LSE Media and Communications

Media@LSE MSc Dissertation Series

Editors: Bart Cammaerts and Nick Anstead

POLICE FACIAL RECOGNITION IN PROGRESS

The Construction of the Notion of Accuracy in the Live Facial Recognition Technology Used by the MET Police in London.

Romina Colman

'POLICE FACIAL RECOGNITION IN PROGRESS': THE CONSTRUCTION OF THE NOTION OF ACCURACY IN THE LIVE FACIAL RECOGNITION TECHNOLOGY USED BY THE MET POLICE IN LONDON

ROMINA COLMAN[1]

[1] rominacolman@gmail.com

Published by Media@LSE, London School of Economics and Political Science ("LSE"), Houghton Street, London WC2A 2AE. The LSE is a School of the University of London. It is a Charity and is incorporated in England as a company limited by guarantee under the Companies Act (Reg number 70527).

Abstract

This research examines the construction of the notion of accuracy in the Live Facial Recognition (LFR) technology deployed by the MET Police in London. Using Fairclough's Critical discourse analysis (CDA), the work examined 5 evidence-based documents created by the MET Police to inform the implementation of the tool. Amongst other results, this study found that accuracy in LFR is constructed as a supervised-human metric that can be managed by "unbiased" Police officers during deployments and through technical interventions. Within this conception, problems in accuracy are understood from a transactional perspective: as "differences in performance" and not as signs of human biases. As a result, algorithms and data are portrayed as objective elements that can be contextually adjusted. Moreover, the notion of accuracy is strictly controlled in the MET Police discourse to construct LFR as a valid technology to fight against crime. To conclude, this work makes it evident why the study of computational concepts in relation to discourse matters. In this sense, a deeper commitment of media researchers is needed to explore facial recognition systems.

INTRODUCTION

Facial Recognition Technologies (FRT) are expanding fast, and so is their use for law enforcement purposes. According to a report of Algorithm Watch (Kayser-Bril, 2019), in 2019, at least 11 out of 25 member states of Europe had Police forces using these biometric identification tools, while 8 had confirmed their plan for incorporation. The radical expansion of these systems has raised concerns amongst experts and civil society groups who argue that governments are increasing surveillance practices on citizens, and targeting people wrongfully (Big Brother Watch, 2018.; EFF, 2007).

In January 2020, the London Metropolitan Police (henceforth MET Police) implemented a Live Facial Recognition technology (LFRT) 'to prevent and detect crime by helping officers find wanted criminals' (MET Police, n.d.-a, para. 2). While traditional FRT capture specific points of the face of individuals in images or videos and compare them against a list of images produced by the Government, here, the procedure is the same but comparisons are made in real-time. The MET Police deploy vans with cameras for special events or in specific parts of the city. The face of those who pass through the cameras is analysed against a particular watchlist prepared by the force. If the system finds a match, an alert is sent to police officers on the field through an app on their phones. After that, staff look for the person and verify its identity. Both traditional, and LFRT are automated systems: they produce decisions based on the operation of algorithms on datasets.

The deployment of LFRT in London created a heated debate and many NGOs condemned its use. The most relevant claims concentrated on problems in accuracy measures (Dodd, 2019; Big Brother Watch, 2018). An independent board of experts from the University of Essex (Fussey and Murray, 2019) that had access to evidence-based data produced by the MET Police concluded that the tool 'was verifiable accurate in just 19% of the cases' (Gayle, 2020, para. 12). Nevertheless, the Police emphasised to the media that false positives were '..one in a thousand' (Gayle, 2020, para. 13) and that they were using the most advanced algorithm of the industry known as NEC-3 (MET Police, n.d.-b).

Accuracy is a key concept in FRT. From a computer science (CS) perspective, low levels of accuracy are usually an indication of discriminatory effects against a certain group of individuals (Boulamwini and Gebru, 2018). In other words, they are signs of a biased system. The centrality of this notion is observed in an annual evaluation of the National Institute of Standards and Technology (NIST) in the United States. Every year, this government agency tests the performance of the most sophisticated facial recognition algorithms around the world. Statistical formulas are applied to measure and evaluate their performance. Although the final document has 3 parts, the accuracy assessment is the one that captures the attention of people inside and outside the CS community. Last year, the report concluded that the NEC-3 algorithm, the one used by the MET Police in their LFR technology, led the ranking as 'the most accurate' algorithm (Grother, Ngan and Hanaoka, 2019a).

As opposed to one-axis analyses from CS, this research adopts a socio-technical approach and brings into the discussion the study of the discursive dimension to understand how the notion of accuracy is constructed in the evidence-based reports of LFR produced by the MET Police. In doing so, this work aims to connect the technical and statistical discourses of accuracy with wider social relations to understand how this notion is constructed.

The originality of this research relies on two main characteristics. Firstly, until now, accuracy has been addressed only in computational terms and although authors such as Selbst et. al. (2019) have called for the adoption of a socio-technical approach, the research done for this project did not find studies on accuracy in FRT, and specifically in LFR, which had adopted this perspective. Secondly, although communication studies have addressed surveillance tools in relation to media publications (Eireiner, 2020; Barnard-Wills, 2011), the exploration of evidence-based reports remains limited.

This dissertation is divided into 5 sections. The first chapter explores CS and facial recognition research to understand how accuracy has been constructed until now. Furthermore, this section also examines the criticisms of the notion. While the conceptual framework explains how key concepts are addressed within the scope of this research, the methodology chapter provides a justification for the use of CDA for this work. The results of the analysis of the MET Police evidence-based reports are presented in the fourth section. This part also addresses some relevant recommendations for future research. Finally, the conclusion offers a summary of the overall findings of this work.

LITERATURE REVIEW

The section starts with an overview of the main characteristics of algorithms and algorithmic bias, two crucial elements to understand accuracy tests. Then, the accuracy definition and its criticisms are examined in depth. Since literature shows that the concept in FRT is inexorably linked to the CS field, the revision presented here is the result of an iterative process between CS and facial recognition research to explain how accuracy has been addressed until now. After that, this chapter explores the general call within the CS field to adopt a socio-technical perspective regarding the study of accuracy. The section ends with a brief conceptualisation of 'the technical' and 'the statistical' as discourses.

Algorithms: main features and criticisms

Algorithms are central to the operation of automated decision-making systems, and therefore, FRT. Within the CS field, they are commonly understood as a set of procedures that are applied 'for transforming input data into a desired output' (Gillespie, 2014: 167; Goffey, 2008). In other words, algorithms perform instructions to address a software problem. Two key dimensions are usually highlighted to describe how they work: a logical part that aims at defining the problem and a control component that executes a specific procedure to manage it (Kowalski, 1979).

Within the CS field, there is a strong tendency to describe algorithms as purely technical and objective (Kowalski, 1979; Sedgewick and Fajolet, 2013; Cormen *et. al*., 2009). This perspective is highly criticised by social science researchers and some computer scientists. Critics argue that it conceives them as independent from the wider societal relations in which they are embedded in (Kitchin, 2017; Seaver, 2018; Gillespie, 2014, Crawford, 2016; Cheney-Lippold, 2011; Barocas and Selbst, 2016).

Critical scholars emphasise that algorithmic operations embody values, assumptions and power relations that exclude and include particular features to make decisions about different aspects of the lives of individuals (Cheney-Lippold, 2011; Kitchin, 2017). Therefore, algorithms are shaped by social relations, but they also shaped the world and the life of people (Seaver, 2018; Kitchin, 2017; Hoffmann, 2019). In this sense, they deploy a performative effect that can amplify existing inequalities or/and create new ones (Eubanks, 2018; O'Neil, 2016). Within this context, algorithms tend to be defined as instruments of power (Beer, 2017; Cheney-Lippold, 2011).

Because of the highlighted features, critical researchers characterise algorithms as 'relational, contingent, contextual in nature' (Kitchin, 2017: 18; Beer, 2017). They argue that they cannot be understood as a simple set of software commands (Beer, 2017). On the contrary, they claim that they should be addressed as 'networked systems' that are 'embedded within complex socio-technical assemblages made up of a heterogeneous set of relations including (...)..individuals, data sets, objects, apparatus, elements, protocols that frame their development' (Kitchin, 2017: 20-14). This intricacy and complexity have led to emphasise algorithms as opaque black boxes elements (Pasquale, 2015; Burrell, 2016).

Algorithmic bias

Although algorithmic bias can have 'overlapping and contradictory meanings' (Crawford, 2017, Timestamp: 9:44), across the CS field it is frequently equated to the definition of computational bias. Bias is commonly defined as a type of discrimination that it is unfair and occurs 'systematically (...) against certain individuals or groups of individuals in favour of others' (Friedman and Nissenbaum, 1996: 332). Within this perspective, the concept does not refer to a single error but to a set of errors that manifest regularly and derive in an 'unfair outcome' (Friedman and Nissenbaum, 1996, p. 333). This discriminatory result can be intentional or unintentional (Wachter, *et. al.*, 2020) and have different sources (Friedman and Nissenbaum, 1996).

Biases are classified in diverse ways, however, they are usually grouped in three main categories: individual or societal, related to technical features or linked to the 'context of use' (Friedman and Nissenbaum, 1996: 335; Springer *et. al.*, 2018). The first category refers to assumptions, practices and attitudes that 'pre-exist' the functioning of the system but are dragged by humans into its design (Friedman and Nissenbaum, 1996). Technical bias is explained as unfair outcomes that arise from hardware and software, and the last category is connected to the systematic discrimination that emerges once the technology is deployed in a specific context (Friedman and Nissenbaum, 1996).

CS scholarship tends to locate the study of algorithmic biases within the technical category and, therefore, as problems of allocation (Friedman and Nissenbaum, 1996; Springer *et.al*., 2018; Barocas *et. al*., 2017). For critical scholars, the focus on biases from a technical view neglects the study of representational harms (Barocas *et. al*., 2017). That is to say, harms that are directed towards the identity of individuals and are 'more difficult to formalise' since they are linked to social structures (Crawford, 2017, Timestamp: 17:12). For instance, in 2018 Buolamwini and Gebru revealed that commercial facial recognition algorithms failed several times to identify 'dark-skinned women' (Hardesty, 2018 para.1) because of the underrepresentation of black women on datasets.

The study of representational harms (Crawford, 2017; Barocas *et. al*., 2017) requires the examination of values, ideas and assumptions that are embedded in software operations (Crawford, 2017). Biases of allocation, on the contrary, are 'easily quantifiable' and just 'a time-bound moment in the decision-making' (Crawford, 2017, Timestamp: 17:10 - 17:02). In other words, they are related to 'discrete and specific transactions' of algorithms (Crawford, 2017, Timestamp: 17:20).

While for a long time CS research has mainly concentrated on the exploration of allocative biases, recent work shows a tendency to address harms considering both categories (Buolamwini and Gebru, 2018; Costanza-Chock, 2018).

Defining accuracy

Accuracy and the fairness field: a brief contextualization

Fairness and accuracy are closely related in CS and facial recognition research. Literature review shows that fairness is a highly contested notion and there is no universal or indisputable definition. In fact, Narayanan (2018) identifies 21 ways in which fairness can be addressed, depending on which technical components are highlighted. However, the research done for this project has detected some common patterns across the construction of this notion. In general terms, fairness is addressed in CS and FRT as the 'counterpart to discrimination' (Peña Gangadharan and Niklas, 2019: 883) and a set of mathematical or statistical formulas to evaluate if data sets and algorithmic models treat groups or individuals in a similar manner (Fischer, 2019; Friedler *et. al*., 2019; Hoffman, 2019; Fish et. al, 2016).

It is relevant to emphasise that this technical conception of the notion is not exempt from criticism. Some social science researchers and critical scholars within the CS field, claim that discrimination is linked to historical and complex structures of oppression that exceed the technical features of software systems (Peña Gangadharan and Niklas, 2019, Ochigame et.al., 2018; Selbst *et.al.*, 2018).

Within CS and FRT literature accuracy is understood as a key measure of fairness. It is explained as a series of metrics and technical procedures to assess the quality of the algorithmic classifications in automated systems (Fischer, 2019). Whereas, in general, accuracy tends to be described as a constitutive aspect of fairness (Fischer, 2019), in some cases the concepts are considered synonyms.

Accuracy as a technical feature: the optimisation logic on algorithms and datasets

As explained, CS and facial recognition literature links accuracy to a set of technical procedures that aim to test the algorithmic classifications (Fischer, 2019; Grother, Ngan, Hanaoka, 2019a). Within this logic, data scientists and developers perform several operations to examine how algorithms classify. In particular, the analysis focuses on evaluating categorisations related to the demographic characteristics of individuals, such as gender and race (Fischer, 2019).

From this perspective, once algorithmic assessments are conducted, if they are not satisfactory, computer scientists introduce modifications in different instances of the process to improve the model that is used to make decisions (Pessach and Shmueli, 2020; Iosifidi, *et.al*., 2019). In this sense, as any other fairness measure, accuracy is 'task specific, and prescribes desirable outcomes for a task' (Friedler *et al*., 2016: 6). Results can therefore be enhanced until they reach a certain standard (Kusner, et. al., 2017; Friedler *et al*., 2019). Within this view, accuracy is described as a technical and internal component of automated tools that can be enhanced and managed. This idea of improvements to gain efficiency is a frequently emphasised characteristic of algorithms across CS literature (Sedgewick and Fajolet, 2013; McLaren, 1969; Hutchinson and Mitchell, 2019).

