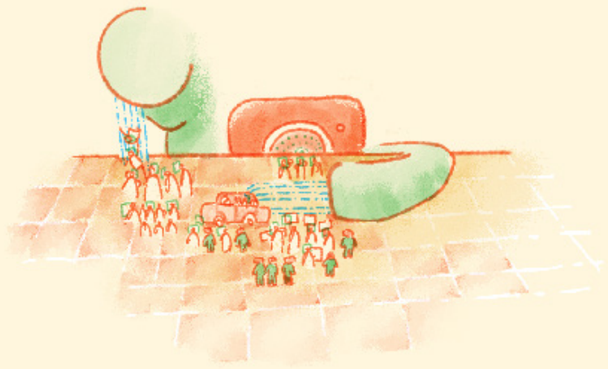


II. Unpacking Algorithmic Harms

AI Observatory

This project was undertaken as a part of Divij Joshi's project as a Mozilla Fellow, and is supported by Mozilla.

Designed and typset by Design Beku

Transparency and Accountability in ADMS

Transparency

Agorithmic systems are frequently referred to as 'black boxes' – as instruments into which inputs and outputs are visible, but the precise mechanism of its function is inscrutable. Such 'black box' Automated Decision-Making Systems pose structural challenges to democratic ideals of transparency, accountability and participation, and consequently, to public trust in the operation of these systems. It is important to interrogate how these constraints arise and what their consequences are.

Transparency is an essential element of a democratic society. Without adequate information about decisions that affect their lives, people cannot comprehend how these decisions are made, how they may affect them, or how they can participate in, and if possible, change such decisions. Similarly, it is impossible to hold the use of ADMS accountable to any legal standard or guiding principles if the mechanisms by which it functions are unknown. Many ADMS pose challenges to these ideals, both by being technically opaque, as well as due to the institutional opacity which often surrounds these systems.

The technical opacity of ADMS is a well-known phenomenon. As explained previously, computer algorithms within ADMS are essentially 'models' or abstractions of the decision-making metrics which are employed by human beings. However, in the process of creating these algorithmic models, many aspects of the decision-making logic may be altered.

One area where technical opacity arises is from the inability to be able to read or parse the computer programme and understand how this decision-making logic has been altered in its translation to software or code.[62] Given that not everyone is able to understand how the decision-making logic is reflected in source code, even where the source code of a computer programme and algorithm is made available, the logic always be obvious or accessible to users or affected persons.[63]

With the increasing scale and complexity of algorithmic tools, which rely on processing hundreds or thousands of inputs through myriad statistical or mathematical processes, the logic of decision-making employed within algorithms becomes even more difficult to interpret, even to the creators of these systems. This may particularly be true for contemporary machine learning practices like image recognition through neural networks, where the algorithmic system can produce accurate outputs, but using metrics which have been 'learned' by the system, that are so abstracted from the initial logic of the designers of the algorithm, as to be technically un-interpretable by humans.[64]

Another source of the opacity of ADMS is institutional – caused by the legal and institutional mechanisms which govern the operation of these systems. Many of the ADMS documented here include computational algorithms developed by and procured from private firms, which have deployed resources into the creation of these systems, and have an economic incentive to retain proprietary ownership and rights over these systems. As such, many algorithmic systems are considered the intellectual property of these private firms, which are protected as trade secrets of copyrights of these firms. These laws prevent the databases, algorithms and other associated components of ADMS from being accessed or scrutinised by the public, and often even by government agencies procuring such systems.

> ***With governments increasingly outsourcing important infrastructure, including ADMS infrastructure, to private firms, there is correspondingly a shift in the norms of transparency – from public and transparent by default, as recognised under a right to information, to private and protected by default, protected both by laws like trade secrets, as well as by the forms and governance practices of private companies.[65]***

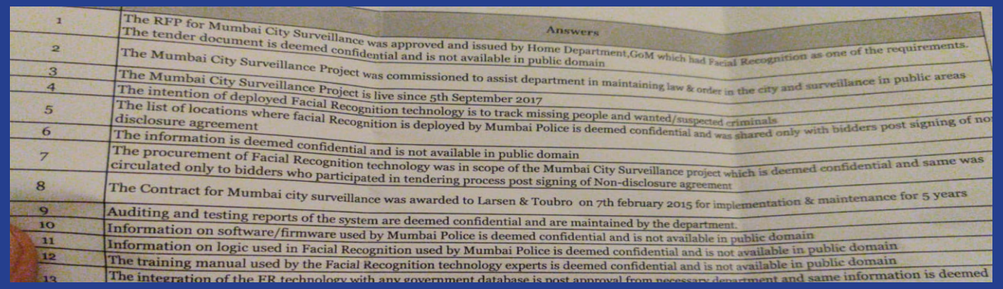
62 Citron DK, 'Technological Due Process' 85 Wash U L. Rev., 1249 (2008)

63 Ananny M and Crawford K, 'Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability' (2018) 20 New Media & Society 973

64 Burrell J, 'How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms' (2016) 3 Big Data & Society

65 Pasquale F., Black Box Society, (Harvard University Press, 2016); Brauneis R., & Goodman E., Algorithmic Transparency for the Smart City, 20 Yale J.L. & Tec.

