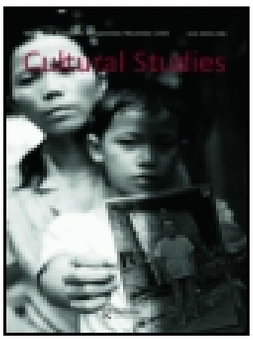


Cultural Studies

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/rcus20

Seeing infrastructure: race, facial recognition and the politics of data

Nikki Stevens & Os Keyes

To cite this article: Nikki Stevens & Os Keyes (2021): Seeing infrastructure: race, facial recognition and the politics of data, Cultural Studies, DOI: 10.1080/09502386.2021.1895252

To link to this article: https://doi.org/10.1080/09502386.2021.1895252

Published online: 26 Mar 2021.

CULTURAL STUDIES
https://doi.org/10.1080/09502386.2021.1895252

Routledge
Taylor & Francis Group

Seeing infrastructure: race, facial recognition and the politics of data

Nikki Stevens [a] and Os Keyes [b]

[a]Human and Social Dimensions of Science and Technology, Arizona State University, Tempe, AZ, United States; [b]Human-Centered Design & Engineering, University of Washington, WA, United States

ABSTRACT

Facial recognition technology (FRT) has been widely studied and criticized for its racialising impacts and its role in the overpolicing of minoritised communities. However, a key aspect of facial recognition technologies is the dataset of faces used for training and testing. In this article, we situate FRT as an infrastructural assemblage and focus on the history of four facial recognition datasets: the original dataset created by W.W. Bledsoe and his team at the Panoramic Research Institute in 1963; the FERET dataset collected by the Army Research Laboratory in 1995; MEDS-I (2009) and MEDS-II (2011), the datasets containing dead arrestees, curated by the MITRE Corporation; and the Diversity in Faces dataset, created in 2019 by IBM. Through these four exemplary datasets, we suggest that the politics of race in facial recognition are about far more than simply representation, raising questions about the potential side-effects and limitations of efforts to simply 'de-bias' data.

KEYWORDS Facial recognition; datasets; surveillance; racialization; critical biometric consciousness; infrastructure

Introduction

Facial recognition technology (FRT) is increasingly deployed in policing, border control and other domains of securitization and control. Beyond general concerns about surveillance, campaigners against FRT have frequently pointed to the *discriminatory outcomes* of that surveillance, or the ways in which the design and use of FRT systems disproportionately subject (often racial) minorities to particular observation and sometimes-violent intervention (Introna and Wood 2004, Buolamwini and Gebru 2018, Snow 2018). One line of response to these issues has been to treat them as an accident; an aberration; a technical problem. Harms are blamed on FRT being inaccurate for populations underrepresented in the datasets used to develop the technologies. As a consequence, the solution is *inclusive*

CONTACT Os Keyes okeyes@uw.edu

representation, and inclusion, with both industry and academic practitioners calling to enroll more Black and brown faces into facial recognition datasets—with the intent of making systems 'work' for marginalized populations (Buolamwini and Gebru 2018).

The writers of this article should be treated as joint first-authors.

A second—and more critical—line of thought begins with situating FRT in context, rather than seeing it as an isolated and ahistorical technology. As Browne (2015) and Beauchamp (2019) note, biometric technologies such as FRT are often *not* 'new' at all—these technologies are merely the latest in a long history of systems of surveillance and control. Browne traces the lineage of FRT as the 'surveillance of blackness,' while Beauchamp explores the way that responses to surveillance that urge cooperation and enrollment serve to legitimise the tools and techniques, all while further marking those who do not or cannot comply (Browne 2015, Beauchamp 2019). Rather than leap on seemingly-simple explanations for harms, Browne calls for a 'critical biometric consciousness … [of] informed public debate around these technologies and their application' requires these technologies to be placed in the context of their history and also requires an investigation of 'who and what are fixed in place—classified, corralled, and/or coerced' (Benjamin 2016).

In this paper, we historically and contextually ground our understanding of FRT datasets—the circumstances of their creation, use and deployment. How can this historicized understanding inform debate over the use and regulation of FRT? With this deeper understanding, can we move past representation-focused protestations that insist that FRT will be less harmful when everyone's face is equally represented within the source data? Taking up Browne's call, and drawing on a range of research publications, government reports and documents retrieved through the Freedom of Information Act, we attend to the history of artifacts *within* facial recognition systems, specifically through examining the creation of 4 exemplary facial recognition datasets.[1]

Facial recognition technology (FRT) is a subset of computer vision, itself a type of artificial intelligence (AI).[2] FRT is not a single technology but an umbrella term for a set of technologies that provide the ability to match an unknown face to a known face. This technology has been used for purposes both eye-catching and seeminglyquotidian, including security, marketing, and assessing classroom attendance (Castillo et al. 2018, Leong 2019). Crucially, there is no unified or universal FRT system. Rather, the technology consists of a shifting web of programmers, algorithms, datasets, testing standards, formatting requirements, law enforcement agents and other operators and users, and consequently, a shifting *form*. FRT is 'a complex structure that has no unifying essence and is continually being 'put together'' (Sellar 2015). One way of examining this process of putting together, and the structures that result, is through examining FRT's component *parts*, which themselves often fall under Star & Ruhleder's notion of 'infrastructure':

standardized, multi-site technologies or tools (Star and Ruhleder 1996, Star 1999, Bowker and Star 2000). By necessity—through easing some flows of life and foreclosing others—infrastructures fundamentally embody, and perpetuate, particular politics or values, something readily apparent in the infrastructure of security (Andersson 2016). While researchers have traced the overall politics of FRT (Introna and Nissenbaum 2010), few have looked in depth at the politics of specific component parts, despite the recognition that (speaking of infrastructures generally) '[E]ach subsystem inherits increasingly as it scales up, the inertia of the installed base of systems that come before' (Bowker and Star 2000, p. 33).

