JTC21 work program to draft technical standards for high-risk AI systems pursuant to Chapter 2 of the EU AI Act.

References

Ada Lovelace Institute (2019) Beyond face value: Public attitudes to facial recognition technology. *Ada Lovelace Institute*. https://www.adalovelaceinstitute.org/case-study/beyond-face-value/#:~:text=In%20September%202019%2C%20the%20Ada,and%20who%20they%20trusted%20to (accessed 22 February 2024).

Ada Lovelace Institute (2022) Countermeasures: The need for new legislation to govern biometric technologies in the UK, London. *Ada Lovelace Institute*. https://www.adalovelaceinstitute.org/report/countermeasures-biometric-technologies/ (accessed 31 October 2022).

American Civil Liberties Union (2021) Coalition letter requesting federal moratorium on facial recognition. *ACLU*. https://www.aclu.org/letter/coalition-letter-president-biden (accessed 29 April 2021).

Benbunan-Fich R (2017) The ethics of online research with unsuspecting users: From a/B testing to C/D experimentation. *Research Ethics 13* (3–4), 200–218. https://doi.org/10.1177/1747016116680664.

Bird S, Barocas S, Crawford K, Diaz F and Wallach H (2016) *Exploring or exploiting? Social and ethical implications of autonomous experimentation in AI*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2846909 (accessed 19 February 2024)

Boyd D (2016) Untangling research and practice: What Facebook's 'emotional contagion' study teaches us. *Research Ethics 12*(1), 4–13. https://doi.org/10.1177/1747016115583379.

Bradford B, Yesberg J, Jackson J and Dawson P (2020) Live facial recognition: Trust and legitimacy as predictors of public support for police use of new technology. *The British Journal of Criminology 60*(6), 1502–1522. https://doi.org/10.1093/bjc/azaa032

Buchanan EA and Ess CM (2009) Internet research ethics and the institutional review board. *ACM SIGCAS Computers and Society 39*(3), 43–49. https://doi.org/10.1145/1713066.1713069.

Buolamwini J and Gebru T (2018) Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research 81*, 1–15.

Buolamwini J, Ordóñez V, Morgenstern J and Learned-Miller E (2020) Facial recognition technologies: A primer. *Algorithmic Justice League*. https://circls.org/primers/facial-recognition-technologies-a-primer (accessed 19 February 2024)

Callon M (1986) The sociology of an actor-network: The case of the electric vehicle. In Callon M, Law J and Rip A (eds), *Mapping the Dynamics of Science and Technology*. Hampshire: Palgrave Macmillan UK.

Cano R (2024) One crash set off a new era for self-driving cars in S.F. Here's a complete look at what happened. *SF Chronicle*. https://www.sfchronicle.com/projects/2024/cruise-sf-collision-timeline/ (Accessed 17 April 2024).

Castelluccia C and Le Métayer D (2020) Impact analysis of facial recognition towards a rigorous methodology. HAL Id: hal-02480647. https://inria.hal.science/hal-02480647/document (accessed 20 February 2024)

Castelvecchi D (2020) Beating biometric bias. *Nature 587*(347–9), 80–90. https://doi.org/10.1038/d41586-020-03186-4.

Chase J (2017) German police seek volunteers for facial recognition surveillance. *DW*. https://www.dw.com/en/german-police-seek-volunteers-for-facial-recognition-surveillance/a-39314012 (accessed 29 April 2021).

CNIL (2019) Facial recognition: For a debate living up to the challenges. https://www.cnil.fr/en/facial-recognition-debate-living-challenges (accessed 20 February 2024)

Connexion Journalists (2019a) French debate on facial recognition. *Connexion*. https://www.connexionfrance.com/French-news/French-debate-on-facial-recognition (accessed 7 April 2021)

Connexion Journalists (2019b) France set to test facial recognition security cameras. *Connexion*. https://www.connexionfrance.com/French-news/France-set-to-test-facial-recognition-security-cameras-says-junior-minister-after-successful-trials-in-Nice (accessed 7 April 2021)

Criado-Perez C (2020) *Invisible Women*. *Vintage*.