The technical features of accuracy assessments are central in empirical research on facial recognition. The Face Recognition Vendor Test (Grother, Ngan, Hanaoka, 2019a), an annual report of the NIST which evaluates facial recognition algorithms of commercial companies, is a clear example of this.

Each year, the NIST assesses the correctness of classification procedures, so that ‘face recognition system developers, and end users should be aware of these differences and use them to make decisions and to improve future performance’ (Grother, Ngan, Hanaoka, 2019b: 3).

This optimisation logic of accuracy tests (Fischer, 2019) usually results in establishing a connection between accurate algorithms, ´algorithmic fairness’ and the ‘fairness of the system’ (Ochigame *et. al.*, 2018). This tendency is particularly observed on recent papers of the ACM Fairness, Accountability and Transparency Conference (ACM FAT Conference), one of the most important conferences in CS (Kleinberg *et. al.*, 2016; Iosifidi, *et.al.*, 2019; Friedler *et. al.*, 2019). There, a vast part of researchers claim that it is possible to create ‘fairness-aware algorithms’ (Iosifidi, *et. al*., 2019), a kind of “pre-packaged algorithm ‘stamped with "fairness" (Selbst *et. al.,* 2019: 66). However, according to critical authors, technical improvements are at least worrisome, since computer scientists wrongly assume that these enhancements are ‘sufficient to achieve social ideals of fairness’ (Green and Hu, 2018, para.1). Moreover, socio-technical approaches claim that algorithms cannot be considered as pure technical elements as human choices are always involved in the data collection process and the decisions that are made to construct the set of rules on which algorithms are based (Selbst *et. al.,* 2018).

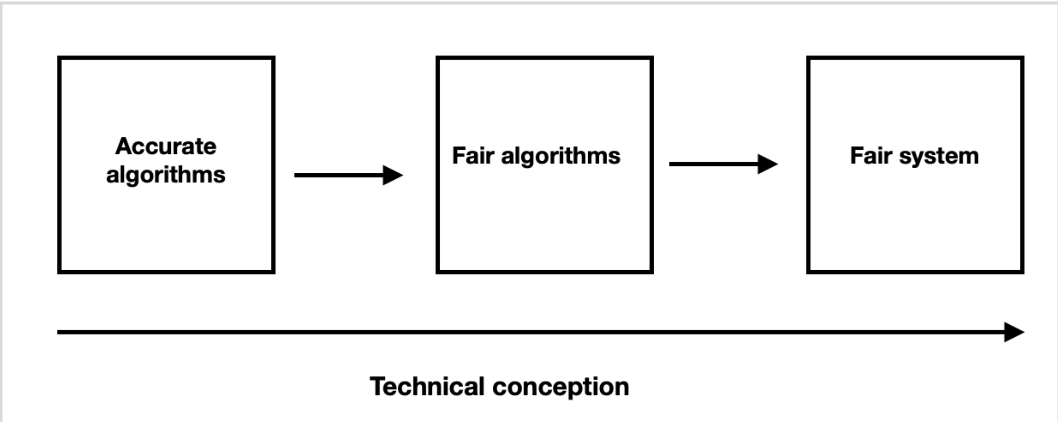


Figure 1: Accuracy as a technical element (based on: Ochigame et al., 2018; Fischer, 2019)

The idea of improvements to gain efficiency is directed not only to algorithmic operations, which are considered objective, (Pessach and Shmueli, 2018) but also to datasets (Friedler *et. al.*, 2019; Dwork *et. al.*, 2012; Fischer, 2019). Across accuracy evaluations, datasets are described as the input of algorithmic classifications. They are commonly addressed as an injection of 'real-world' data into the software system that can be managed and technically adapted to train algorithms and enhance their performance (Albarghouthi *et. al.,* 2016,: 4; Pessach and Shmueli, 2020). For instance, computer scientists usually establish a correlation between accuracy and the size of the dataset (Burrell, 2016), so when problems relating to accuracy measures arise, an increase in the volume of data is considered as the solution.

Objections to this idea of data as input from 'the real world' are raised by experts in the field of social sciences and some computer science researchers who argue that datasets are conceived as neutral elements when instead they embed values, assumptions and ideas that are always connected to a specific social context (Selbst *et. al.*, 2018; Couldry and Mejias, 2019).

The focus on the technical control of datasets is also observed in the case of labels. Data scientists internally categorise datasets using labels. Within FRT, these 'tagged' elements served to classify individuals according to their demographic characteristics (gender, race and ethnicity). In this context, labels can be modified to improve algorithmic operations (Fish, *et. al*., 2016). However, criticism is also directed to this point within literature. Some scholars argue that the commonly used labels in accuracy measures within FRT do not consider relevant aspects such as the colour of the skin or 'variations in pose' (Buolamwini and Gebru, 2018: 4). As a consequence, the data used to deploy algorithmic rules is biased even before algorithms start operating.

A paradox emerges across accuracy literature in terms of the relationship between algorithms and datasets. Both appear as related features in empirical research but are treated as separate elements. As explained at the beginning of this chapter, a relevant part of current literature argues that algorithmic input (data) cannot be understood separately from algorithmic performance since these elements are an assemblage of different components (Selbst *et. al*., 2019; Kitchin, 2017). In this regard, opposing authors (Green and Hu, 2018; Selbst et. al., 2018) claim that computer scientists abstracts social relations by separating algorithms from the 'algorithm's input', that is, the data (Selbst et. al., 2018: 60). As a consequence of this, technology is treated as independent from the 'behaviors and embedded values of the pre-existing system' (Selbst, *et. al*., 2018: 62).

Biases as technical adjustments

Low levels of accuracy are commonly understood as evidence of biases in the algorithmic model. However, due to the general focus of all fairness measures on 'engineering and technical choices' (Peña Gangadharan and Niklas, 2019: 883), they are usually explained as technical issues in data training (pre-processing), or the calibration of the algorithmic classification (post-processing) (Iosifidis *et. al*., 2019). In this sense, biases are addressed as transactional elements that can be identified and removed (Singh and Hind, 2018) through the execution of 'mitigation strategies' and 'remedies' (Wachter *et.al*., 2020: 35; Singh and Hind, 2018).

In the case of datasets, biases are generally associated with failures in sampling procedures, that is to say, the techniques used to collect data (Pessach and Shmueli, 2020). According to critical authors, however, this type of bias is only one amongst others (Selbst *et. al*., 2018). In this regard, they argue that even in those situations where procedures to obtain data are correctly adopted, the data can still embody societal biases (Suresh and Guttag, 2020) since computer scientists 'overlook the historical processes that generate data, the background assumptions of models, and the ethical and political implications of deploying models in specific social contexts' (Ochigame *et. al*., 2018, para.1). Therefore, computational results can only show some effectiveness in addressing bias in terms of 'allocation' of goods or resources – 'who gets what' - but struggle to account for bias as 'representational harm' (Crawford, 2017). In other words, from this perspective, discrimination, subordination and oppression of certain groups of society can still take place even if accuracy measures show no technical biases (Fischer, 2019).

Accuracy as statistical formulas

Apart from its technical characterisation, accuracy is simultaneously addressed as quantified results (Fischer, 2019; Ochigame *et al.*, 2018; Grother, Ngan, Hanaoka, 2019b). The literature review shows that the accuracy of automated systems is assessed through a set of statistical formulas (Fischer, 2019). To perform these tests, developers and data scientists use a table known as confusion matrix that summarises four key classification measures of the algorithmic model (Verma and Rubin, 2018; Google, n.d):

True Positive (TP)	**False Positive (FP)**
False Negative (FN)	**False Positive (FP)**

Figure 2: confusion matrix (Based on: Google, n.d)

True positive (TP) and True Negative (TN) outcomes are correct algorithmic classifications (Verma and Rubin, 2018). For instance, when an individual is correctly identified as someone wanted by the Police (TP) or when a subject is correctly discarded as someone that is not in the Police's list of suspects (TN) (Verma and Rubin, 2018). On the contrary, False Positive (FP) and False Negative (FN) outcomes refer to incorrect categorisations (Verma and Rubin, 2018). As an illustration, a FP occurs when the 'system does match a person's face to an image in a database, but that match is actually incorrect' (EFF, 2007, para. 9). A FN outcome arises 'when the face recognition system fails to match a person´s face to an image that is, in fact, contained in a database' (EFF, 2007, para.8).

Considering the confusion matrix, accuracy analyses of facial recognition focus on the rate of FP and FN, since these metrics indicate the proportion of incorrect classifications in the system (Grother, Ngan, Hanaoka, 2019b). The measures are calculated according to a threshold, a reference number ‘against which all comparisons are made’ (Grother, Ngan and Hanaoka, 2019b: 6). This can be a fixed figure, for instance 0,5, but it can also be adjusted based on the test results (Grother, Ngan and Hanaoka, 2019b). Although both measures are included in accuracy assessments, FP are considered more worrisome across the literature and reports from different organisations (EFF, 2007; Dushi, 2020; Big Brother Watch, 2018). For instance, a document from the European Union Agency for Fundamental Rights (Dushi, 2020) which examines the use of FRT in law enforcement highlights that a FP outcome ‘has crucial consequences on (...) fundamental rights’, since it can produce the arrest of innocent people (p. 7).

FP and FN rates are not only calculated over the general outcomes of a particular FRT. Computer scientists deploy these notations on several variables to make comparisons within each category (Buolamwini and Gebru, 2018; Hoffman, 2019). In FRT, variables are demographic features of individuals: gender, race, color of the skin, etc. For instance, FP and FN are measured in females and males (variable = gender).

From a technical perspective, if the algorithmic model is not biased, false positive rates in females and males are expected to be similar (Fussey and Murray, 2019). Therefore, computer scientists conclude that differences are not statistically significant. If the opposite occurs, results are understood as evidence of a biased system (Barabas *et. al*., 2020). In these latter cases, and from a CS approach, the algorithmic model is modified to enhance its performance.

Concerns about the role of statistics in accuracy measures emerge as a relevant feature across the literature. Ochigame *et.al*. (2018), for instance, highlight that ‘causal inference by itself is insufficient if researchers do not critically engage with the broad spectrum of hypothesis and interventions’ (para. 22). In this sense, the critical literature emphasises that the focus on numbers ‘creates discursive ghettos around marginalized populations via statistical discourse’ (Barabas et.al., 2020: 170). From this perspective, for instance, assessing gender from a binary perspective in accuracy metrics (male-female) reinforces oppression on the trans population (Costanza-Chock, 2018). Furthermore, critical scholars also argue that CS has primarily concentrated on the effects of algorithmic classifications on ‘socially meaningful subgroups’ (Fischer, 2019: 115). That is to say, individuals that are protected by the law. However, there is not much literature that explores other groups that are also marginalised and affected by data-driven systems (Fischer, 2019). As a consequence, even in those cases where statistical notations result in acceptable levels of accuracy - a metric that is mentioned but not defined in the CS literature - the system can still be unfair (Fischer, 2019).

A call for a socio-technical approach

Criticism to how accuracy is examined in CS in general emphasise that 'the conceptual and methodological toolkit used to evaluate the fairness of algorithmic systems remains limited to a narrow set of computational and legal modes of analysis' (Barabas *et. al.*, 2020: 168). In this regard, those with an opposing view call to go beyond technical and statistical conceptions and adopt a socio-technical perspective to explore the interconnections between social and technical features (Walker *et. al.*, 2008; Selbst *et. al.*, 2018; Fischer, 2018; Ochigame *et al.*, 2018).

This necessity appears as fundamental to address 'assumptions, choices, and considerations' that are embedded in automated models (Mitchell *et al.*, 2018: 14). In other words, to unveil power relations and consider human intervention in these processes.

Objections to the CS approach of accuracy also emphasise that individuals are abstracted from computational operations, although 'both humans and machines are necessary in order to make any technology work as intended' (Selbst, *et al.*, 2018: 60). Engineers, data scientists and those who participate in the design of systems or algorithmic models make decisions which are 'subjective, and they often include biases' (Pessach and Shmueli, 2020, para.1).

Despite the general call to adopt socio-technical perspective, the literature review done for this project did not find previous studies that examine the notion of accuracy in FRT, and specifically in LFR, while adopting this framework.

'The statistical' and 'the technical' as discourses

The critical literature on numeric reasoning highlights that statistics is a social phenomenon that produces discursive frames to construct reality (Desrosières, 2002; Porter, 1996, Beer, 2016). Within this perspective, formulas are tools that 'allow the discovery or creation of entities that support our descriptions of the world and the way we act on it' (Desrosières, 2002: 3). Therefore, when they are deployed, social elements are transformed into 'things' to portray them as 'indisputable' objects (Desrosières, 2002). From this approach, statistics relies on a consensus 'among a given group of observers or measurers' (Porter, 1996, p. 96). This agreement provides the sense of objectivity through numbers. Therefore, since figures are obtained following scientific methods, for these scholars, numbers appear to be 'fair and impersonal' (Porter, 1996: 8).