[Case Study: Algorithmic Systems and the Dilution of the Right to Information]



	Answers
1	The RFP for Mumbai City Surveillance was approved and issued by Home Department,GoM which had Facial Recognition as one of the requirements.
2	The tender document is deemed confidential and is not available in public domain
3	The Mumbai City Surveillance Project was commissioned to assist department in maintaining law & order in the city and surveillance in public areas
4	The intention of deployed Facial Recognition technology is to track missing people and wanted/suspected criminals
5	The list of locations where facial Recognition is deployed by Mumbai Police is deemed confidential and was shared only with bidders post signing of no disclosure agreement
6	The information is deemed confidential and is not available in public domain
7	The procurement of Facial Recognition technology was in scope of the Mumbai City Surveillance project which is deemed confidential and same was circulated only to bidders who participated in tendering process post signing of Non-disclosure agreement
8	The Contract for Mumbai city surveillance was awarded to Larsen & Toubro on 7th february 2015 for implementation & maintenance for 5 years
9	Auditing and testing reports of the system are deemed confidential and are maintained by the department.
10	Information on software/firmware used by Mumbai Police is deemed confidential and is not available in public domain
11	Information on logic used in Facial Recognition used by Mumbai Police is deemed confidential and is not available in public domain
12	The training manual used by the Facial Recognition technology experts is deemed confidential and is not available in public domain
13	The integration of the FR technology with any government database is post approval from necessary department and same information is deemed

[Image: Excerpt from an RTI response by the Mumbai Police, on the Mumbai City Surveillance Project]

The Right to Information Act in India was a landmark moment which made concrete the constitutional right to information, recognising the right of a democratic public to demand information from the government, and the responsibility of the public agencies to proactively provide information, including information about how policy decisions are taken; as well as providing information to individuals about decisions taken about them.

Algorithmic systems have denuded the democratic safeguards that laws like the RTI Act provide. While the definition of 'information' under the RTI Act is broad enough to cover source code, or algorithmic models, this information is often not provided by claiming exemptions which exist under the RTI Act, such as Section 8(1)(d), which exempts the disclosure of confidential information and intellectual property in some circumstances; or Section 8(1)(a), which exempts the disclosure of information on grounds of the sovereignty and security of the nation. Further, disclosure of components of an ADMS, like the databases on which an algorithmic system operates, may pose risks to other interests like privacy and personal data protection, which are also recognised in the RTI Act under Section 8(1)(j).

The proliferation of ADMS is systematically incapacitating important democratic norms and values, of which the Right to Information appears to be one casualty. There is an urgent need to reform and bring laws and methods of governmental transparency in line with the use of Automated Decision-Making within government.

Transparency is not an end in itself – it is a necessary but not a sufficient means of ensuring democratic and equitable ADMS use. In particular, it is necessary to understand what forms of transparency can lead to better democratic participation and outcomes, and how transparency can be balanced against other legitimate considerations like privacy, or the need to prevent unintended uses of algorithmic systems.[66]

66 Ananny M and Crawford K, 'Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability' (2018) 20 New Media & Society 973

Accountability

An important and related concern to the transparency of ADMS in India is the lack of clear systems of accountability for the failures or harms caused by ADMS. Accountability for ADMS requires attributing responsibility to an actor, and providing recourse to people affected by a particular outcome of an ADMS.

Algorithmic systems often shift or distribute agency for decisions in ways which have not been previously encountered, and obscures clear lines of responsibility for the outcomes of the system. For example, an algorithm may be designed by one actor, operating on data provided by another, and finally, the ADMS may be used or deployed by a third actor, each of whom would have limited knowledge about the others. Often, this could result in attributing agency and culpability for an outcome of the system to a human actor or institution who did not have control or agency over the decision, or even leading to circumstances where the absence of responsibility implies no accountability or redress for affected persons.[67]

Another reason that ADMS obscures clear accountability is due to the increasing autonomous decision-making capabilities of computer systems – there are many instances where an algorithmic system malfunctions or performs in a manner which could not be reasonably foreseen, making attribution of responsibility difficult.

This is particularly true for contemporary machine learning systems which may make non-intuitive decisions, the logic of which would be difficult to comprehend (as explained in the section above).

Accounting for accountability requires providing clear lines of responsibility for the harms or failures caused by ADMS, including establishing clear liability for the damage caused by ADMS, and clear channels of redress to affected persons.

67 Elish, M.C., 'Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction', Engaging Science,Technology, and Society 5 (2019); 'Responsibility and AI', Council of Europe DG (2019) 05 <https://rm.coe.int/responsability-and-ai-en/168097d9c5>

[Case Study: Aadhaar and Unaccountable Authentication Failures]

The Government of India's Aadhaar Unique Identification system consists of various assemblages of algorithms, including algorithmic systems which identify individuals, known as 'authentication'. Biometric authentication has a documented high failure rate[68] and the failure of identification, or incorrect identification can have devastating consequences for an individual – including the denial of essential public services like ration. Biometric systems are essentially probabilistic, which means that there is always a chance of an error in the matching – and that a mathematical algorithmic threshold must be established for what percentage error is acceptable and what is not. Given this error-prone nature of biometric algorithms, and the consequences of its failure, clear accountability becomes particularly essential.[69]

However, this measure of accountability has not been forthcoming from the Government of India or the agency responsible for Aadhaar – the Unique Identification Authority of India (UIDAI). The legislation governing Aadhaar does not specifically detail who is responsible to resolve failures of the biometric matching algorithm, or how such resolution should occur. While some circulars issued by the UIDAI specify that government agencies should incorporate 'exception handling measures' in case of biometric failures, it has been observed that these measures are often not made available.[70] Moreover, alternatives apart, no clear system of accountability has been framed for the failure of the biometric algorithms.[71]

Aadhaar still remains an outlier, in that its use is governed by a specific legal and regulatory regime, and some measures of accountability like grievance redressal mechanisms exist. Even so, it remains a cautionary example of the consequences of failing to consider and account fault lines within structures of accountability that algorithmic systems and ADMS pose.

ADMS + Surveillance and Profiling

Automated Decision-Making Systems are becoming integral to the proliferation of surveillance enabled through information technology – a method of observing, knowing and governing individuals and populations through information collected about them. Many of the ADMS technologies documented in this project are explicitly used for surveillance – to govern and police populations to reduce risk and undesired behaviour. This section examines how ADMS in India is used to further surveillance and profiling by government agencies.

ADMS requires the behaviours, traits and attributes of individuals and communities to be abstracted as data, in particular ways which are suitable for algorithmic computation. Contemporary ADMS like machine learning technologies require increasingly vast amounts of information in order to draw usable relationships between them through computation. The necessity for huge volumes of data is routinely invoked as a critical first step for the development of Artificial Intelligence and related technologies.[72] ADMS and 'AI' are therefore providing the imperative for amassing vast amounts of data in ways which allow for their algorithmic computation (discussed in the section on 'data and databases in ADMS').

