DISCRIMINATORY BY DEFAULT?

Case studies on the
discriminatory impacts of
algorithmic decision making systems

EQUAL RIGHTS TRUST

INTRODUCTION

In 2023, algorithmic decision-making systems (algorithmic systems)[1] are being used to make decisions in almost every area of human life. These systems are being used to automate tasks, minimise human intervention and reduce cost in areas ranging from recruitment to law enforcement; they are being used to deliver new goods, products and experiences; and they are being deployed to transform the ways in which existing services, such as education and healthcare, are provided.[2] Their influence and impact are pervasive and ever-expanding.

The optimistic once suggested that algorithmic systems — through using data and automating decisions — could create processes which are more efficient, more objective, and free from human bias.[3] Yet there is growing recognition that — far from eliminating discrimination — the use of algorithmic systems frequently results in discriminatory impacts.

Discrimination can and does occur at every point in the lifecycle of these technologies — because of the way in which problems are identified or defined; because of the data used to create or train systems; because of the design of algorithms and systems; and because of the manner in which these systems are deployed and used. There is a growing body of evidence that these systems frequently reinforce existing patterns of discrimination, reflect stereotypical assumptions, and replicate bias. Indeed, because of the way in which algorithmic systems are developed and designed, trained and evaluated, deployed and used, they can be considered to be **discriminatory by default.**

This report aims to illustrate the scope and the scale of what has been termed **algorithmic discrimination** — discrimination which occurs as a result of the design, development or deployment of an algorithmic decision-making system. It does so by presenting fifteen case studies, drawn from existing literature produced by academics, non-governmental organisations, and governmental institutions, which exemplify some of the many different ways in which the use of these systems can cause discrimination.

This report does not seek to be exhaustive, comprehensive, or definitive: the case studies presented here are illustrative. The report does not present findings from original, primary research, but instead compiles and analyses examples and case studies from existing literature. Our aim in collecting and presenting this selection of case studies is to illustrate and exemplify the wide range of ways in which the use of algorithmic systems can result in discriminatory impacts, and thus to demonstrate the scale and the scope of the challenge.

---

[1] We use the term algorithmic decision-making system (algorithmic system) to refer to any system or process through which an automated system is used as part of a decision-making process. Algorithmic system is a broad term to describe any system which uses data and statistical analyses to make decisions or propose solutions. Algorithmic systems include a broad range of tools, systems and processes including both simple automated systems and different types of artificial intelligence (AI), including rule-based AI and machine learning.

[2] Castelluccia, C., Le Métayer, D., "Understanding algorithmic decision-making: Opportunities and challenges", 2019, available at: https://www.europarl.europa.eu/stoa/en/document/EPRS_STU(2019)624261, p. I.

[3] Ackermann, A., "Are Decision-Making Algorithms Always Right, Fair and Reliable or NOT?", *Liberties*, 25 March 2022, available at: https://www.liberties.eu/en/stories/decision-making-algorithm/44109.

Definitions

We use the term **algorithmic decision-making system** (algorithmic system) to refer to any system or process through which an automated system is used as part of a decision-making process. Algorithmic system is a broad term to describe any system which uses data and statistical analyses to make decisions or propose solutions. Algorithmic systems include a broad range of tools, systems and processes including both simple automated systems and different types of artificial intelligence (AI), including rule-based AI and machine learning.

The level of human involvement in algorithmic systems can vary significantly. Some algorithmic systems simply gather data and present it in a readable format to help a human operator come to their own decision, while others produce fully automated decisions, with little or no human intervention. As we use the term, algorithmic systems include both those systems which are fully automated as well as those where humans are in the decision-making loop. This report presents case studies where a variety of different types of algorithmic systems, with varying levels of human intervention, have been used.