On the other hand, a vast part of social researchers within the Information and Communication Technologies (ICT) and the Information Systems (IS) fields agree with the call for socio-technical studies and emphasise that technologies are always embedded in discourse (Orlikowski and Iacono, 2001). They claim that these are artefacts formed by heterogeneous elements and intervened by discursive negotiations that integrate all the components (Orlikowski and Iacono, 2001; Hosein, 2003). In this sense, a review of research within the ICT and IS fields has shown that the exploration of the discursive dimension has been a valid entry point to examine values, ideas, beliefs, assumptions and power relations that are embedded in technology (Avgerou and Bonina, 2019; Thompson, 2004). In other words, the discursive dimension appears in these studies as a valid vehicle to connect 'the technical' and 'the social'. Nevertheless, the examination of discursive features in computational operations is still poor in the media and communications field.

CONCEPTUAL FRAMEWORK AND RESEARCH OBJECTIVE

This work takes a critical approach to analyse the notion of accuracy in the LFR reports produced by the MET Police. In doing so, it adopts a socio-technical perspective and brings into the discussion the study of the discursive dimension to understand how accuracy is constructed in connection to social relations.

Until now, the concept of accuracy in LFR has never been defined beyond the computational field. For the purpose of this study, the notion is understood as an assemblage of two discourses, the technological and the statistical, that are dialectically connected to a wider social context: human practices, ideas, beliefs, assumptions and choices. In this sense, this work assumes that technical and statistical features of accuracy never exist in a pure form. They are always embedded in discourse (Orlikowski and Iacono, 2001; Hosein, 2003, Beer, 2016) and, therefore intervened by social components that contribute to shape and yield the notion of accuracy.

The objective of this study is therefore to unveil the articulation amongst these components to understand the construction of the notion of accuracy in the LFR technology deployed by the MET Police in London.

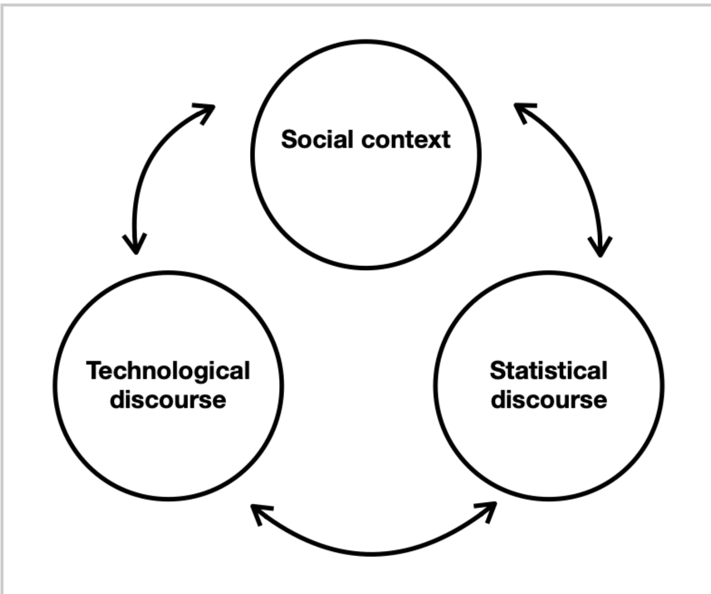


Figure 3: Accuracy from a socio-technical approach

Within this conceptual framework, algorithms as part of the technological discourse are understood as 'networked systems' that are 'embedded within complex socio-technical assemblages made up of a heterogeneous set of relations including (...) individuals, data sets, objects, apparatus, elements, protocols that frame their development' (Kitchin, 2019: 20-14). In this sense, data and labels are part of this 'algorithmic assemblage' and, therefore, also connected to social relations.

In addition, the notion of algorithmic bias is addressed here in a wide meaning. It is understood as the unfair discrimination that emerges from the assumptions, practices, choices and attitudes that are dragged by humans into automated systems before or during the construction of algorithms (Friedman and Nissenbaum, 1996). This 'systematic discrimination' can be grouped in 'harms of allocation' and 'harms of representation' (Friedman and Nissenbaum, 1996: 333; Crawford, 2017). The first type includes those biases that are easier to detect as they are commonly associated with the internal components of the system (Crawford, 2017). The second category comprises harms in connection with 'long term discrimination' which are directed to the identity of people and reinforce the oppressin on certain groups within society (Crawford, 2017, Timestamp: 17:06).

Research questions

Building upon the conceptual framework, this study aims to answer the following research questions:

RQ1: How is the notion of accuracy constructed in the LFR technology deployed by the MET Police in London?

RQ2: How are problems in accuracy explained?

RQ3: How is the notion of accuracy in the LFR tool related to other discourses?

METHODOLOGY

Critical Discourse Analysis: overview and rationale of the methodology

This research uses the Fairclough's Critical Discourse Analysis (CDA) approach to explain how the notion of accuracy is constructed around the LFR technology implemented in London. CDA is a methodological perspective that focuses on the analysis of texts as entry points to explore wider social relations that are articulated in linguistic operations (Fairclough, 2013). In doing so, it addresses the dialectical links between the internal parts of texts and the 'extra-discursive objects' to interpret how social events are constructed (Fairclough, 2013: 20).

From this perspective, discourses interact with society but also act on it (Fairclough, 2013). Thus, the study of ideologies and power relations are key to unpack those meanings that are embedded in texts constructions (Fairclough, 2013). In other words, CDA concentrates on examining the ideas, beliefs, assumptions and representations that are integrated in linguistic dimension to understand how discourses serve to maintain or alter power structures in society (Fairclough, 2013).

The appropriateness of this methodology relies on three strategic factors. Firstly, in accordance with the Fairclough's CDA approach, discourses are understood as 'representations of social life which are inherently positioned' (Fairclough, 2003: 206). In this regard, this research assumes that the notion of accuracy involves a sophisticated and carefully designed discourse that articulates ideas, beliefs, assumptions, choices and human practices that are in connection with the technological and the statistical discourses to control the conception of accuracy in LFR technology. Thus, the dialectical and therefore, relational approach of CDA helps to analyse the discourse of the MET Police in order to uncover the strategies that are deployed to produce a 'rhetorical closure' of 'what it is said' about the accuracy of this surveillance technology (Selbst *et al.*, 2019: 65; Lyon, 2003). As this 'ideological shaping of language texts' (Fairclough, 1992: 2) is a modality of power, the use of CDA will contribute to identify tactics that the Police mobilise to gain control over the definition.

Secondly, one of the most relevant advantages of adopting a CDA approach is that it is a transdisciplinary methodology (Fairclough, 2013). Thus, it is flexible enough to incorporate theoretical elements from other disciplines that are essential to the analysis of the object of study (Fairclough, 2013). As Fairclough (2005) explains, CDA implies a "dialogue between disciplines and theories with each drawing on the concepts, categories and 'logics' of the other" (p. 923). This is an essential feature for the purpose of this work since the examination of the notion of accuracy will necessarily lead to the discussion of some CS and statistical operations to enrich the interpretation of discourses and unveil power relations.

A third strong reason for using Fairclough's relational perspective is that it allows to explore the presence of other discourses or enunciative operations (Fairclough, 2013) that might be related to the notion of accuracy. Thus, the examination of the mechanisms that the MET Police use to emphasise specific features of accuracy and neutralise others may shed light on the legitimation strategies that are deployed to support the implementation of LFR. As Fairclough (2003) explains power and legitimacy are connected practices and 'textual analysis is a significant resource for researching legitimation' (p. 88). Since legitimation is also a central theme in Van Leeuwen (2007) research, this dissertation will take some insights of his work.

CDA criticism

Over the years, CDA has received a large number of criticisms within the academic field (Breeze, 2011; Stubbs, 1997; Carvalho, 2008). Some researchers have accused the methodology of 'lack of rigour' due to the relevance of interpretation procedures in the examination of discourses. Nevertheless, interpretative operations are fundamental to understand how social relations are articulated in texts that - like in the case of this research - draw on apparent objective data and evidence.

The hermeneutic process of CDA is crucial to explore how the MET Police construct meaning around accuracy (Wodak and Meyer, 2009). Iterative operations amongst different 'moments of interests in discourses' (Hosein, 2003: 10) are key to detect and capture the links between key enunciative strategies and wider social structures. In this regard, a content analysis, for instance, would not be suitable to answer the research questions of this project. Disentangling tensions, ideologies and power operations that are embedded in the accuracy notion is also fundamental to understand other constructions such as the idea that LFR is 'one amongst other tools to reduce crime' (MET Police, n.d.). In this regard, only a methodological approach with a focus on explanation - and not only description - can unveil the 'social wrongs' of our society to produce knowledge to change them (Fairclough, 2013: 8). This is the reason why the selection of CDA as a methodological framework also manifests a deep commitment that goes beyond the pages of this dissertation. It is an aspiration of a change for a more just society.

Unit of analysis and selection criteria

The corpus selected for this research comprises 5 key LFR technical documents created by the MET Police: the "MPS Live Facial Recognition Guidance Document" (MET Police, 2020b), the 'MPS the "Data Protection Impact Assessment' (MET Police, n.d.-c), 'The Metropolitan Police Service Live Facial Recognition Trials' (MET Police, 2020c), the 'Equality Impact Assessment' (MET Police, n.d.-b) and the 'MPS Response to the London Policing Ethics Panel Final Report on Live Facial Recognition Technology' (MET Police, 2020a). All of them are available through the institutional website of the MET Police in a special section.

Even though the website contains a total number of 14 documents, the selected reports were chosen for two main reasons. Firstly, because the pilot conducted on these reports showed that the documents address aspects that are crucial to comprehend the notion of accuracy in the LFR technology. They all contain evidence-based information regarding the accuracy of the technology and, therefore, constitute 'information-rich cases' (Patton, 2014: 105).

Evidence is understood here in a broad sense. It can refer to statistical models, evaluation procedures and research (Davies *et al.*, 2000. The corpus, consequently, represents 'organised knowledge' for 'ensuring that what is being done is worthwhile and that it is being done in the best possible way' (Davies *et al.*, 2000: 2-17). In other words, these reports are documents that the MET Police used to justify the implementation of the technology.

Secondly, the selection criteria relies on another strong motivation. The study of the technical reports is still poor within the media and communications studies. Thus, the focus on the evidence-based reports of the MET Police is a political call for media studies to embrace the complexity and expand the study of the discursive dimension.

Operationalisation of the technology

Although in earlier formulations Fairclough (1992) has emphasised an approach to linguistic pieces in three-levels (text, discourse and social practice), due to continuous overlappings amongst the dimensions, the author later substituted this model 'with a conceptualisation of texts and talk as part of a process of articulation" (Jørgensen and Phillips, 2011: 9). The articulation concept, which concentrates on exploring key 'moments of social practice' (Jørgensen and Phillips, 2011: 9), was borrowed from Laclau and Mouffe's theory (1985) to facilitate the interpretation of intertextuality and interdiscursive operations (Jørgensen and Phillips, 2011), in other words, to enhance the understanding of how discourses recontextualise meaning features and other discourses within a text.

This research conducts an analysis that focuses on detecting and explaining those articulation mechanisms in the MET Police reports, since these fragments are privileged instances of the construction of meaning. As the corpus selected for this project consists of evidence-based documents, that in some cases exceed 40 pages, I will be guided by the reviewed research and the conceptual framework and I will concentrate on the 'points of interest' (Hosein, 2003: 10) within these texts. In short, the fragments where the MET Police address the notion of accuracy. In these instances, the analysis will focus on the discursive construction of statistical formulas, algorithms, algorithmic operations, data, labels and biases (the statistical and the technical discourses) and their links to human practices, ideas, beliefs, choices and assumptions. In other words, since theory and method are never separate components (Van Maanen *et al.*, 2007), the diagram introduced in the conceptual framework to explain how the notion of accuracy is understood within this research (see Figure 3) also guides the operationalisation of the methodology.

Ethics and reflexivity

This research follows the Ethics Policy and Procedures of the London School of Economics and Political Science (LSE). All the documents used in this project are of public access.

As a data journalist and a student from the Data and Society programme at LSE, I have a critical background regarding surveillance technologies. Moreover, my role as a researcher is not outside the social relations that I aim to explore (Wodak and Meyer, 2009). In this sense, I am aware that my own ideas or beliefs might have influenced the analysis. During the period of this research I reflected on my interpretative process with my supervisor and I also self-reflected about my pre-conceptions of LFR. As CDA is an interpretative methodology, this work can be contested and informed by future analyses.

ANALYSIS AND RESULTS

Formulas to calculate accuracy: human intervention and control

The False Positive Identification Rate: a human-supervised metric

The MET Police reports introduce statistical notations to explain accuracy. The first discursive operation is observed before describing the formulas. Reports specify that the performance metrics are supported by 'recommendations of the standards ISO/IEC 19795 and ISO/IEC 30137' (MET Police, 2020c:. 14). The fact that these norms are mentioned, serve to deploy the effect that the obtained results are "robust" and meet the technical requirements.