ADMS themselves are utilised for the explicit purpose of surveillance within policing and other public systems – to profile individuals and groups, or to identify and predict people's movements and behaviours and classify them based on their perceived risk and undesirability. As technologies like social media or CCTV cameras allow for the constant production of data about individuals and populations, algorithmic systems are deployed to automatically or programmatically sort and classify such information, at massive scales. Therefore, the act of surveillance itself – what kind of information to acquire, and what kind of behaviour it reveals – becomes automated, continuous and cumulative – bounded by the logic of the particular algorithmic system.[73]

Automated surveillance technologies, like social media surveillance systems and automated facial recognition systems across cities, which are able to collect and process vast amounts of data to draw meaningful inferences about the people being surveilled, are shifting us into a new paradigm of mass surveillance and continuous policing and discipline. In India, departments ranging from state police forces to the Income Tax Department have procured and utilised systems for such unconstrained surveillance, including analysis of interconnected government databases as well as social media surveillance.

72 Committee of Experts on Non Personal Data Governance Framework, supra

73 Andrejevic M and Gates K, 'Big Data Surveillance: Introduction' (2014) 12 Surveillance & Society 185

[Case Study: Automated Facial Recognition Systems for Police Surveillance]

One of the primary modalities of automated surveillance in India today is through the use of Facial Recognition Technologies (FRT). FRT uses various algorithmic techniques to extract features of individual faces captured through photographs or video feeds, and compares them with an existing database of faces, in order to identify whether there is a match between the two.

The use of FRT for policing has grown exponentially in India, in line with the increasing use of video surveillance devices like CCTV cameras, particularly within urban centres and 'smart cities'. Police agencies in India have claimed to employ FRT for purposes ranging from tracking missing children, to identifying protestors and 'rowdy elements'.

Automated Facial Recognition is one of many technologies integrated into ADMS use in policing, which allows for hidden, ubiquitous surveillance. At the time of writing, this toolkit has documented more than 20 implemented or proposed uses of FRT since 2015 alone, each existing without a clear legal basis and without appropriate mechanisms for regulation or oversight. Since 2019, the Government of India has also attempted to create a 'national' Automated Facial Recognition System which will attempt to be a centralised FRT system, built upon CCTNS infrastructure, for all state and central policing and intelligence forces to utilise.

Although the utility and accuracy of these systems is often circumspect, FRT technologies substantially expand the surveillance capabilities of the state. Without any form of regulation or oversight, they are creating the very real possibility of continuous and ubiquitous mass surveillance with very little justification.

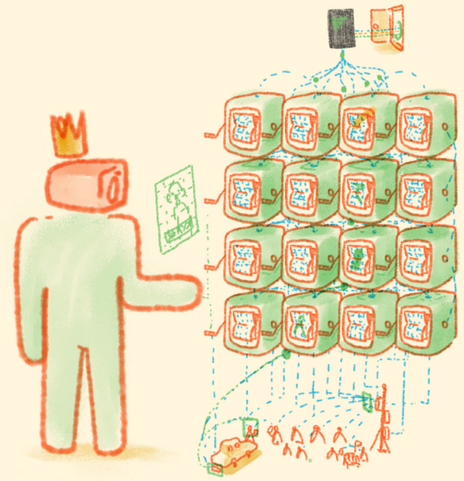
Read more **here** about how Facial Recognition Technologies are intensifying mass surveillance in India.

Another way in which ADMS implicate privacy is through the inference and production of information about individuals or groups in ways which were not consensually disclosed, known as automated profiling.

> ***Automated profiling often concerns seemingly innocuous information which is aggregated and algorithmically computed in ways which can reveal sensitive information, including preferences, attributes and behaviours about individuals and groups.***

The information generated through profiling may be inaccurate or incomplete, or, even when accurate, may reveal information contrary to an individual's agency and self-determination. Further, this inferred information is subsequently used to make decisions concerning individuals.[74]

Profiling through automated means are now a common feature of our online lives – programmatic and behavioral targeting of advertisements on social media, for example, attempts to identify attributes about consumer behaviour in order to send relevant advertisements.[75] Perhaps more concerning, some cases of profiling are explicitly used to aid in the political manipulation of the subjects of profiling, for example, the revelations made about the firm Cambridge Analytica attempting to manipulate voter behaviour on Facebook.[76]



The ADMS documented in this project exhibit forms of behavioural profiling at individual and group levels. For example, 'sentiment analysis' tools are being used by police departments to trawl through social media and understand responses of communities to different events, including political events like Supreme Court judgements on the Ayodhya land dispute, or the abrogation of Article 370 in Kashmir. Such systems attempt to identify the 'emotional' responses of individuals and communities, and provide this information to law enforcement and other government authorities.

74 'Data Is Power: Profiling and Automated Decision-Making in GDPR' (Privacy International) <http://privacyinternational.org/report/1718/data-power-profiling-and-automated-decision-making-gdpr>; Wachter, S. and Mittelstadt, B., 'A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI' (Oxford Law Faculty, 9 October 2018) <https://www.law.ox.ac.uk/business-law-blog/blog/2018/10/right-reasonable-inferences-re-thinking-data-protection-law-age-big>

75 Kingaby H, 'AI & Advertising, A Consumer Perspective' (Medium, 2 October 2020), <https://medium.com/swlh/ai-advertising-a-consumer-perspective-f8cd0fb6893>

76 Hu M, 'Cambridge Analytica's Black Box' (2020) 7 Big Data & Society

[Case Study: Social Media Surveillance and Profiling in India's Policing Agencies]

The surveillance capabilities of government agencies and law enforcement have been transformed with the advent of programmatic and automated surveillance. The internet, and social media in particular, has emerged as a space for gaining all kinds of insights into the behaviours of groups and individuals, which are used to aid policing and law enforcement. Even though these systems rely largely on information which is 'open' and available online, (referred to by law enforcement as Open Source Intelligence or OSINT),[77] these systems jeopardise expectations of privacy because information is used in contexts for which it was not intended.[78]

Social media surveillance systems are widely used in both law enforcement as well as in other government departments. Two major forms of social media surveillance documented in this project including sentiment analysis tools, and social network analysis tools. Sentiment Analysis tools like AASMA scan social media posts for particular words or phrases, and alert agencies when it finds these terms are found, and can also be used to flag what kind of behaviour such speech indicates ('violent', 'anti-national', 'anti-social'). Social Network Analysis Tools like 'X1 Social Discovery'[79] allows law enforcement to automate the process of surveilling particular individuals, and generating profiles of their social networks or 'associates'.