We use the term **algorithmic discrimination** to refer to any discrimination which occurs as a result of the design, development, or deployment of an algorithmic decision-making system. Discrimination is any unfavourable treatment or disproportionate impact that arises in connection with one or more protected characteristics or grounds of discrimination.[4]

Discrimination can arise on any one or any combination of the wide range of ***grounds*** recognised under international law, including — but not limited to: age; descent, including caste; disability; gender expression or gender identity; health status; indigenous origin; political or other opinion; race or ethnicity; religion or belief; sex or gender; sexual orientation and other status.[5] The inclusion of "other status" allows for the recognition of grounds which have not yet been recognised but which are analogous to those explicitly listed in extant legal instruments. Discrimination can occur on the basis of the perception that a person has a characteristic or status related to a ground, due to the association of an individual with a ground, and due to the interaction of multiple grounds (what is referred to as multiple or intersectional discrimination).

Discrimination takes different forms. ***Direct discrimination*** occurs where a person is treated less favourably or subjected to a detriment because of one or more protected characteristics. ***Indirect discrimination*** occurs where a provision, criterion, or practice has or would have a disproportionate negative impact on persons having a status or a characteristic associated with one or more grounds of discrimination. ***Denial of reasonable accommodation*** occurs where there is a failure to make necessary and appropriate modifications or adjustments or support to ensure the participation, on an equal basis with others, of persons sharing a particular protected characteristic or characteristics. ***Harassment*** occurs where unwanted conduct has the purpose or effect of creating an intimidating, degrading, humiliating or offensive environment in connection with a protected characteristic or characteristics. ***Segregation*** occurs where

[4] For a detailed discussion of the definition of discrimination under international law, see: United Nations Human Rights Office, *Protecting Minority Rights: A Practical Guide to Developing Comprehensive Anti-Discrimination Legislation*, New York and Geneva, 2023, Part 2, Section I.A.

[5] For a complete list of grounds recognised at international law, see: United Nations Human Rights Office, *Protecting Minority Rights: A Practical Guide to Developing Comprehensive Anti-Discrimination Legislation*, New York and Geneva, 2023, Part 2, Section I.A.

persons sharing a particular protected characteristic are, without their full, free and informed consent, separated and provided different access to institutions, goods, services, and rights.

As with all forms of discrimination, algorithmic discrimination can be both intentional and unintentional. Discrimination should be prohibited and prevented in all areas of life regulated by law, and both public and private actors should be subject to duties of non-discrimination.

All algorithmic systems go through a **process of development**. While the specific elements of the development process will vary from system to system, there are a number of common "stages" in the process. For the purposes of organising and presenting the case studies in this report, we distinguish five different stages in the development and deployment of these systems: inception; design; training; learning; and deployment and use. While we acknowledge that in many cases, these stages may overlap, run in parallel or be difficult to distinguish, examining the development and use of the systems in this way enables a better understanding of the different ways in which algorithmic discrimination can occur during the process, from inception through to use.

The ***inception*** stage of the development process focuses on identifying the problem that an algorithmic system is developed to solve and determining the scope of the system. At this stage, the purpose of the system is defined and developed, and relevant parameters are determined. Once the purpose and scope of an algorithmic system is determined, the process moves to ***design***. At this stage, those developing the system determine how it will operate and fulfil its function. This stage can include consideration of the variables and factors the algorithm considers or does not consider, where data is drawn from or input and how users interact with the system. In many cases, the inception and design of an algorithmic system are essentially simultaneous, with the purpose of the system defining its design or the design possibilities informing the purpose. While these stages can and do overlap or merge in the development process, it is nevertheless useful to distinguish them when considering discriminatory impacts, as examining these as separate stages enables an understanding of the different ways in which discrimination occurs.

In order to operate, algorithmic decision-making systems are subjected to ***training***, during which data is input, in order that the system can develop and adapt. Many algorithmic systems use machine learning, where the system code enables the system to "learn" over time from data that is presented to it. In this process, the algorithmic system uses data to identify patterns, enabling it to make predictions when presented with new data.[6] Training can be categorised as either supervised or unsupervised. In supervised training, the algorithm is provided pre-labelled data, from which it can identify patterns and rules so that, when it is presented with new, un-labelled data, it is able to classify it. In unsupervised training, the algorithm is presented with data which is not labelled or categorised, and is tasked with finding patterns, commonalities, groups, and structures.[7] This report examines cases where both supervised and unsupervised training have been used, demonstrating that algorithmic systems can have discriminatory impacts, irrespective of the training method.