One of the metrics used to measure accuracy is the False Positive Identification Rate (FPIR). The formula is calculated by subtracting the 'number of confirmed identifications' (alerts that are validated by police officers in the field) from the total number of alerts from the system and dividing the result by the number of faces that are read by cameras (recognition opportunities) (MET Police

2020c: 15). According to reports, an alert is a notification produced by the system that informs about a possible match (MET Police 2020c). The notation is constructed in documents as follows:

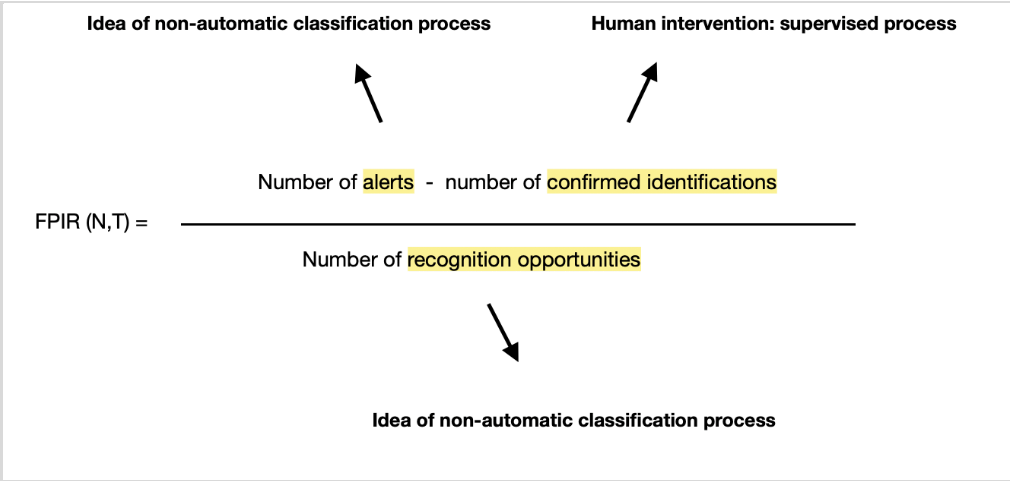


Figure 4: MET Police FPIR formula (MET Police, 2020c: 15)

The presence of police verification in the formula constructs FPIR as a human-supervised metric. In this regard, a governance mechanism known as 'human in the loop' infiltrates the notation to highlight that 'the system identifies and selects decisions, but people perform the key decision-making and actioning role' (Kitchin, 2018: 248). The Police intervention is discursively introduced as a safeguard for calculating accuracy. In this sense, the relationship between figures and police presence contributes to deploy an effect of trust over the outcome and the technology since the number obtained relies on police verification.

The formula denotes another strategy. It constructs algorithmic operations - and therefore technology - as non-automatic by emphasising that classification procedures are mediated by police officers. This 'assumed meaning' (Fairclough, 2003: 58) emerges from the human intervention, but also from the terms 'alerts' and 'recognition opportunities'. In short, 'alerts' do not necessarily imply a 'match' between subjects and those individuals that are part of the Police's watchlist; 'confirmed identifications' involves a police-verification process and, 'recognition opportunities' (faces detected by cameras) refers to 'possible future matches' but are no indication of positive recognition. The use of these terms, once again, erases the fact that alerts are already the result of classification procedures that occur inside the LFR system. Algorithms and data behind the process are obscured. Moreover, this construction abstracts human participation in the previous instances, since its intervention is explained as a post-outcome mechanism (Iosifidis *et al.*, 2019).

How can the formula be interpreted by linking statistical knowledge and discursive features? According to the trial report (MET Police, 2020c), the technology is operationalised in highly populated zones and special events - Notting Hill Carnival, Soho, Remembrance Sunday, etc - thus the number of 'recognition opportunities" (faces detected by cameras) tends to be large. Since false positives - the first part of the formula - is determined by the results of police engagement procedures after alerts, the FPIR is likely to be small. As small numbers are an indication of high levels of accuracy, the 'human in the loop' factor does not only construct the FPIR as a supervised metric, it also contributes to maintaining low error rates. This transdisciplinary dialogue (Fairclough, 2003) between textual features, statistical knowledge and wider conditions unveils how the FPIR is made valid through different operations.

The elimination of the False Negative Identification Rate formula

Although the MET Police mention the False Negative Identification Rate (FNIR) formula, the institution discards its use arguing that "Sometimes it is preferred to talk in terms of 'hit' rates" (MET Police, 2020c: 14). The selection of the word 'sometimes', however, refers specifically to the case of the LFR technology deployed in London.

As Fairclough (2003) explains, ideology and power also involve looking at what it is not said in discourses. In this sense, why does the institution explain the formula but discard its use? According to the MET Police Guidance for deployments (2020b), problems in the FNIR could be a sign of general issues in the system. In other words, the FNIR could question and jeopardise LFR. Moreover, since the formula considers the list of suspects elaborated by the Police (watchlist), the FNIR put under the microscope the dataset.

Despite these observations, the formula is mentioned. The terms 'alert' and 'recognition opportunities' appear in its construction to emphasise once again, that classification processes are non-automatic:

$$\text{FNIR (N,T)} = \frac{\text{Number of recognition opportunities by subjects on the watchlist not generating a correct alert}}{\text{Number of recognition opportunities by subjects on watchlist}}$$

Figure 5: The False Negative Identification Rate (MET Police, 2020c: 14)

The True Positive Identification Rate

The TPIR is the formula that appears in the MET Police reports as a replacement of the FNIR (MET Police, 2020c). The notation is described in CS and FRT's research as 'the hit rate' (Grother, Ngan and Hanaoka, 2019), that is, as a summary of successful matches of the technology. The MET Police (2020c) construct the formula as follows:

Facial images of police officers and staff

$$\text{TPIR} = \frac{\text{Number of correct bluelist alerts}}{\text{Number of bluelist recognition opportunities}}$$

Facial images of police officers and staff

Figure 6: MET Police True Positive Identification Rate Formula (MET Police, 2020c: 15)

A revision of the glossary of the trial report (see Appendix A) shows that the word 'bluelist' refers to police officers' images that are used to train the technology (MET Police, 2020c: 1). The technical appearance of the word and its similarity with another key term, 'watchlist', that is, the list of suspects wanted by the Police (MET Police, 2020c) hides a wider aspect: the formula does not inform 'how efficient' accuracy levels are since it is not being measured against those people who are wanted by the Police. In short, it can only tell 'the effectiveness' of the LFR in categorising police officers. The fact that the formula is 'forced' in reports denotes the relevance of statistical notations and figures in the MET Police discourse of accuracy.

Contradictions and ambivalence are expected features in the recontextualisation of discourses (Fairclough, 2003). Documents of the MET Police explicitly address the problematic use of the bluelist in the TPIR. Once again, the term 'bluelist' is introduced, but now to provide an explanation and legitimate its use:

> Determination of the True Positive Identification Rate is made based on recognition opportunities by bluelist subjects only, as the trial has no way to count the number of people on the operational watchlist that are missed by the LFR' (MET Police, 2020c: 15)

Making the contradiction visible, does not solve the problem but it serves to validate its use. In other words, by mentioning the issue, the formula remains unquestioned and the results are made valid. The construction of 'the transparency effect' is reinforced by highlighting part of the text in bold.

Flexibility and recontextualisation in accuracy formulas to yield metricsitive Identification Rate

The introduction of the words 'bluelist', 'alerts', 'recognition opportunities' and 'confirmed cases' in accuracy formulas show how statistical discourse can be adapted to control 'the accuracy phenomena' and yield the notion. Statistics, therefore, create a discursive frame that constructs accuracy measures as 'indisputable' objects (Desrosières, 2002).

Moreover, the focus on the FPIR and the TPIR and the exclusion of the FNIR show how statistical notations are recontextualised in discourse (Fairclough, 2013) to make visible some narratives and hide others (Beer, 2016).

The Optimisation logic

The Gold-Silver-Bronze command as a human safeguard for accuracy during deployments

Police intervention is not only a key feature in statistical notations. The presence of police officers is also discursively introduced to suggest that accuracy is - and can be - monitored during deployments. The 'Gold-Silver-Bronze' structure (GSB), a traditional Police strategy that defines a chain of command in the field, is key in this regard. The procedure is explained as a combination of roles and hierarchies in decision-making processes that provide strict control in the operation of LFR (MET Police, 2020b).

The GSB strategy assumes that police supervision is sufficient, but also a qualified element to manage the accuracy and technology. The Guidance for deployments (MET Police, 2020b) explains that LFR Engagement Officers do not only assess the notifications generated by the LFR tool but also 'can consider factors which may impact on the accuracy of an alert' (MET Police, 2020c: 11). The way in which the clause is constructed shows accuracy problems as issues that can be detected and anticipated by trained and experienced police officers. Moreover, accuracy incidents are linked to the context (environment) in which the deployment takes place. In this regard, this discursive operation abstracts subjects' intervention, data frames and pre-existing biases of the system (Selbst *et al*., 2019; Friedman and Nissenbaum, 1996).

When 'factors' that affect accuracy emerge, the MET Police highlights that they can be solved by adjusting the threshold: 'if during Deployment a Watchlist image generates more than one False Alert, then consideration will be given to raising the Threshold for Alerts for that Watchlist subject' (MET Police, n.d.-c: 23). As explained in the literature review, in accuracy metrics the 'threshold' is a point of reference to which comparisons are made (Google, n.d.). The clause constructs the idea that problems in accuracy are only a 'threshold incident', and thus, a technical issue that can be solved by adjusting the parameter used to measure false alerts. Therefore, the optimisation logic (Fischer, 2019) is injected in this operation.

Furthermore, in these constructions, police staff are a special kind of human beings. They emerge as 'unbiased' individuals that evaluate and determine solutions to those problems that may arise from the use of technology. This idea is reinforced in one of the reports: 'LFR operators and engagement officers receive awareness training on potential unconscious bias" (MET Police, 2020a: 6). The assumption of neutral intervention is a powerful discursive device to construct accuracy and LFR as elements that can be controlled.

The watchlist: an objective and accurate list

This watchlist comprises images of those wanted by the Police 'for a range of different offences' (MET Police, 2020c: 16). Across reports, the list is treated as a technical element that is objective and inherently accurate because it is specific to each deployment and prepared by experienced police officers (MET Police, 2020c). The idea of 'contextual construction' is portrayed as sufficient to 'ensure the currency, relevancy, necessity and proportionality by which any image is included for potential matching' (MET Police, n.d.-c: 20). The use of the nouns 'currency', 'relevancy', 'necessity' and 'proportionality' emphasise that the data is strictly controlled. However, the operation abstracts the social relations (Selbst, *et al*., 2018) in which the list is produced since the term 'contextual' is linked to 'deployments'. In short, the data frame (Selbst, et.al., 2018) is discursively removed and the dataset arises as an objective and unquestionable element.

According to reports, the watchlist can 'sometimes' be the source of accuracy results above the threshold (MET Police, 2020c). In these cases, the MET Police mobilise the statistical discourse to validate that the size of the list is what affects measures and thus justify that there are no other reasons beyond this particular factor. Therefore, problems in data are constructed as purely technical.

The trial document (MET Police, 2020c) introduces a table with the FPIR and the TPIR results by deployment to support this argument. In performing this discursive operation, the MET Police show how accuracy can be managed (see Figure 7, blue rectangle added for the analysis). However, as Crawford explains (2017), this way of addressing data issues does not account for representational harms. It is just a transactional procedure that cannot eliminate cultural biases dragged into the data by humans (Crawford, 2017, Timestamp: 17:31). For instance, datasets can still have 'societal gender biases' that are 'likely to be propagated throughout' the system and outside it (Buolamwini and Gebru, 2018: 1).

Deployment	Algorithm Version	Watchlist Size	FPIR	TPIR
Notting Hill 2016	S17	266	0.01%	54%
Notting Hill 2017	S17	528	0.09%	71%
Remembrance Sunday 2017	S17/M20	42	0.05%	89%
Hull June 2018	M20	144	0.00%	80%
Stratford Jun 2018	M20	489	0.05%	81%
Stratford Jul 2018	M20	306	0.01%	73%
Soho 17 Dec 2018	M20	2226	0.10%	74%
Soho 18 Dec 2018	M20	2226	0.10%	78%
Romford Jan 2019	M20	2401	0.10%	66%
Romford Feb 2019	M20	1996	0.13%	69%

Table 1 – False Positive Identification Rate and True Positive Identification Rate by deployment

Figure 7: Watchlist size by deployment (MET Police, 2020c: 23-24)

According to the institution, 'with a watchlist size of around 2000, TPIR for the Bluelist is around 70%, and FPIR close to 0.1%' (MET Police, 2020c: 23).Therefore, the Police conclude:

> Increading the watchlist size (…) should proportionately increase the number of subjects of interest found through LFR. To be able to do so without impacting TPIR and FPIR performance is clearly beneficial (MET Police, 2020c: 24)

Since CDA also focuses on highlighting what is assumed in discourses (Fairclough, 2013), the analysis detected a relevant feature. 70% is taken as an optimal result. However, the reason why this percentage is an acceptable number is not addressed. Moreover, some smaller watchlists such as the one used in the Remembrance Sunday and the Stratford deployment have higher levels of accuracy than larger lists of suspects. Also, the discursive operation is contradictory since in other parts of technical reports the Police explain that there is 'no baseline against which to compare performance' because LFR is a 'new technolog' (MET Police, 2020c: 7). The emphasis on the size of the dataset and the deployment of the statistical discourse also serve to neutralise possible questioning of the wider social context in which the lists are produced (Buolamwini and Gebru, 2018).

The NEC-3 algorithm: a pre-packaged element with accuracy

Reports repeat the name of the algorithm used by LFR technology to project the image that the NEC-3 is 'pre-packaged' with accuracy (Selbst, 2018: 66) since it is the best on the market. The use of the name instead of the word 'algorithm' constructs the element as a guarantee of trust for LFR. This is reinforced by the introduction of the NIST voice.

Intertextuality operations (Fairclough, 2003) show that a particular quote from the NIST document is introduced several times: 'NEC-3 is, by many measures, the most accurate we have evaluated' (MET Police, n.d.-c: 7). The repetition of this quote also emphasises another relevant operation: the focus on accuracy as a technical aspect.