Dastin J (2022) Amazon scraps secret AI recruiting tool that showed bias against women. In Martin K (ed), *Ethics of Data and Analytics*, 296–299. Auerbach Publications.

Davies B, Innes M and Dawson A (2018) An evaluation of South Wales police's use of automated facial recognition. https://www.cardiff.ac.uk/__data/assets/pdf_file/0006/2426604/AFRReportDigital.pdf (accessed 7 April 2021)

https://doi.org/10.1017/dap.2025.10019 Published online by Cambridge University Press

Dixon-Woods M and Ashcroft RE (2008) Regulation and the social licence for medical research. *Med Health Care and Philos 11*, 381–391. https://doi.org/10.1007/s11019-008-9152-0.

EDPB (2022) Guidelines 05/2022 on the use of facial recognition technology in the area of law enforcement. *EDPB*. https://www.edpb.europa.eu/our-work-tools/documents/public-consultations/2022/guidelines-052022-use-facial-recognition_en (accessed 19 August 2024).

Eireiner AV (2020) Imminent dystopia? Media coverage of algorithmic surveillance at Berlin-Südkreuz. *Internet Policy Review 9*(1), 1–19. https://doi.org/10.14763/2020.1.1459.

Estrosi C (2019) *Experimentation Reconnaissance Faciale*. Ville de Nice.

European Union Agency for Fundamental Rights (2019) Facial recognition technology: Fundamental rights considerations in the context of law enforcement. https://fra.europa.eu/en/publication/2019/facial-recognition-technology-fundamental-rights-considerations-context-law (accessed 20 February 2024)

Flick C (2016) Informed consent and the Facebook emotional manipulation study. *Research Ethics 12*(1), 14–28. https://doi.org/10.1177/1747016115599568.

Fussey P (2012) Eastern promise? East London transformations and the state of surveillance. *Information Polity 17*(1), 21–34. https://doi.org/10.3233/IP-2011-0253.

Fussey P, **Davies B and Innes M** (2021) 'Assisted' facial recognition and the reinvention of suspicion and discretion in digital policing. *The British Journal of Criminology 61*(2), 325–344. https://doi.org/10.1093/bjc/azaa068.

Fussey P and Murray D (2019) Independent report on the London metropolitan police service's trial of live facial recognition technology. *The Human Rights, Big Data and Technology Project*. https://repository.essex.ac.uk/24946/1/London-Met-Police-Trial-of-Facial-Recognition-Tech-Report-2.pdf (accessed 20 February 2024)

Fussey P and Murray D (2020) Policing uses of live facial recognition in the United Kingdom. In Kak A (ed), *Regulating Biometrics: Global Approaches and Urgent Questions.* AI Now, 78–85. https://ainowinstitute.org/publication/regulating-biometrics-global-approaches-and-open-questions (accessed 20 February 2024)

Gates KA (2004) *The Past Perfect Promise of Facial Recognition Technology.* ACDIS Occasional Paper.

Gates KA (2011) *Our Biometric Future: Facial Recognition Technology and the Culture of Surveillance*. NYU Press.

German Federal Police (2018) *Biometrische Gesichtserkennung: Des Bundespolizeipräsidiums Im Rahmen der Erprobung von Systemen Zur Intelligenten vi- Deoanalyse Durch das Bundesministerium Des Innern, für Bau Und Heimat, das Bundespolizeipräsidium, das Bundeskriminal- Amt Und Die Deutsche Bahn AG Am Bahnhof Berlin Südkreuz.*

Grimmelmann J (2015) The law and ethics of experiments on social media users. *Colorado Technology Law Journal 13*, 219–272.