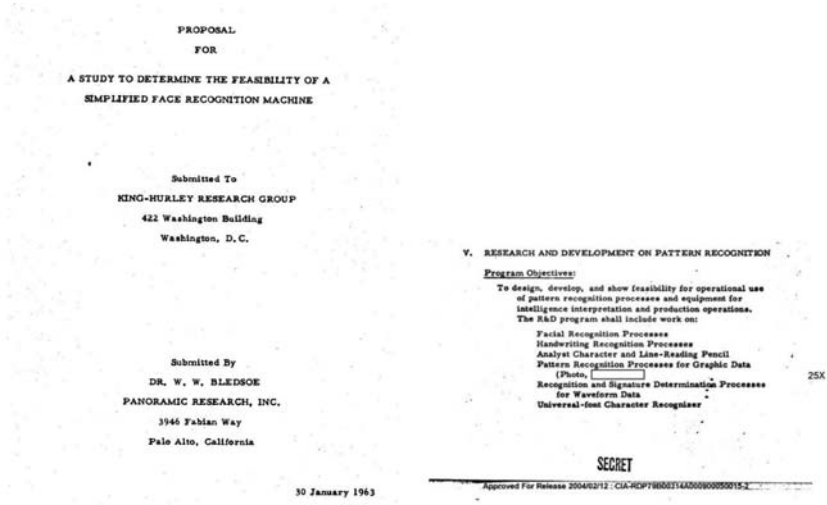
In order to build a facial recognition system, researchers depend on datasets of human faces which they can use to train a model as to what a face 'looks' like and what components of it to prioritize, analyse and consider. These datasets can contain anywhere from hundreds to millions of photographs and unique subjects, in a variety of poses, angles and lighting conditions, and with associated metadata such as the subject's gender, age or race.[3] Due to the complexity and expense of gathering such large datasets, they are frequently shared publicly, discussed and evaluated by researchers (Abate et al. 2007), and often considered a sufficient academic contribution to merit a publication in and of themselves (Guo et al. 2016, Salari and Rostami 2016). They blur boundaries between industry, law enforcement and the academy, with academic datasets often incorporated into commercial and law enforcement products, or vice versa. The reuse of datasets, and their deployment to bridge different worlds, has consequences: as Kitchin & Lauriault note when discussing data more broadly, datasets are 'expressions of knowledge/power, shaping what questions can be asked, how they are asked, how they are answered, how the answers are deployed, and who can ask them' (Kitchin and Lauriault 2018, p. 6). Studying datasets as a site of knowledge/power can help us understand the broader politics of facial recognition, and locate FRT as as cultural vehicle perpetuating a particular trajectory of state power through visuality. With this work, we are not offering a comprehensive understanding of FRT datasets. Instead, we draw on techniques and lenses from science and technology studies and critical data studies to examine the composition of these datasets and the process(es) through which they came to be. By working to (in the words of Leigh Star) 'unearth the dramas inherent in system design' (Star 1999), we demonstrate the value of examining datasets as a key technology informing the broader politics of FRT.[4]

FRT datasets

W.W. Bledsoe and the invention of facial recognition

Cultural theorists have long recognized that 'state power is intimately tied to visuality' (Wood 2016, p. 228), and facial recognition technology can be seen

as the logical result of 'a rationality of government that understands security in terms of visibility' (Hall 2007, p. 320).[5] This intimate connection was visible from the inception of facial recognition research—initiated and funded by the CIA. In 1963, W.W. Bledsoe, Helen Chan and Charles Bisson (hereafter, 'the Bledsoe team') began the first recognized research on facial recognition at Panoramic Research Institute. As noted in Bledsoe's obituaries, and repeated regularly elsewhere, this work was undertaken on behalf of an 'unnamed intelligence agency' (Ballantyne et al. 1996). As shown in Figure 1, the Bledsoe team's original project proposal was addressed to the 'King-Hurley Research Group'—a known CIA front company (Champion 1998). During that time the CIA was also establishing internal programs to 'design, develop, and show feasibility for operational use … for intelligence interpretation and production operations … Facial Recognition Processes' (DFR 116). After a year of work, in 1964, the Bledsoe team delivered their final prototype report to the CIA. Bledsoe himself continued working on the problem at the Stanford Research Institute, with a team including Peter E. Hart, eventually handing it over to them entirely (Ballantyne et al. 1996). After Bledsoe's work, the CIA continued to drive the SRI team's work, while in parallel, the CIA built their own in-organisation hardware and technical



PROPOSAL
FOR
A STUDY TO DETERMINE THE FEASIBILITY OF A
SIMPLIFIED FACE RECOGNITION MACHINE

Submitted To
KING-HURLEY RESEARCH GROUP
422 Washington Building
Washington, D.C.

Submitted By
DR. W. W. BLEDSOE
PANORAMIC RESEARCH, INC.
3946 Fabian Way
Palo Alto, California

30 January 1963

V. RESEARCH AND DEVELOPMENT ON PATTERN RECOGNITION

Program Objectives:

To design, develop, and show feasibility for operational use of pattern recognition processes and equipment for intelligence interpretation and production operations. The R&D program shall include work on:

Facial Recognition Processes
Handwriting Recognition Processes
Analyst Character and Line-Reading Pencil
Pattern Recognition Processes for Graphic Data (Photo, [redacted])
Recognition and Signature Determination Processes for Waveform Data
Universal-font Character Recognizer

25X

SECRET

Approved For Release 2004/02/12 : CIA-RDP78B03714A000900050015-2

(a) W.W. Bledsoe's original project proposal to the Hurley Research Group', a CIA front com-tems.

(b) Proposed CIA Analysis Division program for 'King-1965, including funding for facial recognition syspany.

Figure 1. Documents confirming the Central Intelligence Agency (CIA)'s involvement in directing and funding early facial recognition research. (a) W.W. Bledsoe's original project proposal to the(b) Proposed CIA Analysis Division program for 'King-Hurley Research Group', a CIA front com-1965, including funding for facial recognition syspany. tems.

expertise around FRT throughout the 1960s (DFR 112). The U.S. government has continued actively developing facial recognition technology since this initial project, both through funding academic research in the area and building their own internal systems (Kaufman 1974, Gragg 1997)

When the Bledsoe team began researching computer vision to recognize human faces, they needed a dataset of faces. In the early 1960s, there were no formal datasets of faces other than mugshots. However, the team opted to take their own photographs.[6] Photographs in this first facial recognition dataset were comprised largely of whitepresenting men wearing collared shirts (see Figure 2). We cannot know with certainty why they opted not to use existing photographs, but this choice gave them complete control over the the parameters of the image elements of the photograph (like distance from the camera). The collection of this dataset was performed in their research facility, in low-stakes, high-touch environments. The researchers interacted with the subjects individually, and no effort was made to anonymise them; the photographs were even stored under the subject's last name (Lee-Morrison 2018). These early dataset participants retained their full personhood. They were neither anonymised nor reduced to numbers.