Algorithmic surveillance flips presumptions of innocence, and requirements for specificity or justification in targeting individuals or groups for surveillance, by roping every individual and every action into a matrix of probability which is used to decide undesirable or unlawful behaviour. Ultimately, it is transforming the relationship between the state and the citizen – where all citizens are presumed to be located on this scale of probable guilt, and subject to invasive surveillance.[80]

77 Srivastava, KS, 'Social Media and Privacy: Government Use of Surveillance Tool Raises Concern over Data Protection' <https://scroll.in/article/893015/40-government-departments-are-using-a-social-media-surveillance-tool-and-little-is-known-of-it>

78 Nissenbaum, H., 'A Contextual Approach to Privacy Online', Dædalus Journal of the American Academy of Arts & Sciences, 2011 <https://www.amacad.org/publication/contextual-approach-privacy-online>

79 https://www.x1.com/products/x1-social-discovery/

80 Goldenfein, J., Monitoring Laws, (Cambridge University Press, 2020)

ADMS + Dispossession

ADMS use in India is facilitating the dispossession of people's rights and entitlements without providing adequate recourse to rectify unjust dispossession. This section explains why, and how, ADMS are being used to determine people's legal entitlements in India, and how it is facilitating the dispossession of their legitimate claims to state welfare and access to public goods.

Algorithmic systems are widely used by public agencies in India, in order to identify or screen individuals who may receive government welfare or access to rights; to determine their eligibility for particular entitlements; to sort and classify eligible beneficiaries, as well as to determine the circumstances which can render them ineligible for these entitlements. Each of these uses of ADMS has led to barriers in access to welfare or the exercise of individual rights, and the dispossession of peoples claims and entitlements to varying degrees.

Identification

Welfare distribution in India currently depends extensively on the identification of particular individuals and groups entitled to specific state benefits like ration, housing or credit. Algorithmic systems are being used to identify whether welfare beneficiaries and rights-holders are who they claim to be. There has been a systematic effort to digitise identification systems and utilise ADMS for the purpose of identification, such as in the move from paper documents for identification to the use of digital biometric systems in the case of Aadhaar.

As explained previously, the Government of India's Aadhaar system utilises biometric fingerprint recognition technologies to ensure that beneficiaries are correctly identified. This relies on the faulty assumption that biometric identification is accurate across populations, even as the Government has itself claimed that the biometric authentication mechanism fails at multiple levels. The Government of India is now attempting to utilise iris scanning and facial recognition technologies for biometric authentication within the Aadhaar system, each of which come with their own biases and points of failure.[81]

Eligibility, Screening and Prioritisation

Welfare systems across the world, including in India, depend extensively on filtering and classifying individuals in order to ensure that benefits are claimed according to particular socio-economic circumstances. For example, the eligibility of individuals to central and state government welfare schemes like PDS or is often determined against the enumeration of individuals and households done through the Socio-Economic and Caste Census (last conducted in 2011), or through household survey conducted at the state level.[82]

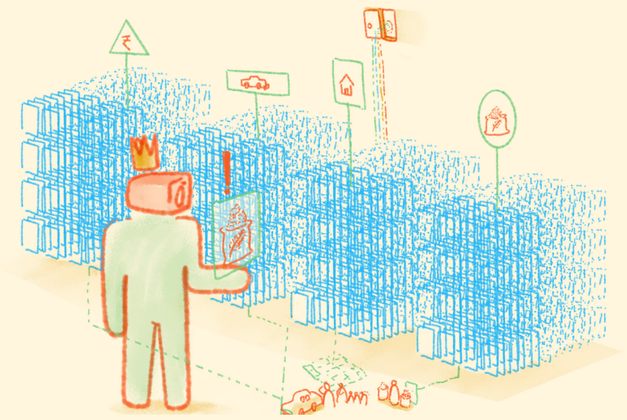
81 Manikandan, A., & Shukla, S., 'Facial recognition, iris scans may be used for welfare scheme payouts', (Aug 26, 2020) <https://economictimes.indiatimes.com/industry/banking/finance/banking/facial-recognition-iris-scans-may-be-used-for-welfare-scheme-payouts/articleshow/77755102.cms?from=mdr>

82 PTI, Government to use SECC data for effectiveness of welfare schemes, (Feb 6, 2017) <https://economictimes.indiatimes.com/news/politics-and-nation/government-to-use-secc-data-for-effectiveness-of-welfare-schemes/articleshow/57001274.cms?utm_source=contentofinterest&utm_medium=text&utm_campaign=cppst>

Increasingly, however, classifications created by algorithmic systems are being used as alternatives to, or in addition to, criteria which have historically determined eligibility for access to welfare or legal entitlements. This is done both to determine their entry into the welfare system (whether they are considered 'eligible beneficiaries'), as well as to classify and sort beneficiaries among themselves, according to some criterion for priority.