Hammersley M (2020) On epistemic integrity in social research. In Iphofen R (ed), *Handbook of Research Ethics and Scientific Integrity.* Switzerland: Springer Nature Switzerland. https://doi.org/10.1007/978-3-030-16759-2_16.

Ho DE, **Black E**, **Agrawala M and Li F** (2020) Evaluating facial recognition technology: A protocol for performance assessment in new domains. *HAI.* https://hai.stanford.edu/sites/default/files/2020-11/HAIFacialRecognitionWhitePaper.pdf (accessed 20 February 2024)

Huang GB, **Ramesh M**, **Berg T and Learned-Miller E** (2008) Labeled faces in the wild: A database for studying face recognition in unconstrained environments. https://vis-www.cs.umass.edu/lfw/s

Humphreys M (2015) Reflections on the ethics of social experimentation. *Journal of Globalization and Development 6*(1), 87–112. https://doi.org/10.35188/UNU-WIDER/2015/903-9.

Introna L and Nissenbaum H (2010) Facial recognition technology: A survey of policy and implementation issues. https://www.research.lancs.ac.uk/portal/en/publications/facial-recognition-technology-a-survey-of-policy-and-implementation-issues(43367675-c8b9-4644-90f2-86815cc8ea15)/export.html (accessed 20 February 2024)

Jiang S, **Martin J and Wilson C** (2019) Who's the Guinea pig? Investigating online a/B/N tests in-the-wild. in *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, pp. 201–210. https://doi.org/10.1145/3287560.3287565

Jouhki J, **Lauk E**, **Penttinen M**, **Sormanen N and Uskali T** (2016) Facebook's emotional contagion experiment as a challenge to research ethics. *Media and Communication 4*(4A), 75–85. https://doi.org/10.17645/mac.v4i4.579.

Junod SW (2008) FDA and clinical drug trials: A short history. In Davies M and Kerimani F (eds), *A Quick Guide to Clinical Trials*, Bioplan, Inc., pp. 25–55.

Justice Sub-Committee on Policing (2020) Facial recognition: How policing in Scotland makes use of this technology. https://archive2021.parliament.scot/parliamentarybusiness/currentcommittees/113103.aspx (accessed 11 March 2021).

Kayali L (2019) How facial recognition is taking over a french city. *POLITICO*. https://www.politico.eu/article/how-facial-recognition-is-taking-over-a-french-riviera-city/ (accessed 11 March 2021).

Kirchhoffer DG and Richards BJ (2019) *Beyond Autonomy: Limits and Alternatives to Informed Consent in Research Ethics and Law.* Cambridge University Press.

Kotsoglou KN and Oswald M (2020) The long arm of the algorithm? Automated facial recognition as evidence and trigger for police intervention. *Forensic Science International: Synergy 2*, 86–89. https://doi.org/10.1016/j.fsisyn.2020.01.002.

Laanstra-Corn A & Sewell T (2024) Algorithmic Surveillance Takes the Stage at the Paris Olympics. *Lawfare*. https://www.lawfaremedia.org/article/algorithmic-surveillance-takes-the-stage-at-the-paris-olympics

Latour B (1993) *The Pasteurization of France*. Cambridge: Harvard University Press.

Lederer S (2007) Research without Borders: The origin of the declaration of Helsinki. In Schmidt U and Frewer A (eds), *A History and Theory of Human Experimentation*. Stuttgart: Franz Steiner Verlag.

https://doi.org/10.1017/dap.2025.10019 Published online by Cambridge University Press

Lemmens T (1999) In the name of national security: Lessons from the final report on the human radiation experiments. *European Journal of Health Law 6*(1), 7–23. https://doi.org/10.1163/15718099920522659.