The NEC-3 is conceived as an element amongst others, an operation that obscures its 'background or setting' and, thus, increases its 'social power' (Beer, 2017, 2-3). In this sense, the Guidance report (2020b) defines the algorithm as a 'system factor' (p.8) and dissociates the NEC-3 from data, institutions, and the biases that individuals might incorporate on its construction.

A further operation is observed. Human intervention is abstracted as a source of biases but incorporated as an unbiased element that can adjust the algorithm. This idea of algorithmic improvements is key on the MET Police documents to show the algorithm, as in the case of datasets, can be adapted to respond correctly to a specific deployment. For example

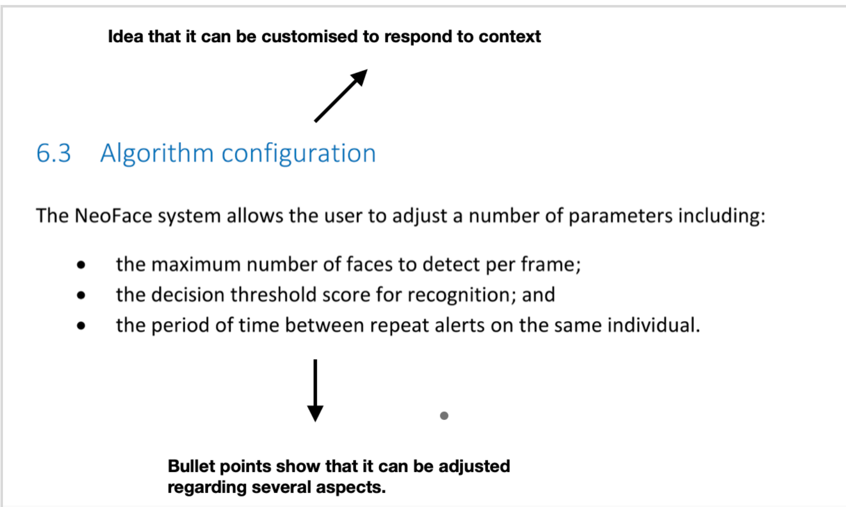


Figure 8: The algorithm as flexible element (MET Police, 2020c: 23-24)

As observed in the image, the use of the bulletpoints and the determiner 'the' emphasise the many aspects of algorithmic classification that can be modified to enhance the NEC-3 performance. However, as Crawford (2017) explains, these adjustments are only 'a time-bound moment in the decision making' and they cannot eliminate representational biases that might be encoded in the algorithm (Timestamp: 17:02).

Another discursive mechanism is also relevant in the fragment. Although no other voices manifest in the text at this particular time, repetition and stress operations can be interpreted as a response to criticisms of the media to LFR (Fairclough, 1992).

The idea of enhancements also serves to construct the NEC-3 as an element that can be modified to respond with objectivity and neutrality. The focus on 'adjustments', therefore, can be interpreted as a strategy to relativise the NEC-3 as the source of biases. This is observed in the MPS Response to the London Policing Ethics Panel (2020a). There, the document states that 'algorithmic injustice/bias within the LFR system (if any) can be effectively mitigated' (MET Police, 2020a: 5). The emphasis on the technical components and the construction of the NEC-3 as just one element of the LFR system, increases the social power of the algorithm (Beer, 2017). Moreover, it also highlights that the discourse only addresses biases as harms of allocation and not as representational harms (Crawford, 2017).

While in statistical formulas algorithmic operations are obscured, in this case, the algorithm is brought into the discussion. The mechanism highlights how visibility and invisibility strategies are deployed, according to each particular discursive context, to yield the notion of accuracy. In other words, the dialectical oscillation between technical aspects and human intervention is a key feature in the construction of accuracy in the LFR tool of the MET Police.

When accuracy fails

Differences in performance, not bias

The trial report indicates that the system produces 'more false positive alerts for men than for women' (MET Police, 2020c: 25). Texts incorporate this issue and relativise it: 'The results relate to *a single test*' (MET Police, 2020c: 27, emphasis added). Furthermore, the problem is described as inherent to FRT in general, and thus not specifically related to the LFR used by the MET Police. In this regard, documents explain that "facial recognition technologies are 'sensitive' or can present 'differences in performance' in terms of ethnicity, gender and age (MET Police, 2020c: 25).

CS research is introduced to support this argument. The MET Police (2020c) highlights two key documents: a study conducted by Klare *et al*. (2012) and the Face Recognition Vendor Test from the NIST (2019a). However, problems in accuracy measures are not addressed as biases. In the trial report, the word is avoided and replaced by the phrase 'demographic differences' (MET Police, 2020c: 27). The analysis detected that the term is made visible in only one circumstance: to respond to media criticism. In this case, journalistic pieces are delegitimised because they question factual evidence (MET Police, 2020c). They are also relativised by introducing quotation marks. For example, in the trial document, the MET Police explains:

> The results of these scientific studies [i.e. the Klare et.al. study and the NIST report] have been characterised in media reports as, for example, 'racial bias' but this is somewhat misleading as the extent of differences in performance vary by algorithm. (MET Police, 2020c: 25)

It is important to emphasise another relevant aspect. Gender classifications are assessed in the MET Police reports in binary terms: female and male. This logic perpetuates stereotypes (Crawford, 2017) and reinforces the oppression of other groups of individuals (Costanza-Chock, 2018).

Blaming the subject, not the algorithm

Research cited in evidence-based documents highlights that there are 'underlying factors which can influence' algorithms (MET Police, 2020c: 10). Those factors are directed to one specific element: the demographic of individuals. Clause constructions, however, denote a more powerful operation: subjects seem to be blamed for problems in accuracy. The trial report (MET Police, 2020c) offers a clear example of this:

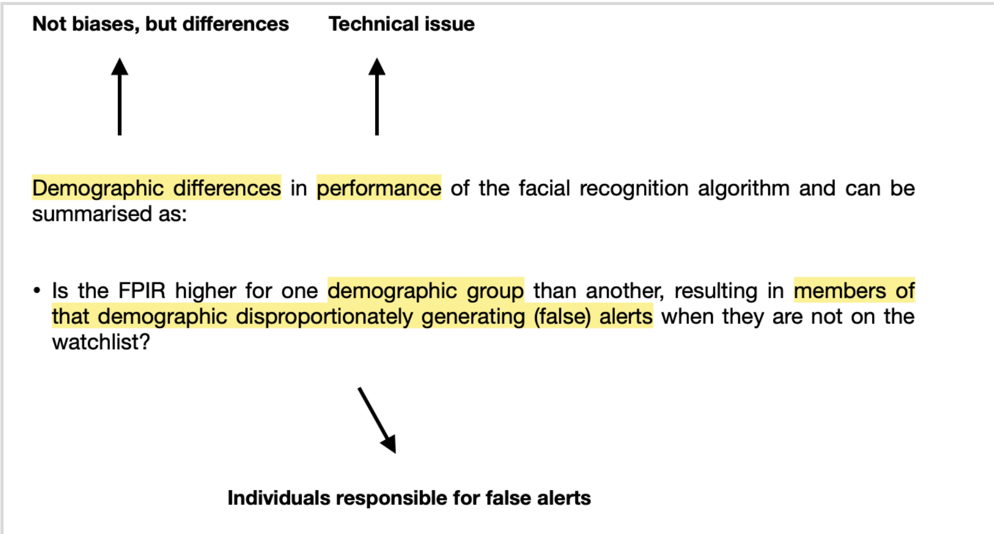
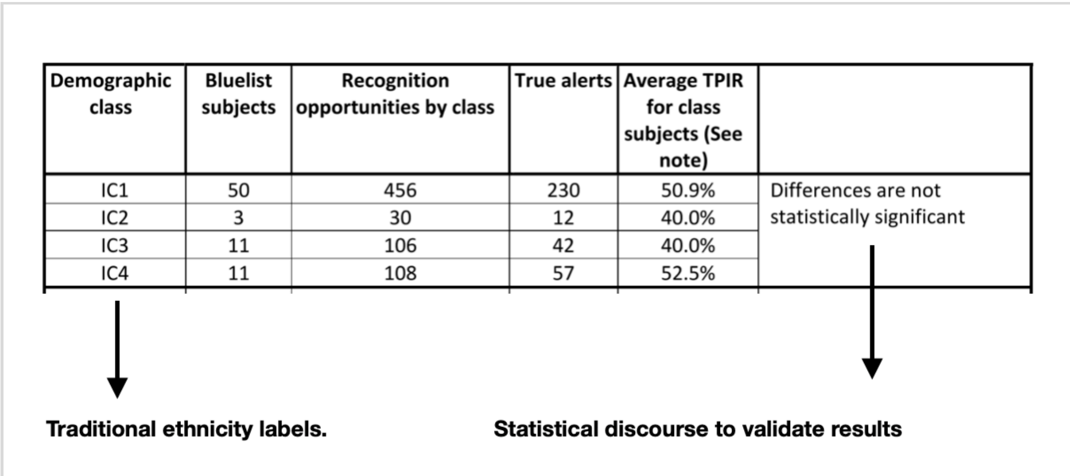


Figure 9: Individuals responsible for false alerts (MET Police, 2020c: 25)

The beginning of the fragment states that what it will be addressed are 'differences in performance' and not biases. Also, the phrase 'members of that demographic' is directly associated with 'false alerts' through the verb 'generating'. The repetition of the term 'demographic' stresses the focus on people and seems to exonerate the system from possible biases. These mechanisms naturalise the social order and inequalities (Fairclough, 2003). Furthermore, although a question mark is introduced to keep some distance from what it is about to be said, the tone of the sentence resembles a declarative sentence rather than an interrogative one (Fairclough, 2003).

Labels to classify ethnicity: not guilty

IC Codes, which are the demographic labels to categorise ethnicity, are mentioned but not conceived as sources of problems. Evidence-based reports turn to statistical discourse to highlight that these 'tags' are neutral elements in accuracy results:



Demographic class	Bluelist subjects	Recognition opportunities by class	True alerts	Average TPIR for class subjects (See note)	
IC1	50	456	230	50.9%	Differences are not statistically significant
IC2	3	30	12	40.0%	
IC3	11	106	42	40.0%	
IC4	11	108	57	52.5%	

Figure 10: IC labels as not responsible for accuracy problems (MET Police, 2020c: 25)

The above table shows some key operations. IC codes are the traditional labels used by the MET Police to classify the ethnicity of individuals (MET Police Authority, 2007). A quick revision of FOI requests of the MET Police shows that these categories are employed for most of the statistics produced by the institution. In this regard, labels are introduced in the MET Police discourse as already validated elements. This legitimation by tradition (Van Leeuwen, 2007) deploys a sense of objectivity in how ethnics are categorised. What is not said in this discursive construction is that IC codes are based on 'Visual Assessment Ethnicity' (MET Police, 2020c: 1). That is to say, officers' perceptions. Moreover, ethnic labels 'are unstable' because they 'are not constant across geographies: even within countries these categories change over time' (Buolamwini and Gebru, 2018: 4).

Although the MET Police recognise that 'assessment of IC codes may be somewhat subjective' (MET Police, 2020c: 1), their use is not questioned. Once again, the discursive visibility of the contradiction and the fact that the MET Police is 'transparent' in this regard, seems to legitimise its use.

Also, the 'authority effect' by tradition (Van Leeuwen, 2007: 96) and its combination with the statistical discourse neutralised possible criticisms to labels. However, once again, there is another contradiction. The bluelist - images of police officers - is used as the variable to calculate the results. Therefore, figures are 'not statistically significant' only in the case of police officers.

Accuracy and other discourses: LFR as a valid tool to fight against crime

Accuracy is generally addressed before or after sections that emphasise the benefits of the tool. The key findings subheading, for instance, is located after the section 'System accuracy'. There, the MET Police highlights: 'LFR will help the MPS stop dangerous people and make London safer' (MET Police, 2020c: 4). This notion is reinforced by 3 bullet points that repeatedly use the word 'help' to emphasise the advantages of the tool. The operation is also strengthened by underlying some key fragments of each clause:

> The trials indicate that LFR will help the MPS stop dangerous people and make London safer. Specifically it will:
>
> - help the police to prevent and detect crime, aiding officers to identify individuals wanted by the police and courts;
> - help the police to improve security and safety on the streets and at public events, particularly when helping to identify persons who pose a significant risk to the public;
> - help the police to protect borders and important infrastructure where criminals and other dangerous persons may try to avoid being identified. (MET Police, 2020c: 4)

The association between accuracy and the image of LFR as an effective technology to "stop dangerous people and 'make London safer' (MET Police, 2020c: 4), repeatedly appears in documents and obscures the fact that LFR is a surveillance tool. Furthermore, the term 'surveillance' appears a few times and only to refer to cameras.

Documents also show that the GSB structure assumes that police practices and knowledge are the primary factors in the decision-making process. In this sense, the use of the GSB command represents LFR as an assistance tool. Therefore, the technology is explained as 'just a tool to aid policing and does not change the core policing role…' (MET Police, 2020a: 6). This link also contributes to abstract the surveillance component of LFR.