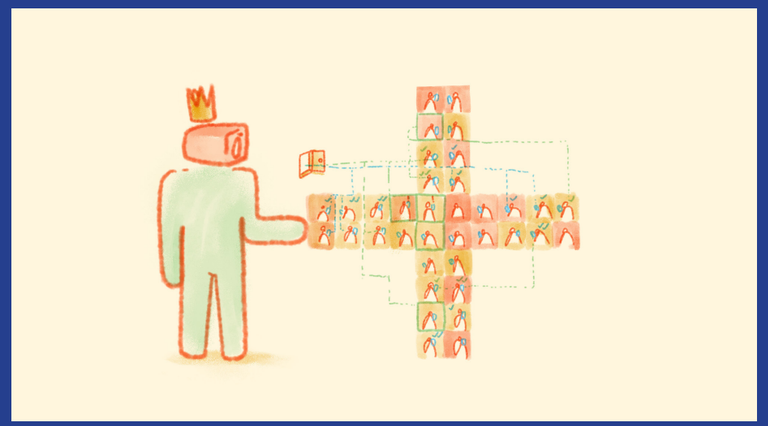
For example, the Government of Telangana's 'Samagra Vedika' scheme is being used to create additional classifications to determine whether an individual or a family qualifies for 'Below Poverty Line' schemes like the Government's subsidised housing scheme, by analysing data across a number of fields, including vehicle registration, electricity bills or property taxes. According to publicly available information, an algorithmic system is being utilised to assess whether an individual should be granted benefits based on an analysis of information located in various databases, without individuals being given the opportunity of a hearing or being informed of how such information is to be used and analysed by the system. The system classifies beneficiaries according to four classes – 'qualify', 'qualify with verification', 'low priority' and 'don't consider'. The benefits are then dispersed according to the classification generated, with different consequences for different classes of individuals (for example, 'don't consider' individuals will not be provided claims, while 'low priority' will only be considered after disbursal to the 'qualified' candidates).

The outputs of algorithmic systems are also being used to 'screen' individuals and determine their access to public spaces. For example, many airports in India have started using facial recognition screening to augment and replace human checks for boarding passes. Similarly, automated systems are determining access to public spaces by screening attendees – including political rallies. In New Delhi, the Delhi Police's Automated Facial Recognition System was used to screen persons who were identified as 'habitual protestors', and disallow them from attending a political rally organised by Prime Minister Narendra Modi.



[Case Study: Algorithmic Health Screening Through Aarogya Setu]

Automated screening tools have been widely utilised to prevent potential outbreaks or contagion during the CoVID-19 pandemic. One such tool has been Aarogya Setu, an ADMS deployed by the Government of India, which is expected to track a person's movement and their associations with other people, in order to model their risk of infection of COVID-19.[83]



Aarogya Setu embodied the failure of legal and government institutions to recognise and account for the failure of ADMS. While the tool was initially used only as a voluntary and private mechanism for individuals to rely upon, the purpose of the tool soon changed. Government notifications made the use of Aarogya Setu mandatory for accessing public spaces and services, including for employees to access their workspaces, as well as for the use of public transport – from flights to trains.

The tool quickly transformed into an automated screening system – the data collected from the system was used to classify individual users based on risk of infection, and to assign particular values (high risk, medium risk and low risk) to users. Access to public spaces was then mediated based on the values assigned to individual users – without individuals being made aware of the specific reason for the value, nor any mechanism to challenge or appeal the decision made by the tool.

The example of Aarogya Setu should caution us not only to the inherent limitations of certain technological systems, but also, importantly, of how reliance on automated digital technologies can exclude populations who do not have access to them and who cannot be 'seen' by the data alone.

83 Joshi, D., Mohan, S., 'A Legal Framework for Digital Surveillance in the COVID-19 Pandemic', Medianama, (July 14, 2020)

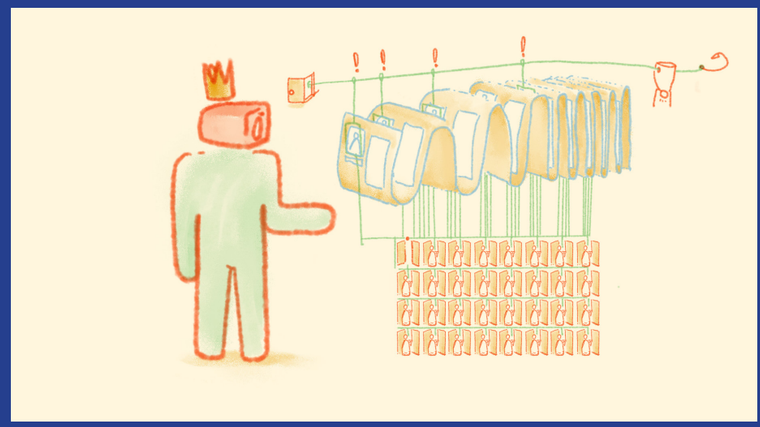
Fraud and Duplicate Detection

A related and common use of the system is to render ineligible and stop the benefits or entitlements of individuals who have been flagged by an automated system as either 'duplicates' or 'fraudulent'. As explained previously, de-duplication of databases in order to ensure the 'uniqueness' of beneficiaries has been a major fixation of contemporary ADMS use in public agencies. These de-duplication algorithms are being utilised across welfare schemes and are leading to the generation of 'clean' lists which often deprive and dispossess individuals of their legal claims, from welfare entitlements like ration or credit, to voting rights.

[Case Study: Disenfranchising Voters Through Algorithmic Purification]

The National Electoral Roll Purification and Authentication Programme, or NERPAP, is a project of the Election Commission of India intended to 'de-duplicate' the electoral rolls for registered voters, in an effort to curtail election fraud. Initially intended to be voluntary, the NERPAP programme has subsequently been expanded and 'piloted' in various jurisdictions without the consent of voters, leading to wide-scale deletion of eligible names from voter lists – and consequently, disenfranchising voters from exercising their right to vote. From Bihar to Telangana – the implementation of the NERPAP ADMS has resulted in the wide-scale disenfranchisement.

The NERPAP programme identified potential duplicate or 'fraudulent' voters through the process of matching names in voter databases (the Electoral Photo ID Card or EPIC database), to citizen databases which were linked with an Aadhaar ID. If the algorithmic system which matched the two databases saw 'duplicates', these names were automatically deleted from the list of voters. The curious process followed in the NERPAP programme assumed fraud if indicated by the name-deletion algorithm, and placed the burden of responding to the deletion on citizens, after the fact. While the algorithmic 'seeding' of Aadhaar with NERPAP was finally halted through the intervention of the Supreme Court of India,[84] the lasting damage to the exercise of a right to democratic participation has neither been acknowledged nor mitigated by the agencies responsible for the system. Rather, according to reports, the Government of India is intending to give legal sanctity to the system through amendments to the law.