Leslie D (2020) Understanding bias in facial recognition technologies: An explainer. *The Alan Turing Institute*. https://doi.org/10.5281/zenodo.4050457 (accessed 22 February 2024)

Liberty (2019) Liberty's briefing on police use of live facial recognition technology. *Liberty*. https://www.libertyhumanrights.org.uk/wp-content/uploads/2020/02/LIBERTYS-BRIEFING-ON-FACIAL-RECOGNITION-November-2019-CURRENT.pdf

Lochner SA (2013) Saving face: Regulating law enforcement's use of mobile facial recognition technology and iris scans. *Arizona Law Review 55*, 201.

London Policing Ethics Panel (2018) Interim Report on Live Facial Recognition. London. http://www.policingethicspanel.london/uploads/4/4/0/7/44076193/lpep_report_-_live_facial_recognition.pdf (accessed 22 February 2024)

London Policing Ethics Panel (2019) Final Report on Live Facial Recognition. London. http://www.policingethicspanel.london/uploads/4/4/0/7/44076193/live_facial_recognition_final_report_may_2019.pdf (accessed 22 February 2024)

MacKenzie D (1989) From Kwajalein to Armageddon? Testing and the social construction of missile accuracy. In Gooding D, Schaffer S and Worrall J (eds), *The Use of Experiment: Studies in the Natural Sciences*. Cambridge: Cambridge University Press, pp. 409–436.

MacKenzie D (1990) *Inventing Accuracy: A Historical Sociology of Nuclear Missile Guidance*. The MIT Press.

MacKenzie D (1999) Nuclear missle testing and the social construction of accuracy. *The Sciences Studies Reader*, 343–357.

Marres N and Stark D (2020) Put to the test: For a new sociology of testing. *British Journal of Sociology 71*(3), 423–443. https://doi.org/10.1111/1468-4446.12746.

Merton R (1973) *The Sociology of Science: Theoretical and Empirical Investigations*. University of Chicago Press.

Metropolitan Police Services (2021) Live facial recognition: Legal mandate. https://www.met.police.uk/SysSiteAssets/media/downloads/force-content/met/advice/lfr/policy-documents/lfr-legal-mandate.pdf (accessed 22 February 2024)

Miller F and Wertheimer A (2010) *The Ethics of Consent: Theory and Practice*. Oxford University Press.

Misuraca G and van Noordt C (2020) Artificial intelligence in public services. *JRC*. https://publications.jrc.ec.europa.eu/repository/handle/JRC120399 (accessed 22 February 2024)

Mody CCM and Lynch M (2010) Test objects and other epistemic things: A history of a nanoscale object. *British Journal for the History of Science 43*(3), 423–458. https://doi.org/10.1017/S0007087409990689.

Mordini E and Tzovaras D (2012) *Second Generation Biometrics: The Ethical, Legal and Social Context*. Springer Science & Business Media.

National Physical Laboratory (2020) Metropolitan police service live facial recognition trials. *National Physical Laboratory* https://www.met.police.uk/SysSiteAssets/media/downloads/central/services/accessing-information/facial-recognition/met-evaluation-report.pdf (accessed 22 February 2024).

Parkhi OM, Vedaldi A and Zisserman A (2015) Deep face recognition. https://www.robots.ox.ac.uk/~vgg/publications/2015/Parkhi15/parkhi15.pdf (accessed 22 February 2024)

Pascu L (2020) France looks to establish legal framework to deploy biometric video surveillance. *Biometric Update*. https://www.biometricupdate.com/202001/france-looks-to-establish-legal-framework-to-deploy-biometric-video-surveillance (accessed 29 April 2021)

Persaud N (2010) Protocol. In Salkind NJ (ed). *Encyclopedia of Research Design*. Vol. *2*. SAGE Publications, 1132–1135.

Pinch T (1993) Testing - one, two, three… Testing!: Toward a sociology of testing. *Science, Technology and Human Values 18*(1), 25–41. https://doi.org/10.1177/016224399301800103.

Polonetsky J, Tene O and Jerome J (2015) Beyond the common rule: Ethical structures for data research in non-academic settings. *Colorado Technology Law Journal 13*, 333–368.