In order to help this early computer learn to 'see' the faces of these willing subjects, the faces needed to be translated into something that the computer could understand. As a result, the Bledsoe used x,y coordinates to mark the location of facial features ('landmarks') on each photo they took, spending an average of 40 s examining and marking each photo (Lee-Morrison 2018). The systems failed with even minor lighting differences between the photos (Wayman 2007), and because of the manual classification required, were highly onerous to use. This coordinate-based classification and manual labelling resonates strongly with the history of photography in policing— particularly the use of (racialised) anthropometric measurements rather than faces themselves (a use we will see reappear), and the motivation to not replace but *replicate*, albeit more efficiently, human faculties (Cole 2009). The researchers expected that the computer would 'see' faces in the same way that they believed humans saw faces. Beyond their control of photographic conditions, the Bledsoe team did not perform computational interventions to standardize their subjects or their photographs; they simply chose subjects who fit standardized profiles.

The Bledsoe team passed more than their ways of seeing into the system. They also passed their racial biases. 'Despite being a depiction of variability, the selection is focused on a particular sociological grouping: younger to middle-aged Caucasian men' (Lee-Morrison 2018). The Bledsoe team used the best photographic technology available at the time, including Shirley cards—images of white women used to calibrate light meters.[7,8]

Figure 2. A photograph of the (surprisingly intimate) process Bledsoe's team used to gather their dataset of faces. (DFR 092)

FERET

Over the next twenty years, computer vision researchers made significant progress in developing FRT systems. Rather than a single project by a single computer science lab, facial recognition research grew to include a range of teams at different universities, each pursuing their own ways of designing and testing systems. The result was a range of semi-independent approaches, each laboratory undertaking their own inquiries and using their own datasets. This proved both a boon and a hindrance. Most positively,

the wide range of efforts produced many different techniques for tackling technical challenges. But this broad range of techniques was of limited applicability when the different methods could not be put into conversation with each other. Because the teams used widely different datasets, annotation processes and testing regimens, comparing the resulting models was less 'apples to oranges' and more apples to automobiles.

The U.S. government was a common sponsor to many of these projects and in 1995, launched a program to standardize FRT development and testing. They wanted to allow for different algorithms to be meaningfully compared, and put into conversation, in order to both establish the state of the technology and identify (and incentivise collaboration around) particularly promising ways forward. The vehicle for this standardization was *FERET*, a collaboration between the Defense Advanced Research Projects Agency (DARPA) and Counterdrug Technology Development Program was launched in 1995. Run by the Army Research Laboratory, with P. Jonathon Phillips serving as technical agent (DFR 002), the *FacE REcognition Technology* (FERET) project's purpose was to develop (and apply) a standardized testing methodology— and accompanying it, produce a standardized dataset for both that testing, and the initial and further development of a system.

The collection process for this dataset (known simply as the 'FERET dataset') produced 8,525 images, covering 884 individuals, with photography undertaken at both U.S. Army facilities and George Mason University. The practices around photography showed a distinct shift in intimacy and humanization compared to Bledsoe's work; each subject was faced with not a close, conversational photographer, but a staff of three, assembling equipment in a standardized way for a standardized series of shots and instructing the subject on pose and position. For variation, subjects were instructed to remove or add glasses, or rearrange their hair. Once taken, the photographs were shipped to Kodak for digitization and stored—or as the reports put it, 'sequestered by the government'—not under the subject's name but under an Armyprovided ID number (Phillips et al. 1997, 2000, Moon and Phillips 2001). As with Bledsoe, participation was voluntary and (to a certain degree) consensual, although there are questions about how consent works where one party consists of government officials and university faculty, and the other students and individuals who have signed an oath to obey military orders. Also as with Bledsoe, issues of race and representation appeared; 'some questions were raised about the age, racial and sexual distribution of the database'—questions that were dismissed due to the emphasis on algorithmic performance at scale (DFR 002).

Collecting such an expansive dataset was not just a matter of greater resourcing, but of new techniques for automating the process of annotating photographs. While Bledsoe attempted to automate the recognition of

human faces, their work never reached the level of human-independent automation. Teams working after Bledsoe also focused on removing human intervention in the recognition process.[9] In these early attempts, researchers were modeling computer vision to see faces *like humans saw faces* and focused on the human evaluation of facial features like 'shade of hair, length of ears, lip thickness' (Turk and Pentland 1991, p. 72). The manual labour involved in FRT systems was drastically reduced in 1987, when Sirovich and Kirby developed 'eigenfaces': essentially, a mechanism by which a large dataset of human faces could be used to probabilistically generate the locations and nature of common features. Rather than manually tagging where particular facial features were located on each and every image, researchers could train an algorithm to identify common, prominent locations, and treat those as features (Sirovich and Kirby 1987).

An eigenface algorithm works by creating an expected 'face' based essentially on the 'average' location of each major facial feature: eyes, nose, mouth. This was a major departure technologically—it allowed the computer to work 'independently' and pushed the computer away from working as an 'automated human' to working as a computer. As is visible in Figure 3, the 'eigenface' result is not designed to be legible to a human eye, nor to be 'identifiable' as a particular individual. Techniques such as eigenface generation played an important role in enabling the generation and use of FERET, removing a large amount of manual labour from the process of annotating images. But they simultaneously moved computer vision toward the logics of big data— where faces are only useful in aggregate with others. Additionally, the development of the eigenface algorithm allows humans enrolled in face datasets to become simply 'data.' As the eigenface image demonstrates, individuals were 'standardized' and the 'success' of removing the need for human markup also removed the intimacy between researcher and researched. Developers and researchers no longer needed to look at every face (for nearly a minute, as the Bledsoe team did), but could simply collate the faces into fodder for the algorithm.

MEDS-I and MEDS-II

While FERET provided a dramatic improvement in the scale and consistency of available FRT datasets, the demands of new algorithms quickly outstripped what it and new eigenface algorithms could provide. As models became more complex, demanding a wider array of variables and values computed from each photograph, it fell victim to what researchers call the 'curse of dimensionality'. More variables means more *possible values*, and in order to adequately cover all of them, 'the demand for a large number of samples grows exponentially' (Li and Jain 2015, p. 207)

Generating those exponentially-expanding datasets *de novo* would involve a constant, monumental expenditure of resources for subject