As observed in the analysis, intertextuality operations (Fairclough, 1992) show that there is a dialectical oscillation between 'human intervention' and 'technical optimisation' to produce a 'control of topic' (Fairclough, 2013: 35). The way in which formulas are described and how algorithms, data and labels are conceived construct the image that LFR 'is doing nothing wrong' because accuracy can be perfectly managed. Human supervision and technical optimisation, however, abstract the social causes that underpin accuracy issues so that the technology remains unquestioned. Moreover, the results of statistical formulas contribute to validate the system as a technology that does not harm people, because 'differences are not statistically significant' and when they are, they are inherent to FRT in general, due to the size of the watchlist or to the demographic of individuals. In this regard, Algorithms, labels and data are exonerated as sources of bias, and when biases manifest they are relativised. In addition, the idea of 'non-automatic classification' also serves to construct the technology as non-automatic: the word 'live' is therefore dissociated from facial recognition.

The term fairness does not appear in any of the examined documents. Moreover, fairness issues are reduced to accuracy results that can always be adjusted (Grother, Ngan and Hanaoka, 2019b). Harms are only "harms of allocation". Representational harms have no place within the MET Police logic. The 'rhetorical closure' (Selbst et. al., 2019: 65) of accuracy maintains the discussion over the technical components and avoids addressing human biases.

Future research

This work revealed some unforeseen but rich discursive features that future work should consider to expand the study of accuracy in LFR technologies. One of those aspects is the relationship between accuracy and governance mechanisms. The analysis of the MET Police reports made visible how human oversight is a powerful discursive machine that infiltrates the notion of accuracy to construct it as something that can be managed, but also erase human practices in data, algorithms, labels and biases. In this sense, the role of governance elements in accuracy constructions needs to be explored in depth.

It is also relevant for future studies to maintain the focus on the vocabulary of statistical formulas. As the analysis showed, words used in their construction not only served to make numbers valid but also to sustain other discourses, such as the representation of the classification procedures as a non-automatic.

The transdisciplinary approach of CDA was key to establish a rich dialogue between statistical notations, CS features and social relations. It highlighted the importance of a more critical conversation between computer and social sciences. As this work showed, knowledge from both fields is needed to strengthen meaning interpretations and understand how FRT are legitimised for its operational deployment.

CONCLUSION

This research explored the construction of the notion of accuracy in the LFR technology used by the MET Police in London. The analysis revealed that accuracy is explained as a supervised-human metric that can be controlled and enhanced because of police intervention. Within this logic, police officers are introduced in the discursive constructions as unbiased individuals that guarantee that classifications produced by technology are accurate and non-automatic.

Police intervention is, therefore, a powerful discursive device. It obscures human practices and, therefore, biases that might be embedded at different instances of the design of the technology (Selbst, *et al.*, 2018), and serves to support the idea that technical improvements are sufficient to guarantee that the technology is accurate.

Problems in accuracy are explained as 'differences in performance' and not as biases. Moreover, they are not attributed to the algorithm used by the MET Police since it is conceived as the best on the market (MET Police, n.d.-b). On the contrary, when issues in accuracy levels emerge, they are explained as inherent to FRT, due to the size of the watchlist or to the demographic of individuals.

The analysis of the notion of accuracy showed the relevance of discursive features in supporting the deployment of the LFR technology in London. The construction of the system as a valid tool to fight against crime is intrinsically connected to the idea that technology is accurate. Moreover, this strategy obscures other narratives, such as the fact that LFR is a surveillance tool.

To conclude, this work made it evident why the study of computational concepts in relation to discourse matters. Furthermore, it emphasised the importance of addressing technical reports since these are also part of the public discourse of LFR. In this sense, a deeper commitment of media researchers is needed to explore FRT systems. Articles in the media are always one part of the story. This is why this research is also a political call to those working in the field. Media scholars need to embrace complexity, and they need to do it now, before it is too late.

REFERENCES

Albarghouthi, A., D'Antoni, L., Drews, S., & Nori, A. (2016). Fairness as a program property. *arXiv preprint arXiv:1610.06067*.

Avgerou, C., & Bonina, C. (2020). Ideologies implicated in IT innovation in government: A critical discourse analysis of Mexico's international trade administration. *Information Systems Journal*, 30(1), 70-95.

Barabas, C., Doyle, C., Rubinovitz, J. B., & Dinakar, K. (2020, January). Studying up: reorienting the study of algorithmic fairness around issues of power. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 167-176).

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *Calif. L. Rev*., 104, 671.

Barocas, S., Crawford, K., Shapiro, A., & Wallach, H. (2017). The problem with bias: from allocative to representational harms in machine learning. Special Interest Group for Computing. *Information and Society (SIGCIS).*

Barnard-Wills, D. (2011). UK news media discourses of surveillance. *The Sociological Quarterly*, 52(4), 548-567.

Beer, D. (2016). How should we do the history of Big Data? *Big Data & Society.* https://doi.org/10.1177/2053951716646135

Breeze, R. (2011). Critical discourse analysis and its critics. *Pragmatics*, 21(4), 493-525.

Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77-91).

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1), 2053951715622512.

Carvalho, A. (2008). Media (ted) discourse and society: Rethinking the framework of critical discourse analysis. *Journalism studies*, 9(2), 161-177.

Cheney-Lippold, J. (2011). A new algorithmic identity: Soft biopolitics and the modulation of control. *Theory, Culture & Society*, 28(6), 164-181.

Couldry, N., & Mejias, U. A. (2019). *The costs of connection: How data is colonizing human life and appropriating it for capitalism*. Stanford University Press.

Crawford, K. (2016). Can an algorithm be agonistic? Ten scenes from life in calculated publics. *Science, Technology, & Human Values*, 41(1), 77-92.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2009). *Introduction to algorithms.* MIT press.

Costanza-Chock, S. (2018). Design justice, AI, and escape from the matrix of domination. Journal of Design and Science. https://doi.org/10.21428/96c8d426

Davies, H. T., & Nutley, S. M. (Eds.). (2000). What works?: Evidence-based policy and practice in public services. Policy Press..

Desrosières, A. (2002). *The politics of large numbers: A history of statistical reasoning.* Harvard University Press.

Dushi, D. (2020). The use of facial recognition technology in EU law enforcement: Fundamental rights implications.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012, January). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference* (pp. 214-226).

Eireiner, A. V. (2020). Imminent dystopia? Media coverage of algorithmic surveillance at Berlin-Südkreuz. *Internet Policy Review*, 9(1), 1-19.

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor.* St. Martin's Press.

Fairclough, N. (1992). *Discourse and social change* (Vol. 10). Cambridge: Polity press.

Fairclough, N. (2003). *Analysing discourse: Textual analysis for social research.* Psychology Press.

Fairclough, N. (2005). Peripheral vision: Discourse analysis in organization studies: The case for critical realism. *Organization studies*, 26(6), 915-939.

Fairclough, N. (2013). *Critical discourse analysis: The critical study of language*. Routledge.

Fischer, F. (2019). The Accuracy Paradox of Algorithmic Classification. In *Proceedings of the STS Conference Graz 2019*.

Fish, B., Kun, J., & Lelkes, Á. D. (2016, June). A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the 2016 SIAM International Conference on Data Mining* (pp. 144-152). Society for Industrial and Applied Mathematics.

Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2016). On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236.*

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems* (TOIS), 14(3), 330-347.

Green, B., & Hu, L. (2018, July). The myth in the methodology: Towards a recontextualization of fairness in machine learning. In *Proceedings of the machine learning: the debates workshop.*

Gillespie, T. (2014). The relevance of algorithms. *Media technologies: Essays on communication, materiality, and society*, 167(2014), 167.

Goffey, A.(2008) "Algorithm." Software Studies.*Software Studies*, 2008-04-18.

Hoffmann, A. L. (2019). Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society*, 22(7), 900-915.

Hosein, I. (2003). A research note on capturing technology: toward moments of interest. In *Global and organizational discourse about information technology* (pp. 133-153). Springer, Boston, MA.

Hutchinson, B., & Mitchell, M. (2019, January). 50 years of test (un) fairness: Lessons for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*(pp. 49-58).

Iosifidis, V., Fetahu, B., & Ntoutsi, E. (2019, December). FAE: A Fairness-Aware Ensemble Framework. In *2019 IEEE International Conference on Big Data* (Big Data) (pp. 1375-1380). IEEE.

Jørgensen, M. W., & Phillips, L. J. (2011). Discourse analysis as theory and method. In *Discourse Analysis as Theory and Method*. Sage.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807.*

Kitchin, R. (2017). Thinking critically about and researching algorithms. *Information, Communication & Society*, 20(1), 14-29.

Kitchin, R. (2018) Governance. In *Ash, J., Kitchin, R. and Leszczynski*, A. (eds) Digital Geographies. *Sage, London*. pp. 238-249

Kowalski, R. (1979). Algorithm= logic+ control. *Communications of the ACM*, 22(7), 424-436.

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. In *Advances in neural information processing systems* (pp. 4066-4076).

Laclau, E., & Mouffe, Chantal. (1985). *Hegemony and socialist strategy : Towards a radical democratic politics*. London: Verso.

Lyon, D. (2003). *Surveillance as social sorting privacy, risk, and digital discrimination*. London: Routledge.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2018). Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. *arXiv preprint arXiv:1811.07867*.

MacLaren, M. D. (1969). The Art of Computer Programming—Volume 1: Fundamental Algorithms (Donald E. Knuth). *SIAM Review*, 11(1), 89-91.

Narayanan, A. (2018, February). Translation tutorial: 21 fairness definitions and their politics. In *Proc. Conf. Fairness Accountability Transp*., New York, USA (Vol. 1170).

Ochigame, R., Barabas, C., Dinakar, K., Virza, M., Ito.J.(2018). Beyond Legitimation: Rethinking Fairness, Interpretability, and Accuracy in Machine Learning. In *ICML 2018: The Debates, at the 35th International Conference on Machine Learning*. ICML 2018.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.

Orlikowski, W. J., & Iacono, C. S. (2001). Research commentary: Desperately seeking the "IT" in IT research—A call to theorizing the IT artifact. *Information systems research*, 12(2), 121-134.

Patton, M. Q. (2014). *Qualitative research & evaluation methods: Integrating theory and practice*. Sage publications.

Pasquale, F. (2015). *The black box society*. Harvard University Press.

Peña Gangadharan, S., & Niklas, J. (2019). Decentering technology in discourse on discrimination. *Information, Communication & Society*, 22(7), 882-899.

Pessach, D., & Shmueli, E. (2020). Algorithmic fairness. *arXiv preprint arXiv:2001.09784.*

Porter, T. M. (1996). *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton University Press.

Seaver, N. (2018). What should an anthropology of algorithms do?. *Cultural anthropology*, 33(3), 375-385.

Sedgewick, R., & Flajolet, P. (2013). *An introduction to the analysis of algorithms*. Pearson Education India.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 59-68).

Springer, A., Garcia-Gathright, J., & Cramer, H. (2018, March). Assessing and Addressing Algorithmic Bias-But Before We Get There. In *AAAI Spring Symposia*.

Stubbs, M. (1997). Whorf's children: Critical comments on critical discourse analysis (CDA). *British studies in applied linguistics*, 12, 100-116.

Crawford, K.. (2017). The Trouble with Bias - NIPS 2017 Keynote - Kate Crawford #NIPS2017 [Video]. YouTube. https://www.youtube.com/watch?v=fMym_BKWQzk

Thompson, M. P. (2004). ICT, power, and developmental discourse: A critical analysis. *The Electronic Journal of Information Systems in Developing Countries*, 20(1), 1-25.

Van Leeuwen, T. (2007). Legitimation in discourse and communication. *Discourse & communication*, 1(1), 91-112.

Van Maanen, J., Sørensen, J. B., & Mitchell, T. R. (2007). The interplay between theory and method. *Academy of management review*, 32(4), 1145-1154.

Verma, S., & Rubin, J. (2018, May). Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)* (pp. 1-7). IEEE.

Wachter, S., Mittelstadt, B., & Russell, C. (2020). Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *Available at SSRN*.

Walker, G. H., Stanton, N. A., Salmon, P. M., & Jenkins, D. P. (2008). A review of sociotechnical systems theory: a classic concept for new command and control paradigms. *Theoretical issues in ergonomics science*, 9(6), 479-499.

Wodak, R., & Meyer, M. (2009). Critical discourse analysis: History, agenda, theory and methodology. *Methods of critical discourse analysis*, 2, 1-33.

Non-academic references:

Big Brother Watch. (2018). Face off: The lawless growth of facial recognition in UK policing. Retrieved from https://bigbrotherwatch.org.uk/wp-content/uploads/2018/05/Face-Off-final-digital-1.pdf [Last consulted: September 1, 2020].

Dodd, V. (2020). Met police to begin using live facial recognition cameras in London. *The Guardian*, 24 January: https://www.theguardian.com/technology/2020/jan/24/met-police-begin-using-live-facial-recognition-cameras [Last consulted: September 1, 2020].

Electronic Frontier Foundation.(2007). Street-level surveillance. Face recognition. Electronic Frontier Foundation, 24 October, URL: https://www.eff.org/es/pages/face-recognition [Last consulted: September 1, 2020].

Fussey, P., & Murray, D. (2019). Independent Report on the London Metropolitan Police Service's Trial of Live Facial Recognition Technology. Retrieved: http://repository.essex.ac.uk/24946/

Gayle, D. (2020). Met police deploy live facial recognition technology. *The Guardian*, 11 February: https://www.theguardian.com/uk-news/2020/feb/11/met-police-deploy-live-facial-recognition-technology [Last consulted: September 1, 2020].