Power M (1997) *The Audit Society: Rituals of Verification*. Oxford: OUP Oxford.

Purshouse J and Campbell L (2019) Privacy, crime control and police use of automated facial recognition technology. *Criminal Law Review 3*, 188–204.

Purshouse J and Campbell L (2022) Automated facial recognition and policing: A bridge too far? *Legal Studies 42*(2), 209–227.

R (Bridges) v Chief Constable of the South Wales Police [2020] EWCA Civ 1058

Resnik DB and Hofweber F (2023) Research ethics timeline. *National Institute of Environmental Health Sciences* https://www.niehs.nih.gov/research/resources/bioethics/timeline (accessed19 August 2024)

Restrepo ML (2023) She was denied entry to a Rockettes show — Then the facial recognition debate ignited. *NPR*. https://www.npr.org/2023/01/21/1150289272/facial-recognition-technology-madison-square-garden-law-new-york (accessed 19 August 2024)

Reuter M (2017) Bundesregierung: Test am südkreuz wird auf jeden fall ein erfolg. *Netzpolitik.Org*. https://netzpolitik.org/2017/bundesregierung-test-am-suedkreuz-wird-auf-jeden-fall-ein-erfolg/ (accessed 9 May 2021).

Reverby SM (2013) *Examining Tuskegee: The Infamous Syphilis Study and its Legacy*. Chapel Hill: The University of North Carolina Press.

Robinson M (2020) Hundreds of innocents face being grabbed in the street by police as Scotland yard admits new facial recognition system gives false alerts one in every thousand faces - as it introduces cameras to spot criminals in London's busiest public places. *Daily Mail*. https://www.dailymail.co.uk/news/article-7924733/Scotland-Yard-introduces-facial-recognition-cameras-hunt-watchlist-2-500-suspects.html (accessed 18 August 2024)

Rice v Connolly [1966] 2 QB 414

https://doi.org/10.1017/dap.2025.10019 Published online by Cambridge University Press

Rooksby J, **Rouncefield M and Sommerville I** (2009) Testing in the wild: The social and organisational dimensions of real-world practice. *Computer Supported Cooperative Work (CSCW) 18*, 559–580.

Roth CL (2021) New surveillance Technologies in Public Spaces: Challenges and perspectives for European Law at the example of facial recognition. *Urban Agenda for the EU*. https://futurium.ec.europa.eu/sites/default/files/2021-06/Action%203%20-%20Final%20Report%20v0.2.pdf (accessed 27 March 2024).

Roy A (2024) Driverless cars covered 5x more test miles in California in 2023. *Reuters*. https://www.reuters.com/business/autos-transportation/driverless-cars-covered-5x-more-test-miles-california-2023-2024-02-02/ (accessed 17 April 2024).

Ryder M (2022) The Ryder Review: Independent legal review of the governance of biometric data in England and Wales. *Ada Lovelace Institute*. https://www.adalovelaceinstitute.org/report/ryder-review-biometrics/

Sheriff L (2023) Endless fallout: The Pacific idyll still facing nuclear blight 77 years on. *The Guardian*. https://www.theguardian.com/environment/2023/aug/25/endless-fallout-marshall-islands-pacific-idyll-still-facing-nuclear-blight-77-years-on (accessed 29 May 24).

Spiegelhalter D (2020) Should we trust algorithms? *Harvard Data Science Review 2*(1). https://doi.org/10.1162/99608f92.cb91a35a.

Stilgoe J (2018) Machine learning, social learning and the governance of self-driving cars. *Social Studies of Science 48*(1), 25–56.