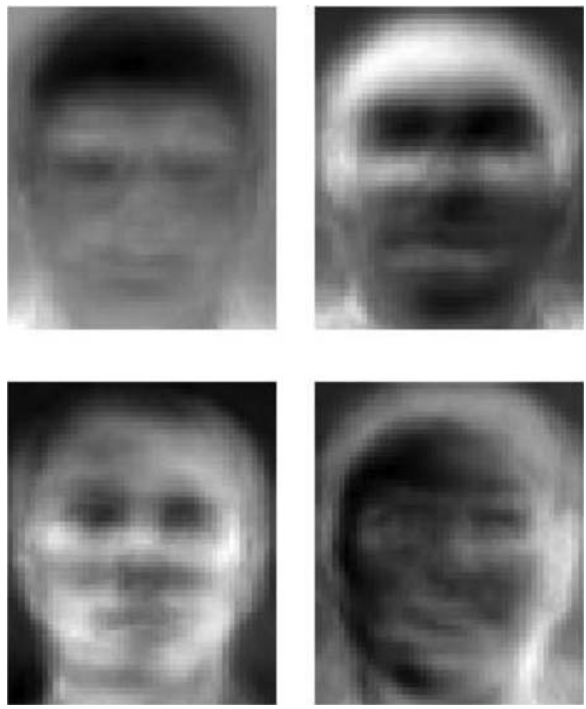


Figure 3. A diagram produced by the Army Research Laboratory team to advertise how many FRT companies had sprung from participation in FERET (DFR 420).

recruitment alone. Far easier would be redirecting existing data collection efforts, and existing flows of captive bodies. Over the late 1990s and early 2000s, researchers worked on precisely that; redesigning existing infrastructures and workflows to meet the requirements of FRT systems. In particular, they focused their efforts on a location where photography and carceral practices meet: mugshots.

In the United States, mugshots are taken of newly-arrested people as a matter of course. The specifications for these photographs in the 1990s—which were 'very widely implemented and used within and between the law enforcement communities of the United States and the many other countries' (Li and Jain 2015, p. 318)—were the 'Type 10' standards. Initially fairly broad, the standards were heavily modified by a committee that included National Institute of Standards and Technology (NIST) delegates and representatives of various companies with an interest in FRT. The new version specifically incorporated information (such as pose, angle and facial feature coordinates) that were useful in facial recognition (DFR 491). These were adapted as required practice by the FBI in 2000 (Li and Jain 2015, p. 316). Far from researchers manually annotating painstakingly collected

images as Bledsoe was forced to, photographs would now be regularly produced with values, variables and feature locations calculated by street-level bureaucrats, long before researchers encountered them. Table 1.

These adaptions directly connected the US police state to facial recognition technology by making existing state surveillance and photography work a viable source of FRT training data—as evidenced by our next example, the innocuously-named 'MEDS dataset'. Prepared by the MITRE Corporation for NIST, the MEDS dataset was released in two chunks (MEDS-I and MEDS-II) in 2009 and 2011 respectively (DFR 061, DFR 068). Taken together, the datasets released contained 518 people, represented by 1,217 photographs. This was a small portion of the overall dataset, which contained images of 1.8 million people and was prepared for a NIST-run biometric test program in 2010 intended to 'assist the FBI and partner organizations refine tools, techniques and procedures for face recognition' (DFR 043, DFR 068). Subjects in MEDS were not voluntarily enrolled, nor did they consent after the fact—such a thing would be impossible, as revealed by the database's full name: the 'Multiple Encounters (Deceased Subjects) Dataset.' The MEDS dataset is a curated collection of mugshots culled from state and federal databases. Subjects were included if they had been arrested multiple times and were deceased at time of curation. In other words, deceased, arrested people—whether convicted or not, under laws that may or may not have been valid in the mid-2000s—had their mugshots taken and reused, in some cases up to 40 years after their arrest, for the purpose of further refining law enforcement tools of surveillance and control.

Given the racialised nature of the U.S. carceral system, it is unsurprising that race appears prominently in the composition and discussion of the dataset. In contrast to some claims that facial recognition datasets do not contain enough Black faces, MEDS *over*represented African-American subjects, as seen in Table 2. While AfricanAmericans make up 12.8% of the U.S. population, they make up 40% of the subjects and, with disproportionate targeting and re-arrest rates, 47% of the photographs. This data troubles notions of 'fixing' facial recognition through increasing the presence of faces of people of color—taken on the surface, such proposals would treat the

Table 1. An example set of facial recognition datasets, from (Masi et al. 2018).

Name	Subjects	Pictures/Videos	Availability
Facebook	4,030	4.4 Million	Private
CASIA	10,575	494,414	Public
Google	8 Million	200 Million	Private
VGGFace	2,622	2.6 Million	Public
UMDFaces	8,277	367,000	Public
MS-Celeb-1M	100,000	10 Million	Public
VGGFace2	9,131	3.31 Million	Public
IMDb-Face	1.7 Million	59,000	Public

MEDS dataset as a *success* story. Yet taken within its context (the U.S. carceral system), it is clearly nothing of the sort; indeed, as Amade M'Charek notes in her discussion of race in genomics, the oppressive context of this data itself gives rise to questions of 'phenotypic othering', the 'racialization of specific groups of people based on a heightened visibility in specific political situations' (M'charek 2014, M'charek 2020).In other words, the harms of facial recognition can neither be understood as caused by, nor solveable through, addressing dataset representativeness. The issue is not the data alone, but the context of heavily racialised policing and incarceration that both motivates this technology and shapes its deployment.[10]

Demographic overrepresentation was hardly the only way in which racialised differences appeared. The process and formatting of the images was also racialised. Unlike the Bledsoe and FERET datasets, the MITRE researchers were uninvolved in collecting the images, which were instead provided by local police departments and in many cases decades old. As a consequence, images were of inconsistent format and quality, and so subject to 'normalization and correction:' reorienting the images and ensuring the metadata was consistent. This was largely undertaken with 'a combination of government, commercial, and custom … tools' (DFR 068), and allowed the researchers to ensure the dataset was standardized.

This drive for standardization required the elimination or coercion of inconsistent data. In the case of biographic data – race, gender, age–provided with the images, the researchers noted that there were inconsistent values for the same subject, 'presumably due to input error or inconsistent information collection from subjects who may not have cooperated with the process' (DFR 068). When presented with such inconsistencies, the researchers 'corrected' the race and gender values, based on their own judgment. While these inconsistencies may have been due to input error, they may also have been caused by the fluid and changing way that individuals may experience their gendered, raced identities (Jones and McEwen 2000).

Further, the automated matching of images also produced racial differences, to the point where—when investigating examples of false positives—MITRE notes that 'all [false positives] in the set are African-American males', with their software disproportionately more likely to identify African-American men as 'looking the same' even when there is no

Table 2. Racial composition of the MEDS-I database used in MBE 2010.