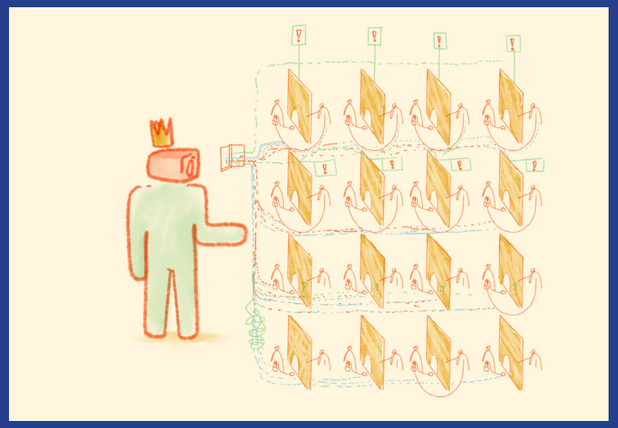
84 Election Commission of India, Circular, <https://rti.eci.nic.in/public/images/cpio_uploaded/3769/13.08.2015(NERPAP-ban%20on%20Aadhaar%20linking%20with%20ER%20database).pdf>

Apart from efforts at removing 'digital duplicates', progressively sophisticated 'automated fraud detection systems' abound in public agencies – from the Income Tax department to public health insurance through the Ayushman Bharat scheme. Increasingly, these agencies are adopting machine learning systems for identifying fraudulent behaviour, including systems analyse information across a range of databases, from government databases to social media.

[Case Study: Automated Fraud Detection in the Ayushman Bharat Universal Health Scheme]

Ayushman Bharat is the Government of India's ambitious universal healthcare scheme, which intends to provide health insurance coverage to 40% of the lowest-income population in India. Fraudulent claims towards health insurance have been described as a major obstacle in the implementation of the scheme. To overcome this obstacle, the National Health Authority, responsible for the administration of Ayushman Bharat, has implemented a 'Fraud Analytics Control and Tracking System', developed by the firm SAS.

According to the NHA, the FACTS system will use Artificial Intelligence and Machine Learning in order to "identify suspect transactions & entities. Using advanced tools such as Natural Language Processing and Optical Character Recognition and Image Analytics, unstructured data such as images, documents and clinical notes submitted are analysed to detect cases of potential fraud and abuse."[85] As per the Anti-Fraud Guidelines issued by the National Health Authority, the fraud identification software will be retrospective – to assess patterns of fraud from historical claims, as well as assessing claims on a case-by-case basis. Therefore, the software is possibly being leveraged to accept or deny insurance claims under the scheme. It is as yet unclear what particular algorithmic models or datasets will be utilised in the fraud analytics software.



85 National Health Authority, Annual Report 2018-19, <https://pmjay.gov.in/sites/default/files/2019-09/Annual%20Report%20-%20PMJAY%20small%20version_1.pdf>

What are the consequences of the software flagging a transaction or an individual as 'fraudulent'? According to the Anti-Fraud Guidelines, the Anti-Fraud Cell to whom the fraud has been notified is supposed to ascertain whether there is prima facie evidence of fraud, and if this is found, then to conduct a full investigation which can result in the rejection of an insurance claim and disciplinary action.

Systems like FACTS appear to be effective methods of achieving a legitimate aim – of curtailing fraudulent behaviour. However, in the absence of transparent processes and accountability measures where wrong decisions can be challenged, these systems can result in individuals becoming embroiled in invasive surveillance and pecuniary processes, at times when they are most vulnerable – such as during a medical emergency. These systems can compromise important democratic values and human rights – including the right to privacy and the right to challenge an adverse decision. Recognising this, a similar fraud analytics system, known as the System Risk Indicator or SyRI, was struck down on grounds of violating the European Convention on Human Rights by a Dutch Court in the Hague.[86]

The use of ADMS to determine the scope and extent of rights and entitlements can lead to massive disentitlement and dispossession, without providing adequate justification, and in ways which can be difficult to uncover or report. With government agencies wholeheartedly endorsing 'data-based decision making' for welfare, we must remain cautious of the ways in which these systems can cause injustice at scale, particularly to populations who are the most dependent on the state for their social security and safety.

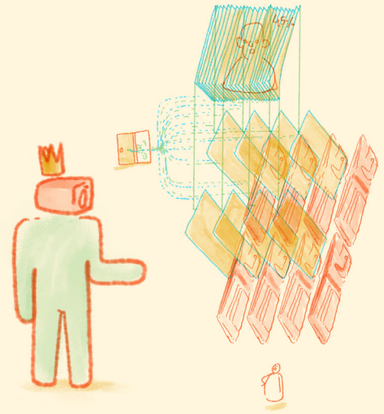
ADMS + Discrimination

Automated Decision Making Systems are often used to generate classifications about individuals or communities, for example, by ascribing certain labels or classifications like 'fraudulent', 'creditworthy', 'criminal' or 'trustworthy' to groups that share common attributes. The use of ADMS to classify and rank information and data, can be systematically biased and discriminatory against individuals or communities which possess certain characteristics, for example, along lines of gender, class, caste or ethnicity, which can lead to unjust social outcomes for these groups. This section interrogates how ADMS in India is used to classify among people and populations and how this can lead to systemic and structural discrimination.

ADMS can be systematically biased against, or in favour of, groups of people who share particular attributes. Often, these biases reproduce socially-embedded and historical discrimination along lines of gender, caste or ethnicity. Due to the

86 Meuwese, A. (2020). Regulating algorithmic decision-making one case at the time. Case note on: District Courtof The Hague , 5/02/20, ECLI:NL:RBDHA:2020:865 (NJCM vs the Netherlands (SyRI)). European Review of Digital Administration & Law, 1(1), 209-211.

scale at which algorithms used in various ADMS are applied, these biases can quickly become pervasive and socially consequential. However, the sources of discrimination can be unaccounted for or overlooked, and often are obscured due to the challenges of transparency in ADMS, making it difficult to identify or rectify.[87]



Multiple sources of bias can exist within ADMS, which can lead to discriminatory outcomes – from the technical architecture of the algorithmic model, the data on which it operates, or the context in which it is used. For example, the historical biases in the kind of data collected within policing databases and used within the CCTNS system can become embedded within ADMS use in policing more generally. Historical police records over-represent particular communities along lines of class and caste, and an automated system relying upon such data is more likely to over-represent such communities, for example, in making decisions about which areas to police.