Google.(n.d.) Machine learning crash course. Classification: True vs False and Positive vs Negative. Retrieved from https://developers.google.com/machine-learning/crash-course/classification/true-false-positive-negative [Last consulted: September 1, 2020].

Grother, P., Grother, P., Ngan, M., & Hanaoka, K. (2019a). Face Recognition Vendor Test (FRVT) Part 2: Identification. US Department of Commerce, National Institute of Standards and Technology.

Grother, P., Ngan, M., & Hanaoka, K. (2019b). Face Recognition Vendor Test (FVRT): Part 3, Demographic Effects. National Institute of Standards and Technology.

Hardesty, H. (2018). Study finds gender and skin-type bias in commercial artificial-intelligence systems [Blog post]. MIT News, 11 February, URL: https://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212 [Last consulted: September 1, 2020].

Kayser-Bril, N. (2019). At least 11 police forces use face recognition in the EU, AlgorithmWatch reveals. AlgorithmWatch. Retrieved from https://algorithmwatch.org/en/story/face-recognition-police-europe [Last consulted: September 1, 2020].

MET Police Authority (n.d.).The code systems used within the Metropolitan Police Service (MPS) to formally record ethnicity. Retrieved from http://policeauthority.org/metropolitan/publications/briefings/2007/0703/index.html [Last consulted: September 1, 2020].

MET Police (n.d.-a). Live Facial Recognition. Retrieved from https://www.met.police.uk/advice/advice-and-information/facial-recognition/live-facial-recognition/ [Last consulted: September 1, 2020].

MET Police (n.d.- b). Equality Impact Assessment. Retrieved from https://www.met.police.uk/SysSiteAssets/media/downloads/central/advice/met/facial-recognition/equality-impact-assessment.pdf [Last consulted: September 1, 2020].

MET Police (n.d-c). Data protection impact assessment (DPIA). Retrieved from https://www.met.police.uk/SysSiteAssets/media/downloads/central/advice/met/facial-recognition/data-protection-impact-assessment.pdf [Last consulted: September 1, 2020].

MET Police (2020a).MPS Response to the London Policing Ethics Panel Final Report on Live Facial Recognition Technology. Retrieved from https://www.london.gov.uk/sites/default/files/mayor_of_london_-_lfr.pdf [Last consulted: September 1, 2020].

MET Police (2020b). MPS LFR Guidance document. Guidance for the MPS Deployment of Live Facial Recognition Technology. Retrieved from https://www.met.police.uk/SysSiteAssets/media/downloads/force-content/met/advice/lfr/mpf-lfr-guidance-document-v1-0.pdf [Last consulted: September 1, 2020].

MET Police (2020c).Metropolitan Police Service Live Facial Recognition Trials. Retrieved from https://www.met.police.uk/SysSiteAssets/media/downloads/central/advice/met/facial-recognition/met-evaluation-report.pdf [Last consulted: September 1, 2020].

Singh, A. & Hind, M. (2018, September 7). "AI Fairness 360: Attacking Bias from All Angles!". [Blog post].IBM Developer. Retrieved from https://developer.ibm.com/blogs/2018/09/19/ai-fairness-360-attacking-bias-from-all-angles/ [Last consulted: September 1, 2020]

APPENDIX A: LFR Glossary of terms

1 Glossary of Terms

The terms defined within this section and as used throughout this report apply to this joint report only. The terminology and definitions may not apply to those adopted elsewhere by the MPS.

Alert: A notification of a possible match between a facial image from an individual present, and a facial image on the watchlist, where the comparison score exceeds the specified threshold.

Adjudication: The process of human assessment of an alert to decide whether to engage further with the individual matched to a watchlist image.

Bluelist: A partition of the watchlist containing facial images of police personnel (officers and/or staff) for use in the setup and evaluation of LFR comprising a set of subjects with controlled presence in the Zone of Recognition.

Live Facial Recognition (LFR): Real-time, automated facial recognition using video surveillance cameras.

False Negative Identification Rate (FNIR): The proportion of recognition opportunities of subjects who are on the watchlist that do not generate a correct alert. FNIR effectively indicates the number of [subjects] `missed' by the LFR system. The complement of FNIR is the **True Positive Identification Rate**

False positive alert: An alert that is not confirmed as a correct match for a subject's true identity. In this report false positive alerts include the following cases: alerts refuted by corroborative police checks, alerts dismissed in adjudication, and cases where requested engagement with the subject failed.

False positive identification rate (FPIR): The frequency of false positive alerts among recognition opportunities for individuals not included in the watchlist. Note that, in the operational context, it can be assumed that only a very small proportion of recognition opportunities will be for individuals on the watchlist.

IC code: Visual assessment ethnicity code, used by the MPS to record the perceived ethnicity of people [1]:

IC1: White – North European
IC2: White – South European
IC3: Black
IC4: Asian – Indian subcontinent
IC5: Chinese, Korean, Japanese, or other Southeast Asian
IC6: Arab or North African
IC9: Unknown

Assessment of IC codes may be somewhat subjective; different observers may sometimes assign a different IC code to the same individual. IC codes do not necessarily correspond to self-defined ethnicity and/or declared ethnicity.

APPENDIX B: Evidence of the analysis

As explained in the methodological chapter, the corpus selected for this project consists of evidence-based documents that, in some cases, exceed 40 pages. For this reason, the material of this appendix represents fragments of those "moments of interests" (Hosein, 2003) in the analysed documents. Each report is introduced by its name to demonstrate that the fragments correspond to different documents.

Text 1: Metropolitan Police Service. Live Facial Recognition Trials

SCARE-QUOTES
RELATIVISATION
CLAUSE CONSTRUCTION
INDIVIDUALS CAUSE DIFFERENCES IN PERFORMANCE
AUTHORITY

7 The effect of subject demographics

Several studies (e.g.,[16][17]) report that performance of facial recognition is sensitive to demographics such as ethnicity, gender and age. The results of these scientific studies have been characterised in media reports as, for example, 'racial bias' but this is somewhat misleading as the extent of differences in performance vary by algorithm, and by application and these differences are observed with respect to both FPIR and TPIR. Further studies (e.g.,[18][19][20]) report that, in an operational context, differences in performance can be as much to do with image quality issues, as the underlying demographics of the subjects. These concerns lead the MPS to consider the potential demographic effects on performance within the MPS LFR evaluation and undertake a (non-operational) trial, specifically to measure the effect of subject demographics on system performance.

ACADEMIC RESEARCH IS INTRODUCED TO VALIDATE WHAT IT WILL BE SAID

THE VOICE IS INTRODUCED TO RELATIVISE IT.

NOT BIASET

Demographic differences in performance of the facial recognition algorithm and can be summarised as:

QUESTION MARK BUT DECLARATIVE CLAUSE

- Is the FPIR higher for one demographic group than another, resulting in members of that demographic disproportionately generating (false) alerts when they are not on the watchlist?
- Is the FNIR higher for one demographic group than another, resulting in members of that demographic being less likely generate an alert, even when they are on the watchlist?

BLAMING THE SUBJECT

EMPHASIS

7.1 Subject demographics and FPIR

Demographic details (IC codes) of individuals present in the area of the LFR trial were recorded for the 6th to the 10th deployments. This allows an estimation of the FPIR by demographic (IC code or Gender) as shown in Table 2. Statistical analyses[2] show that the observed differences in FPIR by perceived ethnicity are not statistically significant. However, the gender difference in FPIR is statistically significant with significantly more false alerts for men than for women.

TRADITIONAL ETHNICITY LABELS USED BY THE MET POLICE

*

Demographic class	Estimated recognition opportunities by class	False alerts by class	Observed FPIR for class	Comment on statistical significance
IC1	24,800	23	0.09%	Differences are not statistically significant
IC2	2,300	2	0.09%	
IC3	6,100	5	0.08%	
IC4	6,700	2	0.03%	
IC5	2,900	0	0.00%	
IC6	400	0	0.00%	
Female	21,550	10	0.05%	Difference is statistically significant
Male	21,650	22	0.10%	
Total	43,200	32	0.07%	

Table 2 – False Positive Identification Rate by demographic

STATISTICAL DISCOURSE

IDEA THAT LABELS ARE NEUTRAL BECAUSE OF STATISTICAL RESULTS

BINARY CLASSIFICATION

* LEGITIMATION BY TRADITION

THEY REPRODUCE INEQUALITIES

7.2 Subject demographics and TPIR

The operational trial deployments used a small bluelist, comprising only staff involved in the setting up and running of the trial. A proposal was made to expand the bluelist in terms of size and ethnic diversity to provide the additional data needed to help in consideration of the effect of subject demographics on system performance.

The initial intention had been to use an expanded watchlist within one or more of the remaining operational deployments, where bluelist subjects would join the flow of the public to be representative of the operational use case and the other trial parameters. It proved impossible to coordinate availability of a suitable bluelist

CONTEXTUAL ADJUSTMENTS

WHAT ABOUT TRANSGENDER PEOPLE, FOR INSTANCE

[2] For analysing whether differences in observed FPIR between different demographic classes are statistically significant a chi-squared test for homogeneity of proportions was used. In this report, the MPS has used a significance level of 0.1 for reporting that a difference in performance is statistically significant.

STATISTICS

25

Idea of improvements

ACCURACY → MORE THAN FIGURES

4.2 Performance metrics[1]

ACADEMIC RESEARCH TO LEGITIMIZE THE ARGUMENT

The evaluation has been conducted to follow the requirements and recommendations of the standards
ISO/IEC 19795 on Biometric performance testing and reporting [9][10][11], and
ISO/IEC 30137 Part 1 & Part 2 on the Use of biometrics in video surveillance systems [12][13].

There is no single figure that can be used to describe the 'accuracy' of a facial recognition system in any meaningful way. The standards mandate reporting performance of identification systems in terms of the frequency of **two** error conditions of the identification process; false positives and false negatives. The error rates will be measured over recognition opportunities, i.e. the period that a subject is walking through the Zone of Recognition.

IDEA OF NON-AUTOMATIC

The **False Positive Identification Rate** (FPIR) is the proportion of recognition opportunities of subjects who are not on the watchlist which generate an alert:

QUANTIFICATION

$$\mathrm{FPIR}(N,T) = \frac{\text{Num. recognition opportunities of subjects not on the watchlist that generate an alert}}{\text{Num. recognition opportunities of subjects not on the watchlist}}$$

*

where N represents the size of the watchlist, and T the threshold that the comparison score must exceed for an alert to be generated.

STATISTICAL DISCOURSE

The **False Negative Identification Rate** (FNIR) is the proportion of recognition opportunities of subjects who are on the watchlist which don't generate the correct alert.

*

$$\mathrm{FNIR}(N,T) = \frac{\text{Num. recognition opportunities by subjects on the watchlist not generating a correct alert.}}{\text{Num. recognition opportunities by subjects on watchlist}}$$

FNIR states the "miss" rate. Sometimes it is preferred to talk in terms of "hit" rates. The complement of FNIR is the **True Positive Identification Rate** (TPIR).

$$\mathrm{TPIR}(N,T) = 1 - \mathrm{FNIR}(N,T).$$

"IN THIS CASE"

WHY?

NO SPECIFICATION

* This is not the formula they used. They have adopted it (see next page)

They discarded this formula

* FORMULA INTRODUCED BUT NOT USED.

TECHNICAL DISCOURSE

[1] Performance results given in this report pertain to a single FR vendor and one particular model for LFR implementation. Performance for other facial recognition software may be different not least as this is an evolving technology.

USE OF THE NAME OF THE SYSTEM → GUARANTEE OF TRUST

to achieve optimal parameters in the desired Zone of Recognition (for example feedback on the inter-eye distance in pixels, and whether this meets recommendations).

The NeoFace system utilised for the trial was the Rapid Deployment Alien Ware laptop configuration that can support two concurrent camera streams. It was observed that when the crowd density increased the real time processing could start to 'lag'. Although higher resolution cameras are available, the hardware configuration is not sufficient to continuously process the stream (detect faces, create a template and search against the watchlist) in near real time. With advances in processing technology, it is expected that this issue would be resolved.

→ USE OF THE NAME OF THE FACIAL RECOGNITION SYSTEM

AUTHORITY

6.3 Algorithm configuration

→ FLEXIBILITY

The NeoFace system allows the user to adjust a number of parameters including:

- the maximum number of faces to detect per frame;
- the decision threshold score for recognition; and
- the period of time between repeat alerts on the same individual.

BULLET POINTS AND "THE" TO REINFORCE IDEA OF ADJUSTMENT

For all deployments the system was configured to detect a maximum of 5 faces per frame. Non-operational system tests had established that adjusting this number any higher would result in the processing servers reaching maximum capacity and system lag.

→ TECHNICAL CONCEPTION

Adjusting the decision threshold downwards may increase the TPIR but can have the effect of increasing the number of alerts that must be dealt with. Likewise, adjusting the threshold upwards can mean that subjects might walk past the camera but not generate an alert. For operational deployments, the threshold used was the default recommended by the manufacturer and this proved effective at maintaining a manageable level of false positive alerts against a sufficient True Positive Identification Rate.

NON-AUTOMATIC CLASSIFICATION

When there is an alert between an individual and a watchlist subject, additional or repeated alerts for that watchlist image are supressed for a configurable period of time to allow the individual to clear the Zone of Recognition. This means that the comparison score logged for each alert is that of the first recognition opportunity that scores above threshold. This may not be the highest comparison score possible for the recognition opportunity and as a consequence, the effect of using a higher decision threshold cannot be inferred from the logged comparison scores.