Race	U.S. Population (2016, %)	Subjects (%)	Photographs (%)
Asian-American	5.2%	0.78%	0.70%
African-American	12.6%	40.62%	47.47%
Native American	0.8%	0.26%	0.28%
Unknown	NA	1.82%	1.40%
White	73.3%	56.51%	50.14%

resemblance whatsoever.[11],[12] A racial difference also made itself known in the software used to correct and reorient photographs: the error rate for images with White faces was less than 10%; the error rate for African-American faces, over 30% (DFR 068).

Images that could not be automatically corrected and marked were manually adjusted and annotated—particularly in the case of 'extreme subject expressions.' In the MEDS dataset documentation, the document authors demonstrated 'extreme subject expression' with an image of a (now deceased) middle-aged African American man screaming and looking away from the camera.[13] By including this image as a problem of tagging and correction, the authors demonstrated what we might have suspected— the instrumentalisation and dehumanization of the photographed subjects. The image of the screaming man was problematized for its inability to be automatically corrected, rather than as a smoke signal pointing to the fires of injustice and abuse within the U.S. carceral system. The MEDS dataset contains many images of arrestees who are bloody, bruised and bandaged. This is not remarked on, and evidently poses no issue for the purposes of the dataset's developers and users—blood and bruises are not part of their remit unless it interferes with the algorithmic gaze.

Diversity in faces

In the decade after the MEDS datasets' creation, facial recognition became both increasingly widely deployed (Smith 2016, Bud 2018), and increasingly widely criticized, particularly on the subject of race (Stark 2018, Bacchini and Lorusso 2019, Cook et al. 2019). From a variety of angles and sites of inquiry, these critiques pointed to the ways that computer vision and technical classification systems create oppressive racializing results. As these race-centered critiques grew, multi-national corporations attempted to offer 'solutions' to the racialized impacts of facial recognition surveillance, specifically the demographic makeup of facial recognition training data. Perhaps ignorant of it's own role as purveyor of surveillance technology for genocide (Black 2001), IBM created a product with promotional material asking a set of very simple questions: 'Have you ever been treated unfairly? How did it make you feel? Probably not too good.'

So opens the paper *Diversity in Faces*, describing the dataset of the same name.

Accompanied by a 23-second video of cartoon faces with a variety of skin colors, Diversity in Faces (DiF) was created in 2019 as a response to the perceived limitations in facial recognition datasets, directly referencing Gebru & Buolamwini's work as an inspiration. IBM's research team articulates the problem of bias and discrimination in facial recognition, and argues that 'the heart of the problem is not with the AI technology itself, per se … .for

[FRT] to perform as desired … training data must be diverse and offer a breadth of coverage' (Smith 2019). The 'diversity' they desire shifts the 'heart' of the problem from the algorithms to issues with the content of the dataset. In their way, they echo critical data studies scholarship that data are not pre-factual or pre-existing, but are 'situated, contingent, relational, and framed, and used contextually to try and achieve certain aims and goals' (Kitchin and Lauriault 2018, p. 4). To achieve their diversity goal and solve the problem, the researchers released DiF, containing one million images labelled with a range of data, from skin tone to craniofacial dimensions.[14] Tailored to provide a 'diverse' range of faces and labels, their 'initial analysis [shows] that the DiF dataset provides a more balanced distribution and broader coverage compared to previous datasets.' They end by noting that 'IBM is proud to make this available and our goal is to help further our collective research and contribute to creating AI systems that are more fair' (Smith 2019).

At first glance, the DiF dataset appears very different from the other corpora we have described; produced by a private entity, tested to ensure a wide distribution of faces, DiF was motivated by fairness—by *diversity*. But looking deeper, it becomes apparent that far from being distinct from prior datasets, DiF embodies and transforms the practices we have been tracing throughout the history of facial recognition. There are two primary resonances we point out—the sourcing of images and consent for inclusion; and the process of image tagging.

First, the sourcing of the data: where did the photographs come from, and how were they captured? In prior databases we have seen a range of practices from direct photography to the secondary consumption of policing data—but with DiF, that consumption was in some ways *tertiary*. Photographs were neither taken directly nor from an existing system: instead, researchers reprocessed the contents of another, pre-existing dataset named YFCC100M. Produced by Yahoo researchers, YFCC100M consists of openly-licensed photographs from the sharing site 'Flickr,' and was released online (and widely reused) for general use by computer vision researchers (Thomee et al. 2016). Because of the open licensing (and very general nature of what YFCC100M was to be used for), Yahoo's researchers saw no need to ask subjects for consent prior to including their photographs: the choice to openly license a photo was seen as sufficient evidence of consent. We see a shifting notion of consent throughout the history of FRT, as datasets become larger—from a model of explicit consent (Bledsoe and other early teams), to a model of capturing those who deviate from social norms (arrestees), to capturing those who simply consented to have photos they took reproduced.

IBM's researchers took a different tack. Their process *did* include a mechanism through which photographs could be withdrawn—but this was very

different from a consent process. To begin with, the right of withdrawal was *post hoc*: it was only after the public release of the dataset that photographs could be removed, since prior to that there had been no effort to ensure photographs' subjects or authors were aware of their presence in the dataset. This right was also not about the subjects at all; rather, it was extended to the *authors* of the photographs, who could write to IBM and request the removal of photographs of theirs within the dataset. Such an authorial, post-hoc redaction approach is common in the United States, and originates not in any concern for people, but, as D.E. Wittkower has demonstrated, for *property*: for the legal status of the photograph itself, and the intellectual property rights of its authors (Wittkower 2016).

Anyone attempting to do this faced an uphill battle; it was 'almost impossible … IBM requires photographers to email links to photos they want removed, but the company has not publicly shared the list of Flickr users and photos included in the dataset, so there is no easy way of finding out whose photos are included' (Solon 2019). If one did somehow manage to access the user and photo data, IBM promised only to 'consider [the] request … and remove the related [images] … as appropriate' (IBM 2019). Participation had thus been inverted, shifting from voluntary participation (in the case of Bledsoe) to involuntary and unknown participation (as in MEDS) through to participation that, while appearing voluntary and consensual, shifted the burden to the photographer and provided the *subject* with no rights or notification at all. Rather than being a matter of identity or self, the face becomes the property of the individual who *captures* it—further alienating the subject and dehumanising the photograph by ensuring that, even in the event a photograph is deanonymised, the subject has no right to speak for or on it. Additionally, the selection criteria for images to remain in the DiF dataset is for them to pass processing 'correctly': 'if there was any failure in the image processing steps, we excluded the face from further consideration' (Merler et al. 2019). This statement begs the question—whose images were excluded? While we do not consider enrollment in this dataset a privilege, it is noteworthy that a dataset created for diversity, a close kin to 'inclusion', immediately excluded 'failing' images.