Historical patterns of discrimination are particularly likely to be reproduced in the functioning of machine learning systems. Since these types of algorithmic systems learn patterns from underlying historical data and apply these patterns in making decisions about future behaviours or phenomenon, they are more likely to recognise and reproduce existing inequalities and discrimination, often in a way which is posited as 'neutral' or 'objective'.[88] Processes which have relied upon machine learning, like modern image recognition and facial recognition systems, or 'predictive policing' systems have been shown to reproduce such historical biases due to underlying biases present in the datasets upon which they operate. For example, studies of facial recognition technologies have consistently shown how their performance varies according to ethnicity – in part because the images on which they are 'trained' are not diverse.[89]

87 Barocas S and Selbst AD, 'Big Data's Disparate Impact', 104 California Law Review 671 (2016).

88 Chouldechova, A., 'Fair prediction with disparate impact: A study of bias in recidivism prediction instruments', 5(2) Big Data, 153–163, 2017.

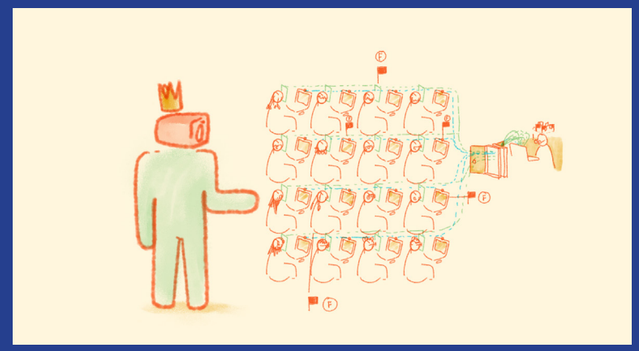
89 Buolamwini J and Gebru T, 'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification', Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR 81:77-91, 2018

[Case Study: Automated Facial Recognition System and the NIST Standards]

In July, 2019, the National Crime Records Bureau began the process of inviting bids for the installation of an 'Automated Facial Recognition System' which would connect to policing and law enforcement databases around the country, in order to build a centralised, national facial recognition system. In the procurement documents like the Request for Proposal, the NCRB has relied upon specific technical standards in order to indicate the proposed 'efficiency' and reliability of the software. In particular, the NCRB has relied on the test for the accuracy of FRT conducted by the National Institute of Standards and Technologies, along with technical demonstrations of how 'accurate' the technology is, based on data provided by the NCRB.

> ***Accuracy, however, is a highly contingent metric. As per the NIST itself, facial recognition algorithms perform differently across different demographics - along lines of gender and ethnicity. Additionally, these systems perform very differently under 'test conditions' as against when they are deployed in real-world scenarios.***

The NIST itself recognises this and creates distinct benchmarks for its Facial Recognition Vendor Test (FRVT), relied upon by the NCRB. The 'accuracy' ultimately depends on the underlying data presented to the system, as well as the new data on which it is expected to operate. Therefore, a single test or number, such as the one indicated under the RFP, is insufficient to understand whether the system will function 'correctly' across demographics. A testing process which does not account for potential discrimination could, therefore, falsely categorise a system as being 'accurate' based only on its performance on a particular group.



Discriminatory FRT use in law enforcement can have severe consequences – not only for individuals who may be falsely flagged and subject to invasive policing procedures, but particularly when reproduced at a social level, they can replicate, enhance, and potentially and perversely justify discriminatory police practice against minorities and marginalised populations.

Another source of discrimination is in the modelling of the algorithmic system and the biases inherent in the task of selecting the inputs and outputs of the algorithmic system. The choice of selecting particular data as relevant factors in classification or prediction is key to the task of algorithmic modelling, and can embed assumptions which lead to discrimination. For example, a software known as COMPAS, used in the USA to determine risk for releasing undertrial prisoners on bail, had been shown to systematically discriminate against people of colour. One possible reason for this discrimination could be the factors which were chosen to indicate 'risk of re-offending' – including prior arrests and arrests of close families or friends. These data or 'features' were more likely to occur for people of colour, and consequently, these factors could have influenced the algorithm to systematically indicate higher risk for these groups.[90] Similarly, a study of the Delhi Police's 'Crime Mapping' software C-MAPS indicated that the filters used to identify areas for policing or crime hotspots are classified using filters for immigrant settlements or minority areas.[91] Such biases may also be embedded within algorithmic design unintentionally, through the lack of testing or considerations of diverse populations.[92] For example, the Aadhar systems fingerprinting and biometric pattern matching algorithm has been shown to underperform on particular demographics, including based on age and gender. [93]Examples like this also show how proxy characteristics – or indirect discrimination – can be built into algorithmic systems, even if the algorithm does not directly consider protected attributes like race or gender as an input in making decisions.

The various sources of biases inherent withing ADMS, coupled with the failures of accountability and transparency outlined in this toolkit, make discrimination within the large technological systems used by public agencies difficult to uncover and challenge. Moving ahead with governance through ADMS without critically reflecting on how these challenges of discrimination can be resolved is antithetical to principles of substantive equality and non-discrimination that we value.