→ TECHNICAL ENHANCEMENT

ACCURACY PROBLEMS

"THE THRESHOLD CAN BE ADJUSTED"

The decision threshold score for recognition can be adjusted per watchlist or per watchlist subject image. The advantage of the latter feature is that the threshold can be adjusted upwards for watchlist subject images that generate a high number of alerts (for example due to Pose, Illumination or Expression factors) as observed at Notting Hill Carnival in 2017.

→ TECHNICAL DISCOURSE

6.4 Facial recognition performance

FPIR and TPIR (the False-Positive and True-Positive Identification Rates) for each deployment are shown in Table 1. For the later deployments, with a watchlist size of around 2000, TPIR for the Bluelist is around 70%, and FPIR close to 0.1%. It should be noted that the environmental conditions, and camera parameters (e.g. face size for subjects in Zone of Recognition) were fairly similar in these deployments, and that the performance levels achieved will differ in other conditions.

TECHNICAL CONCEPTION

Deployment	Algorithm Version	Watchlist Size	FPIR	TPIR
Notting Hill 2016	S17	266	0.01%	54%
Notting Hill 2017	S17	528	0.09%	71%
Remembrance Sunday 2017	S17/M20	42	0.05%	89%
Hull June 2018	M20	144	0.00%	80%
Stratford Jun 2018	M20	489	0.05%	81%
Stratford Jul 2018	M20	306	0.01%	73%
Soho 17 Dec 2018	M20	2226	0.10%	74%
Soho 18 Dec 2018	M20	2226	0.10%	78%

USE OF THE STATISTICAL DISCOURSE TO SHOW THAT IF THE WATCHLIST IS INCREASED, ACCURACY IMPROVES

HOWEVER, SOME CONTRADICTIONS APPEAR WHEN LOOKING AT NUMBERS.

Text 2: Equality Impact Assessment

NOT PROTECTIVELY MARKED

GUIDANCE

STEP 5d. EIA Action Plan Template Service Delivery Impacts (External)

Potential positive / negative issues / impacts	Activity	Role Holder	Action By Date	Progress/Timescale/ Monitoring
AGE: Impacted LFR Alerts are based on the mapping of key facial indicators when comparing an inputted face image to an individual's face entering or passing through a Zone of Recognition. Therefore, the functionality, accuracy and performance of LFR may be less effective if changes to facial appearance have occurred between the time the Watchlist image was taken, and the time the subject enters the Zone of Recognition. The age of criminal responsibility starts at 10 years old. Image capture via Custody Imaging, on which LFR technology is reliant, is dependent on the age, date and time at which the custody image was taken. In addition, the European Union's Agency for Fundamental Rights 'Facial Recognition Technology Fundamental Rights Considerations in the Context of Law Enforcement Report 2019' highlights that as a child grows and time passes, the accuracy of a biometric match can diminish. The risk of a failure to match increases when facial images recorded at a young age are compared more than five years after they were collected. The report further indicates that the accuracy of facial recognition technology *is in general* significantly lower for children younger than 13 years old. They associate this to "rapid growth and change in facial appearance". The images uploaded to LFR can be from a wide variety of sources. For example, an image of a	**THE MPS ACCEPT THIS IMPACT BASED ON THE FOLLOWING CRITERIA:** • The MPS use of an algorithm that the National Institute of Standards and Technology (NIST) have tested. They noted that "NEC-3, is on many measures, the most accurate we have evaluated". • The criteria is set out in the Proportionality, Legality, Accountability, Necessity 'P.L.A.N' Assessment • LFR is a tool that assists police officers locate wanted people. LFR does not qualify as formal identification and does not make decisions that result in any person being spoken to. It provides a guide to officers about which people passing the system may be of interest to them. Officers then consider the Alert using their experience and training before the Engagement Officer makes any decision to Engage with a person. This includes consideration about whether age is a factor in generating an Alert. Even where an Engagement occurs, further action is not an automatic consequence, the officer would need a lawful basis to take any further action (such as an arrest). • The pre-requisites on which LFR Deployments must meet a legitimate aim. These will be articulated in the MPS LFR Form 1 and are set out in the LFR Legal Mandate and Section 4 of the LFR Standard Operating Procedure.	LFR Team	Complete	

→ VOICE OF THE NIST

→ REPEATED FOCUS ON THE ALGORITHM AND NIST EVALUATIONS

SCARE-QUOTE TO VALIDATE ARGUMENT.

→ ASSISTANT TOOL

DISCOURSE THAT APPEARS WHEN ACCURACY IS MENTIONED.

→ NON-AUTOMATIC CLASSIFICATION

→ NON-AUTOMATIC TOOL

→ Police officers assess the alerts (unbiased human beings)

Page 32

IDEA THAT THE NEC-3 PRODUCES "STABLE" FALSE POSITIVES

NOT PROTECTIVELY MARKED

This is the third in a series of reports on ongoing face recognition vendor tests (FRVT) executed by NIST.

As part of the Aims and Scope of the report NIST has conducted tests to quantify demographic differences in contemporary face recognition algorithms. The report concludes;

→ BIASES AS MITIGATION

Identification Algorithms: The presence of an enrolment database affords one-to-many algorithms are source for mitigation of demographic effects that purely one-to-one verification systems do not have. We note that demographic differentials present in one-to-one verification algorithms are usually, but not always, present in one-to-many search algorithms. One important exception is that some developers supplied identification algorithms for which false positive differentials are undetectable. Among those is Idemia, who publicly described how this was achieved. **A further algorithm, NEC-3, is on many measures, the most accurate we have evaluated.** Other developers producing algorithms with stable false positive rates are Aware, Toshiba, Tevian and Real Networks. These algorithms also give false positive identification rates that are approximately independent of the size of enrolment database.

→ NAME OF THE ALGORITHM → AUTHORITY → GUARANTEE OF TRUST

Home Office Biometrics Strategy

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/720850/Home_Office_Biometrics_Strategy_-_2018-06-28.pdf

	STATISTICAL DISCOURSE	REPETITION OF THE QUOTE ON SEVERAL FRAGMENTS OF THE REPORT EVERY TIME ACCURACY IS QUESTIONED NIST LEGITIMATION STRATEGY TO MAKE THE STATEMENT VALID	noted that "NEC-3, is on many measures, the most accurate we have evaluated". ▪ During trials that the MPS conducted when referencing demographics, there were no statistical significance to suggest a demographical deferential. ▪ During trials the MPS did identify differentials with regards to gender. The results relate to a single test and show that the LFR system is less likely to trigger alerts in relation to women who pass an LFR camera. The MPS's operational use of LFR uses a more recently released algorithm which has been tested by NIST. The NIST tests show universal statements about 'bias' are not supported by testing. HUMAN INTERVENTION ▪ Social observation indicates women change (e.g. make-up) and/or obscure (use hoods/umbrellas) their appearance more frequently and significantly than men. ▪ Highlighted the reporting of racial bias within the National Press. However, this bias generally relates to FR algorithms for categorizing subjects by ethnicity, not to LFR algorithms matching passer-by images against a watchlist. BIAS "THE MEDIA" ▪ The ability to monitor and review diversity data to analyse disproportionality.
Frontline Policing Headquarters	Consult on the implementation of LFR on their impacted business area.	14/01/2020	Supported the concept of LFR in meeting safeguarding policing priorities. Consulted to attain statistical data on crime area to check whether the Necessity criteria for LFR implementation is met.
Violent Crime Taskforce	Consult on the implementation of LFR on their impacted business area.	14/01/2020	Supported the concept of LFR in meeting safeguarding policing priorities. Consulted to attain statistical data on crime area to check whether the Necessity criteria for LFR implementation is met.
Safeguarding Command	Consult on the implementation of LFR on their impacted business area.	14/01/2020	Supported the concept of LFR in meeting safeguarding policing priorities. Consulted to attain

Text 3: MPS Response to the London Policing Ethics Panel Final Report on Live Facial Recognition Technology

Community engagement	The MPS will undertake a programme of community engagement to continue to build trust, help people understand how LFR is used, and the safeguards in place to ensure LFR is not used disproportionately. The LPEP report identified that public support for the use of LFR in relation to serious crime, was between 81% and 83% of respondents and, that in general terms, more than half thought that the police use of LFR could be acceptable. The MPS's community engagement strategy seeks to build on this position – the LPEP findings help confirm that in the right circumstances, the public recognise that LFR is a valuable policing tool.

PUBLIC SUPPORT OF THE TOOL

IMPORTANCE OF ORDER OPERATIONS ?
ACCURACY DISCOURSE
LINKED TO

2. **Building trust by making trial data public**

How the MPS has addressed this

14. The MPS will publish an evaluation report regarding its LFR trials. This report will include data relating to each trial, algorithmic bias and the learning points identified. The MPS will also publish the legal mandate, guidance document, standard operating procedures and other relevant documentation to allow the public to see how the MPS has carefully responded to the trials and implemented the points of learning identified.

USE OF THE WORD "DIFFERENCES" NOT BIAS.

15. The MPS has carefully considered issues regarding bias and algorithmic injustice. With regards to the algorithm and software used in the MPS trials, there was no disproportionality across any particular ethnic group with regards to the generation of false alerts. The MPS has observed differences in the way the LFR algorithm responds to gender. The results relate to a single test and show that the LFR system is less likely to trigger alerts in relation to women who pass an LFR camera. Wider independent testing has been undertaken by the National Institute of Standards & Technology (NIST) on facial recognition algorithms with respect to demographic differentials (Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects, NISTIR 8280, December 2019). The NIST tests show universal statements about 'bias' are not supported by testing. Instead they found that it is very important that a highly accurate algorithm is selected for use. NIST tests have reported that "the NEC-3 algorithm is, on many measures, the most accurate we have evaluated". The algorithm used by the MPS is the NEC NeoFace algorithm.

NIST VOICE IS INTRODUCED

"THE ALGORITHM IS ACCURATE"

RELATIVISATION STRATEGY

16. In light of the MPS trials and the NIST report, the MPS has carefully considered the issue and has concluded that algorithmic injustice / bias within the LFR system (if any) can be effectively mitigated so that policing decisions are made in compliance with the Public Sector Equality Duty. In this respect the MPS has built-in to its processes a number of important safeguards. These include:

IDEA EMBEDDED IN STATISTICAL FORMULAS

MPS Measure	How the measure addresses the LPEP Condition
The adjudication process	When the LFR system generates an alert, it is not an automatic consequence that any member of the public will be engaged by the MPS. Adjudication means that an officer, and not the LFR system, makes the decision about whether to engage a member of the public. An officer's decision to carry out an engagement will utilise their training, policing experience, and the information from the LFR system tempered by their awareness of the factors that impact on accuracy of an LFR alert. Relevant factors include camera configuration, camera location, lighting conditions, the distance at which people will pass the LFR system, and points relating to an individual's age and appearance.

5

	Ultimately, no engagement will be made unless an officer is content that it is underpinned by a lawful policing purpose. Furthermore, even when an engagement with a member of the public does occur, it does not follow that further police action will result – the officer requires a legal basis for any action taken.
Police officer experience GUARANTEE OF ACCURACY POLICE OFFICERS AS UNBIASED HUMAN BEINGS	The police are trained to engage with the public in the pursuit of their enquiries. There is a well-established legal basis by which officers can ask questions of members of the public. In this sense, LFR is doing nothing new. It is just a tool to aid policing and does not change the core policing role or necessarily result in a disadvantage where an engagement occurs. ASSISTANCE TOOL UNBIASED VIEW When deciding to engage with a member of the public, officers are still required to exercise their judgement based on their experience. This element of judgement is the same as when an officer decides to approach a member of public without the aid of LFR. Officers are also supported by their command structure. This ensures they make the decision to approach members of the public in the right circumstances, in the right way.
Indicators of unconscious bias NON-AUTOMATIC CLASSIFICATION	LFR operators and engagement officers receive awareness training on potential unconscious bias. Given the importance of the adjudication process to the Public Sector Equality Duty, it is important that MPS personnel do not unduly bring a bias into the deployment which could disproportionately impact certain members of the public. Further details relating to this will be made public in the MPS's LFR documents via its website.
Post-deployment review	The LFR post-deployment review provides an opportunity to review LFR deployments. The reviews can be used to identify where LFR system performance may not have met expectations, and will consider the effectiveness of the safeguards that were in place, identifying changes that can be implemented to improve the LFR system. This is an ongoing process to help ensure compliance with the Public Sector Equality Duty.

17. Recognising the need for continued public engagement, the MPS will make data available to the public via our public facing website on a deployment-by-deployment basis so that the benefits and results of LFR can be scrutinised.

3. Necessity and proportionality

How the MPS has addressed this

18. The MPS has carefully considered the legal position clearly articulated in R (on the application of Edward Bridges) v The Chief Constable of South Wales [2019] EWHC 2341 (Admin). In all areas the MPS has created a framework to use LFR that allows it to comply with the Bridges decision.

Media@LSE MSc Dissertations Series

The Media@LSE MSc Dissertations Series presents high quality MSc Dissertations which received a mark of 76% and above (Distinction).

Selected dissertations are published electronically as PDF files, subject to review and approval by the Editors.

ISSN: 1474-1938/1946