The 'passing' images were subject to attribute generation—the labelling of faces' genders, ages, skin colours and craniofacial structures—which is the site of the *second* resonance between DiF and the earlier FRT datasets it aims to problematize. As we have discussed, an ongoing thread of work through the history of FRT has been to remove the human from the loop; to get as close to fully-automated image annotation as possible, enabling larger datasets with more mechanically objective annotations. With FERET, subjects' gender and race labels were provided through researcher assessment and interaction with the photographs subjects. Because of the structure of FERET data gathering, photographers could ask subjects about their demographic profile while images were being taken. With the MEDS dataset, law

enforcement assessment and interaction provided demographic labels. We are not suggesting that law enforcement is a credible source of information about a person, rather that the mechanism of data gathering was one predicated on human interaction, providing at least a theoretical opportunity for a subject to have input into their labels.

Interacting with subjects would have required identifying and contacting them (and presumably obtaining their explicit consent to be included in a facial recognition dataset); as a result, DiF researchers did not use any subject-provided information to tag images. Instead, they used pre-existing computer vision tools. For skin colour and craniofacial structure, they used automated, algorithmic tools— systems designed for color determination and structure selection. That they intended for skin color and facial structure features to be used to approximate race is not in doubt; their associated paper explicitly notes, in unpacking their choice of attributes, that 'skin color alone is not a strong predictor of race, and other features such as facial proportions are important' (Merler et al. 2019), and the researcher who led the team leaves open the possibility of '[returning] to some of these subjective categories' such as race (Coldewey 2019). The use of 'subjective' here is noteworthy. That the researchers identify race as a subjective category implies a contrast with their 'objective' anthropometric tools. While it is generally accepted that phrenology is a racist pseudoscience, we would do well to remember that anthropometry has its origins in racial classification systems. Pieter Camper's (1722–1789) 'facial angle' was 'useful as a means of distinguishing and ranking the races of man' and Edward Long's measurement-based classification system (1774) was used in the United States as a justification for the enslavement of kidnapped Africans (Bates 1995, p. 4). Our history shows us that systems used to measure human bodies are frequently leveraged as tools of colonialism and criminalization and must never be considered *prima facie* 'neutral' or 'objective'. (Cole 2009)

Conclusion

Through analysing four different FRT datasets—their contents, their use, and the practices and motivations surrounding their creation—we have sought to demonstrate (in a small way) how deeply interwoven FRT is with practices of racialization and dehumanisation. Beginning with the work of W.W. Bledsoe in inventing FRT, we have tracked—through datasets produced by defence research programs, the (formally) independent U.S. government technical standards agency, and a private company seeking to *explicitly address* dataset bias—the ways in which racial exclusion has consistently haunted facial recognition research. Further, we have articulated how efforts to achieve larger, more representative datasets have depended on increasingly dehumanising approaches to dataset subjects and their enrollment.

Facial recognition is inherently built around a degree of dehumanisation. Its monodirectionality 'leads to a qualitatively different way of seeing … .[the subject is] not even fully human. Inherent in the one way gaze is a kind of dehumanization of the observed' (Brighenti 2007, p. 337). This dehumanisation inherently raises questions of race and racialization, since to be 'less-than-human' is often racially-contingent.[15] As scholars such as Ellen Samuels have shown, many predecessors to current biometric technologies originated from fears that Black people would, post-Civil War, be able to evade scrutiny; to pass; to appear as (white) people (Samuels 2014, pp. 27–49). Similarly, the reduction of people to parts (and its frequently-racialised nature) can be seen throughout technoscientific history, from J. Marion Sims' gynecological experiments on Anarcha, Betsey, Lucy or other (unnamed) enslaved people, to the reuse and reproduction of Henrietta Lacks' cellular line (Wald 2012, Snorton 2017). We find, through this examination of FRT datasets, a similar lineage of reduction. As the technology matured into its role as a tool of the state, researchers became increasingly removed from the dataset subjects; compare the Bledsoe team's high level of engagement to the MEDS team engaging with the images only to remove subject names.

As discussed, a large area of focus for US-based FRT critics concerned with how it is implemented centre on *datasets*—specifically, how representative the datasets used to train the software are of the U.S. population. Critics contend that datasets used for FRT systemically underrepresent people of colour, particularly African-American people, and urge the creation of more 'diverse' and 'balanced' datasets as a way to prevent harmful and/or discriminatory consequences; we can see this as a driver for IBM's Diversity in Faces dataset. This criticism is often framed as an issue of 'biased datasets'—but of course, the datasets themselves are not so much *biased* as they are *reflective of their sites of use*. These histories do not show a straightforward underrepresentation of people of colour: rather, how race appears in these datasets is often contingent on their purpose. Datasets produced by and for the carceral state, such as MEDS-I and MEDS-II overrepresent people of colour compared to the U.S. population, but represent them accurately when looking at incarcerated people. Datasets produced for surveillance capitalism, marketing and neoliberal logics of extraction underrepresent people of colour—but may represent them perfectly proportionately in terms of their *purchasing power*. There is no simple story of (mis/under)representation leading to bias: it is the logics and systems of inequality that lead to the datasets' purposes, and so naturalize the datasets' demographic skews. Recognising this should make us highly sceptical about efforts to 'improve' FRT by 'debiasing' datasets. If the datasets that underrepresent people of colour are being used to train models for policing, border control and other forms of state control and violence, an improved model will only provide greater

accuracy for law enforcement agencies. If they are 'only' being used to train systems for surveillance capitalism, then efforts to increase representation are merely efforts to increase the ability of commercial entities to exploit, track and control people of colour (Spence 2015). Notions of 'visibility' and the ethical consequences of them are highly contingent and not universally received (Agostinho 2018). More generally, this article demonstrates the value in examining the ways in which infrastructures and their components are designed to understand the cultural values and politics that circulate within and between them.