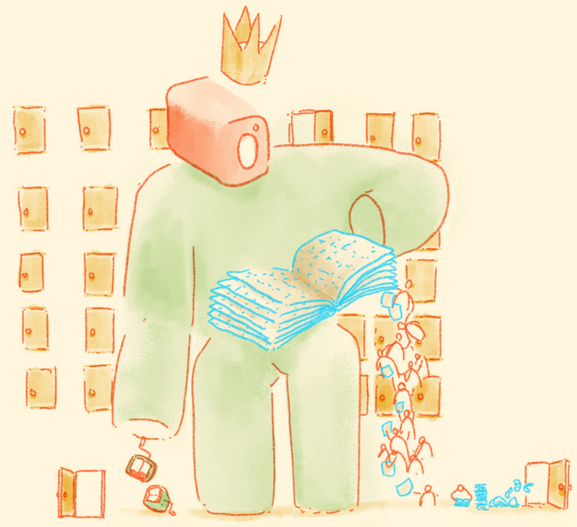
90 Machine Bias', ProPublica (2016) <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

91 Narayan, S, and Marda, V, 'Data in New Delhi's Predictive Policing System, Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, <https://dl.acm.org/doi/abs/10.1145/3351095.3372865>

92 Drozdowski P and others, 'Demographic Bias in Biometrics: A Survey on an Emerging Challenge' [2020] IEEE Transactions on Technology and Society (Volume: 1, Issue: 2, June 2020)

93 Rao U and Nair V, 'Aadhaar: Governing with Biometrics' (2019) 42 South Asia: Journal of South Asian Studies 469

Glossary of Terms



Artificial Intelligence

Artificial Intelligence or 'AI' has no legal or standard definition. It usually refers to the use of computational systems in a context where their performance is considered as displaying intelligence and agency normally attributed to people. 'AI' nowadays often refers to Machine Learning systems, which use data to make predictions and classifications.

Algorithm

An algorithm refers to a series of steps to convert a particular input into a specified output. For the purpose of this resource, 'algorithms' are the instructions or logic which is used by a computer to perform a specific task.

Within many Automated Decision-Making Systems, the algorithm is part of the specific software or computer programme used to process information from a database and generate a 'classification'.

The term 'algorithm' here is often invoked to refer to a wider practice of making decisions and producing knowledge through the use of computational systems, which requires information about people and things to be converted into computational data, and which formalises social phenomenon as mathematical or quantifiable problems which can be solved through effective computation.

Example: Facial Recognition Technology used by the Delhi Police utilises a machine learning algorithm, to create a digital 'copy' of a person's face and find a match within police databases.

Automated Decision-Making Systems
An Automated Decision-Making System (ADMS) here refers to any system which uses computational and algorithmic tools to automatically process information and generate a decision which is of consequence to an individual or a community.

ADMS use algorithms to automatically process information in different ways, but it is a broader term than 'algorithm', which also capture the underlying technologies and computing infrastructures, as well as the human processes and the specific context in which algorithms are used.

Example: The Government of Telangana's 'Samagra Vedika' system uses data analysis to place people in categories for the purpose of allocating welfare resources. This analysis informs the decisions of the Government for the purpose of welfare administration.

Biometric Data
Biometric data refers to data which measure biological human characteristics, such as fingerprints, genetic information, gait, blood type, or facial maps in facial recognition technologies.
Example: Fingerprints and iris scans used in the Aadhaar system are examples of biometric data, which is captured and stored in a digital form.

Bias
In statistics and machine learning, bias is the systematic preference that a particular algorithm displays for some kinds of outputs over others. There are multiple sources of biases in algorithms and ADMS.

Depending on the context in which it is deployed, biases in algorithmic systems can lead to harmful or illegal forms of discrimination.

Example: If a targeted advertising algorithm consistently prefers to show higher-earning job opportunities to men over women (with all else being equal), the system can be said to be biased in favour of men.

Data
Data is an abstraction of any real world phenomenon into a fixed and usable format which can be read by humans or by computers, as digital data.

Database
A database is an organisation of data points in a usable manner, particularly of digital data.

Example: The Centralised Identities Data Repository is a digital database of all biometric information collected within the Aadhaar system.

Data Protection
Data protection refers to various practices and efforts towards ensuring individual or collective control over forms of data to prevent misuse.

Discrimination
In law, discrimination is the systematic unfair or unequal treatment of an individual or a community based on certain attributes like caste, race, class, gender or sex, and unrelated to a legitimate objective.

Example: The refusal to provide government services, or entry into a public house or establishment, on the basis of religion, caste or race is a form of discrimination prohibited by the Constitution of India.

Due Process
Due Process is a procedural threshold which must be satisfied when a government agency makes any consequential decision about an individual. The procedural elements of due process vary across contexts, but generally include the right to notice, the right to a hearing, the right to present countervailing evidence, and the right to an appeal.

Privacy, Right to
The Right to Privacy is a fundamental right protected under Part III of the Constitution of India. The Right to Privacy was explicitly affirmed by a 9-judge bench of the Supreme Court of India in KS Puttaswamy v Union of India. Privacy over information is closely related to data protection.

Example: You have a right to privacy in personal communications, which means that governments cannot spy on your personal communications without legitimate reasons and only within specific exceptions and limitations.

Machine Learning
Machine Learning is a property whereby an algorithm's performance in a defined output improves with 'experience' in the form of more data or more computational power. Machine learning systems are computerised algorithmic systems, which can produce a particular output without explicitly being programmed to do so by a human.

Machine learning is widely considered a form of Artificial Intelligence, and is used to make predictions and classifications about human behaviour and phenomenon.

Example: The Income Tax Department uses Machine Learning to classify potential instances of tax fraud' based on historical examples of 'fraudulent behaviour'.

Profiling
Profiling refers to the process of using algorithmic systems to sort individuals into certain classifications or categories, or to predict attributes which were not

disclosed.

Example: Policing agencies use algorithms deployed on social media to link individuals to 'potential criminal associates'. Inferring possible criminal associations through social media behaviour is an example of 'profiling'.

Classification
Classification is an algorithmic process by which input data is processed and sorted into different categories or classes according to the function of the algorithm.

Source Code
Source Code is the written description which specifies the actions performed by a computer programme.

Equality, Right to
The Right to Equality is a fundamental right under the Constitution of India. It requires the State to provide equal protection of the law to every person, and prohibits discrimination on certain protected grounds. It also protects against arbitrary and unreasonable government action, particularly in the creation of unreasonable and irrational classifications used to discriminate against people.

Big Data
Big Data generally refers to the use of computational analytics to find patterns among large and diverse sets of data.