Surveillance Camera Commissioner (2019) The police use of automated facial recognition technology with surveillance camera systems. https://assets.publishing.service.gov.uk/media/5fc75c5d8fa8f5474a9d3149/6.7024_SCC_Facial_recognition_report_v3_WEB.pdf (accessed 22 February 2024)

Surveillance Camera Commissioner (2020) Facing the camera: Good practice and guidance for the police use of overt surveillance camera systems incorporating facial recognition technology to locate persons on a watchlist, in public places in England & Wales. https://assets.publishing.service.gov.uk/media/5fc75c5d8fa8f5474a9d3149/6.7024_SCC_Facial_recognition_report_v3_WEB.pdf (accessed 20 February 2024)

Tene O and Polonetsky J (2016) Beyond IRBs: Ethical guidelines for data research. *Washington and Lee Law Review Online 72*(3), 458.

The World Medical Association (2018) WMA declaration of Helsinki– Ethical principles for medical research involving human subjects. *The World Medical Association*. https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects (accessed 21 May 2021).

Theodore C and Trotry A (2020) French white paper on internal security makes several proposals for the use of facial recognition in France. *AI-Regulation.Com*. https://ai-regulation.com/french-white-paper-on-internal-security-makes-several-proposals-for-the-use-of-facial-recognition-in-france/ (accessed 7 April 2021)

Trigg A and Milner L (2023) Three Arrests after Force Trials Facial Recognition. *BBC*. https://www.bbc.com/news/articles/c51jzy42p19o (accessed 19 August 2024)

Trigueros DS, **Meng L and Hartnett M** (2018) *Face Recognition: From Traditional to Deep Learning Methods*. arXiv: 1811.00116 [cs.CV]. https://arxiv.org/abs/1811.00116 (accessed 22 February 2024)

Untersinger M (2019) Reconnaissance faciale: La CNIL tique Sur Le bilan de l'expérience niçoise. *Le Monde*. https://www.lemonde.fr/pixels/article/2019/08/28/reconnaissance-faciale-la-cnil-tique-sur-le-bilan-de-l-experience-nicoise_5503769_4408996.html (accessed 29 April 2021).

Van de Poel I (2016) An ethical framework for evaluating experimental technology. *Science and Engineering Ethics 22*(3), 667–786. https://doi.org/10.1007/s11948-015-9724-3.

Van de Poel I (2017) Moral experimentation with new technology. In Asveld L and Mehos DC (eds), *New Perspectives on Technology in Society: Experimentation beyond the Laboratory*. Routledge

Van de Poel I, **Lotte A and Mehos DC** (2017) *New Perspectives on Technology in Society: Experimentation beyond the Laboratory*. Routledge.

Vertesi J (2015) *Seeing like a Rover: How Robots, Teams, and Images Craft Knowledge of Mars*. Chicago: University of Chicago Press.

Vollmann J and Winau R (1996) Informed consent in human experimentation before the Nuremberg code. *BMJ 313* (7070), 1445–1447. https://doi.org/10.1136/bmj.313.7070.1445.

Yeung K (2019) Why worry about decision-making by machine?'. In Yeung K and Lodge M (eds), *Algorithmic Regulation*. Oxford: Oxford University Press.

Yeung K (2023) Dispelling the digital enchantment: How can we move beyond its destructive influence and reclaim our right to an open future? *Prometheus 39*(1), 8–27. https://doi.org/10.13169/prometheus.39.1.0008.

Yeung K and Bygrave LA (2022) Demystifying the modernized European data protection regime: Cross-disciplinary insights from legal and regulatory governance scholarship. *Regulation & Governance 16*(1), 137–155.

Zafeiriou S, **Zhang C and Zhang Z** (2015) A survey on face detection in the wild: Past, present and future. *Computer Vision and Image Understanding 138*, 1–24. https://doi.org/10.1016/j.cviu.2015.03.015.

Cite this article: Yeung K and Li W (2025). From 'wild west' to 'responsible' AI testing 'in-the-wild': lessons from live facial recognition testing by law enforcement authorities in Europe. *Data & Policy*, 7: e59. doi:10.1017/dap.2025.10019