Notes

1. Documents from our corpus will be referenced using the abbreviation 'DFR' followed by their index number, rather than author name and year, which are not always easily available. Note to reviewers: the full corpus will be available online at the culmination of our research project.
2. For critical perspectives on AI more broadly, please see Alison Adams' *Gender and the Thinking Machine*.
3. The subject's metadata may come from a variety of sources that are *not* the subject including third-party assessments of the subject's identity.
4. For more work on data as politically charged and/or negotiated, see (Ribes 2017, Shilton 2018, Maienschein *et al.* 2019, Williams 2018)
5. For more on the connection between imagery and disciplinary control see Sekula (1986), Tagg (1993).
6. Mugshots may have been the first dataset of faces. See Finn (2009) for an analysis of the mugshot. As police departments discovered, mugshot books did not scale—they became less useful the more faces they contained.
7. The slow process of replacing Shirley cards did not start (until 1995). (Roth 2009)
8. The history of photography is also intertwined with anthropometry and phrenology, both of which were/are mobilized for social control (Cole 2009).
9. See (Turk and Pentland 1991) for a review of early automation attempts between Bledsoe and eigenfaces.
10. For information on the disproportionate harms maintained by the US carceral system, see Kristian Williams' *Our Enemies in Blue: Police and Power in America* and Alex S. Vitale's *The End of Policing*. For more information on technology in the carceral system see *Captivating Technology: Race, Carceral Technoscience, and Liberatory Imagination in Everyday Life* edited by Ruha Benjamin
11. A 'false positive' is when the system incorrectly identifies two photographs as containing the same subject
12. We are unwilling to reproduce the nonconsensual sharing of people's images. Interested readers can learn more and download the datasets at https://www.nist.gov/itl/iad/image-group/special-database-32-multipleencounter-dataset-meds
13. The documentation for MEDS and other FRT datasets is not a neutral reporting of the contours of a technical object, but is itself a technical object produced through a variety of sociocultural interactions. We hope other researchers will analyze these documentation technologies.
14. IBM does not offer their definition of diversity. We assume, based on the imagery and language, that they are referencing racial, gender, and age diversity.

15. For information on the history of racialization as a concept, please see (Murji and Solomos 2005)

Acknowledgements

We are tremendously grateful to advisors, reviewers and friends, past and present, including Jacqueline Wernimont, David Ribes, Adam Hyland, Kate Crawford, Danya Glabau, Anna Lauren Hoffman—and each other. Our thanks further go to the editors for their precise and painstaking work in putting together this special edition.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was partially supported by an Ada Lovelace Fellowship, funded by Microsoft Research.

Notes on contributors

Os Keyes is a researcher and writer at the University of Washington, where they study gender, technology and (counter)power. They are a frequently-published essayist on data, gender and infrastructures of control, and a winner of the inaugural Ada Lovelace Fellowship.

Nikki Stevens is a PhD candidate at Arizona State University and a research associate at Dartmouth College. Their background as a software engineer informs their research on proxy surveillance, corporate data collection and the affective power of data.

ORCID

Nikki Stevens http://orcid.org/0000-0003-3811-9245
Os Keyes http://orcid.org/0000-0001-5196-609X

References

Abate, A.F., et al., 2007. 2d and 3d face recognition: A survey. *Pattern recognition letters*, 28 (14), 1885–1906.

Agostinho, D., 2018. Chroma key dreams: algorithmic visibility, fleshy images and scenes of recognition. *Philosophy of photography*, 9 (2), 131–155.

Andersson, R., 2016. Hardwiring the frontier? the politics of security technology in Europe's 'fight against illegal migration'. *Security dialogue*, 47 (1), 22–39.

Bacchini, F., and Lorusso, L., 2019. Race, again. how face recognition technology reinforces racial discrimination. *Journal of information, communication and ethics in society*, 17 (3), 321–335.

Ballantyne, M., Boyer, R.S., and Hines, L., 1996. Woody bledsoe: His life and legacy. *Ai magazine*, 17 (1), 7–20.

Bates, C. 1995. Race, caste and tribe in central india: The early origins of indian anthropometry.
Beauchamp, T., 2019. *Going stealth*. Durham, NC: Duke University Press.
Benjamin, R., 2016. Catching our breath: critical race sts and the carceral imagination. *Engaging science, technology, and society*, 2, 145–156.
Black, E., 2001. *IBM and the holocaust: The strategic alliance between Nazi Germany and America's most powerful corporation*. New York: Random House Inc.
Bowker, G.C., and Star, S.L., 2000. *Sorting things out: classification and its consequences*. Cambridge, MA: MIT Press.
Brighenti, A., 2007. Visibility: A category for the social sciences. *Current sociology*, 55 (3), 323–342.
Browne, S., 2015. *Dark matters: On the surveillance of blackness*. Durham, NC: Duke University Press.
Bud, A., 2018. Facing the future: The impact of apple faceid. *Biometric technology today*, 2018 (1), 5–7.
Buolamwini, J., and Gebru, T., 2018. Gender shades: intersectional accuracy disparities in commercial gender classification. *Conference on fairness, accountability and transparency*, 81, 1–15.
Castillo, R.V.C., et al. 2018. Class attendance generation through multiple facial detection and recognition using artificial neural network. *In*: *Proceedings of the 2018 International Conference on Machine Learning and Machine Intelligence*. ACM, 38–42.
Champion, B., 1998. Subreptitious aircraft in transnational covert operations. *International journal of intelligence and counter intelligence*, 11 (4), 453–478.
Coldewey, D. 2019. Ibm builds a more diverse million-face data set to help reduce bias in ai. Available from: https://techcrunch.com/2019/01/29/ibm-builds-a-more-diverse-millionface-dataset-to-help-reduce-bias-in-ai/.
Cole, S.A., 2009. *Suspect identities: A history of fingerprinting and criminal identification*. Cambridge, MA: Harvard University Press.
Cook, C.M., et al., 2019. Demographic effects in facial recognition and their dependence on image acquisition: An evaluation of eleven commercial systems. *Ieee transactions on biometrics, behavior, and identity science*, 1 (1), 32–41.
Finn, J.M., 2009. *Capturing the criminal image: from mug shot to surveillance society*. Minneapolis, MN: University of Minnesota Press.
Gragg, S., 1997. Facial recognition at the cia. *In*: W. Ishimoto, ed. *Terrorism and counter-Terrorism methods and technologies*. Boston, MA: International Society for Optics and Photonics, 2933, 43–47.
Guo, Y., et al., 2016. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. *In*: B. Leibe, J. Matas, N. Sebe, and M. Welling, eds. *European Conference on computer vision*. Amsterdam, NL: Springer, 87–102.
Hall, R., 2007. Of ziploc bags and black holes: The aesthetics of transparency in the war on terror. *The communication review*, 10 (4), 319–346.
IBM. 2019. Diversity in faces privacy agreement. Available from: https://web.archive.org/web/20190709194832/https://www.research.ibm.com/artificialintelligence/trusted-ai/diversity-in-faces/documents/privacy-notice.html.
Introna, L., and Nissenbaum, H. 2010. Facial recognition technology a survey of policy and implementation issues.
Introna, L.D., and Wood, D., 2004. Picturing algorithmic surveillance. *Surveillance and society*, 2 (2-3), 177–198.
Jones, S.R., and McEwen, M.K., 2000. A conceptual model of multiple dimensions of identity. *Journal of college student development*, 41 (4), 405–414.

Kaufman, G.J., 1974. *Computer recognition of facial profiles*. Columbus, OH: Ohio State University.
Kitchin, R., and Lauriault, T., 2018. Toward critical data studies: charting and unpacking data assemblages and their work. *In*: J. Thatcher, A. Shears, and J. Eckert, eds. *Thinking big data in geography: New regimes, new research*. Lincoln, NE: University of Nebraska Press, 3–20.
Lee-Morrison, L., 2018. A portrait of facial recognition: tracing a history of a statistical way of seeing. *Philosophy of photography*, 9 (2), 107–130.
Leong, B., 2019. Facial recognition and the future of privacy: I always feel like … somebody's watching me. *Bulletin of the atomic scientists*, 75 (3), 109–115.
Li, S.Z., and Jain, A., 2015. *Encyclopedia of biometrics*. New York: Springer Publishing Company, Incorporated.
Maienschein, J., et al., 2019. Data management and data sharing in science and technology studies. *Science, technology, & human values*, 44 (1), 143–160.
Masi, I., et al. 2018. Deep face recognition: A survey. *In*: 2018 *31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, Oct. 471–478.
M'charek, A., 2020. Tentacular faces: race and the return of the phenotype in forensic identification). *American anthropologist*, 122 (2), 369–380.
Merler, M., et al., 2019. Diversity in faces. *Arxiv preprint arxiv:1901.10436*.
Moon, H., and Phillips, P.J., 2001. Computational and performance aspects of pca-based facerecognition algorithms. *Perception*, 30 (3), 303–321.
Murji, K., and Solomos, J., 2005. *Racialization: studies in theory and practice*. Oxford: Oxford University Press on Demand.
M'charek, A., 2014. Race, time and folded objects: the hela error. *Theory, culture & society*, 31 (6), 29–56.
Phillips, P.J., et al., 1997. The FERET September 1996 database and evaluation procedure. *In*: J. Bigün, G. Chollet, and G. Borgefors, eds. *International Conference on audio-and video-based biometric person authentication*. Crans-Montana: Springer, 395–402.
Phillips, P.J., et al., 2000. The feret evaluation methodology for face-recognition algorithms. *Ieee transactions on pattern analysis and machine intelligence*, 22 (10), 1090–1104.
Ribes, D. 2017. Notes on the concept of data interoperability: Cases from an ecology of aids research infrastructures. *In*: *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 1514–1526.
Roth, L., 2009. Looking at shirley, the ultimate norm: colour balance, image technologies, and cognitive equity. *Canadian journal of communication*, 34 (1), 111–136. Available from: https://www.cjconline.ca/index.php/journal/article/view/2196 [Accessed 2019-09-06].
Salari, S.R., and Rostami, H., 2016. Pgu-face: A dataset of partially covered facial images. *Data in brief*, 9, 288–291.
Samuels, E., 2014. *Fantasies of identification: disability, gender, race*. New York: NYU Press.
Sekula, A., 1986. The body and the archive. *October*, 39, 3–64.
Sellar, S., 2015. Data infrastructure: a review of expanding accountability systems and largescale assessments in education. *Discourse: studies in the cultural politics of education*, 36 (5), 765–777.
Shilton, K., 2018. Engaging values despite neutrality: challenges and approaches to values reflection during the design of internet infrastructure. *Science, technology, & human values*, 43 (2), 247–269.

Sirovich, L., and Kirby, M., 1987. Low-Dimensional procedure for the characterization of human faces. *Journal of the optical society of america a*, 4 (3), 519–524.
Smith, F., 2016. Law and order: the twin impact of mobile and biometrics. *Biometric technology today*, 2016 (9), 5–7.
Smith, J.R. 2019. Ibm research releases 'diversity in faces' dataset to advance study of fairness in facial recognition systems. Available from: https://www.ibm.com/blogs/research/2019/01/diversity-in-faces/.
Snorton, C.R., 2017. *Black on both sides: A racial history of trans identity*. Minneapolis, MN: University of Minnesota Press.
Snow, J. 2018. Amazon's face recognition falsely matched 28 members of congress with mugshots. *Aclu*. Available from: https://www.aclu.org/blog/privacytechnology/surveillance-technologies/amazons-face-recognition-falsely-matched-28.
Solon, O. 2019. Facial recognition's 'dirty little secret': Millions of online photos scraped without consent. Available from: https://www.nbcnews.com/tech/internet/facial-recognition-sdirty-little-secret-millions-online-photos-scraped-n981921.
Spence, L.K., 2015. *Knocking the hustle*. Santa Barbara, CA: punctum books.
Star, S.L., 1999. The ethnography of infrastructure. *American behavioral scientist*, 43 (3), 377–391.
Star, S.L., and Ruhleder, K., 1996. Steps toward an ecology of infrastructure: design and access for large information spaces. *Information systems research*, 7 (1), 111–134.
Stark, L., 2018. Facial recognition, emotion and race in animated social media. *First Monday*, 23, 9.
Tagg, J., 1993. *The burden of representation: essays on photographies and histories. vol. 80*. Minneapolis, MN: University of Minnesota Press.
Thomee, B., et al., 2016. Yfcc100m: The new data in multimedia research. *Communications of the acm*, 59 (2), 64–73.
Turk, M., and Pentland, A., 1991. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3 (1), 71–86.
Wald, P., 2012. Cells, genes, and stories. *In*: K. Wailoo, A. Nelson, and C. Lee, eds. *Genetics and the unsettled past: The collision of dna, race, and history*. Rutgers: Rutgers University Press, 247–265.
Wayman, J.L., 2007. The scientific development of biometrics over the last 40 years. *In*: K.M.M. de Leeuw and J. Bergstra, eds. *The history of information security*. Amsterdam, NL: Elsevier, 263–274.
Williams, R., 2018. Bloody infrastructures!: exploring challenges in cord blood collection maintenance. *Technology analysis & strategic management*, 30 (4), 473–483.
Wittkower, D.E., 2016. Lurkers, creepers, and virtuous interactivity: from property rights to consent to care as a conceptual basis for privacy concerns and information ethics. *First Monday*, 21 (10).
Wood, S., 2016. Police body cameras: emotional mediation and the economies of visuality. *In*: S.Y. Tettegah, and S.U. Noble, ed. *Emotions, technology, and design*. Emotions and Technology. Boston, MA: Academic Press, 227–239.