Algorithmic systems continue to develop after the initial training, through ongoing analysis of the data which they receive during use. This is what we term the ***learning*** stage of the development process. In this

[6] See above, note 2., p. 4.

[7] *Ibid.*

stage, an algorithmic system can continue to learn how to respond to data, based on its environment and how users interact with it. For example, an AI chat-bot — which simulates a conversation with a real person by allowing users to input text and receive an algorithmically generated response — will learn from its users how real people respond to certain questions and use this information to improve its own responses.[8] Again, as with the overlap between inception and design, we distinguish initial training from ongoing learning because it enables a better understanding of the way in which discriminatory impacts occur — and thus the safeguards which need to be taken to prevent it.

The final stage of the development process as we define it is ***deployment and use*** of the system. The way in which algorithmic systems are deployed and the context in which they are used varies significantly, including in ways which system developers do not envisage or predict. The context in which systems are deployed includes aspects of the sector in which the system is used, the geographic region, or the culture or characteristics of the groups and individuals who use it or who are subjected to its use. As the case studies in this report demonstrate, contextual factors play a significant role in whether and how algorithmic systems result in discriminatory impacts.

Purpose, methodology and structure

This report was developed by the Equal Rights Trust, an independent international non-governmental organisation that works in partnership to advance equality through law. The Equal Rights Trust's vision is an equal world: a world in which everyone — irrespective of their identity, status or beliefs — can participate in every area of life on an equal basis with others. We work towards this vision by addressing one of the root causes of inequality: discrimination. We focus our efforts on eliminating discrimination, its consequences and its legacies. We do this through the law. Our mission is to work in partnership to support the development, adoption, implementation and use of equality laws.

This report was developed to accompany the launch of the ***Principles on Equality by Design in Algorithmic Decision-Making***, a standard setting document, developed by the Equal Rights Trust and endorsed by a group of the leading global equality organisations — organisations representing women; national, ethnic and religious minorities; indigenous peoples; persons with disabilities; older persons and LGBTI+ persons. The Principles elaborate why and how states and businesses must adopt a proactive, pre-emptory and precautionary approach to identifying, assessing and addressing the equality impacts of algorithmic decision-making systems, if they are to comply with their obligations, responsibilities and commitments to eliminate discrimination and ensure equal enjoyment of rights and freedoms.

As noted above, this report does not present primary research. Instead, it compiles and analyses case studies presented from a range of different existing sources. Work to develop the report was undertaken in three phases.

First, we conducted desk-based research and consulted with experts in equality and non-discrimination and experts in digital rights. The aim of this stage was to scope and map evidence on the actual and potential discriminatory impacts of algorithmic systems, in order to identify patterns and trends and to establish what kind of evidence base existed.

---

[8] *Ibid.*

In the second phase, we launched a global call for information and evidence of cases or patterns of actual, emerging, and anticipated or potential discrimination arising from the use of AI and algorithmic systems. The call was issued to the Trust's global network of equality activists and organisations, and sought evidence from any country, but with a particular interest in evidence from those working outside the Global North and West.[9]

Once the window for submissions had closed, we moved on to a third stage, in May 2023. In this stage, our aim was to analyse information gathered and provided during the previous stages and through further desk-based research, in order to identify and select case studies for inclusion in this report. Time and resource constraints meant that we were unable to include every case study and example which was shared with us. In selecting the fifteen case studies for inclusion, we sought to illustrate and exemplify the myriad ways in which algorithmic systems can result in discriminatory impacts. Accordingly, in selecting case studies from the information gathered, we sought to include a diverse range of examples of different forms of algorithmic discrimination, arising on the basis of different grounds, in different sectors and areas of life and in different global regions.

The case studies include examples of algorithmic discrimination arising in areas of life ranging from social security to law enforcement, and from healthcare provision to social media. We present examples of algorithmic discrimination arising on the basis of sex and gender, race and ethnicity, gender identity, sexual orientation, and disability, among other characteristics. Case studies are drawn from a wide variety of different countries and context.

We find evidence of discriminatory impacts arising at every stage in the lifecycle of an algorithmic system, from inception to operation. Indeed, these stages provide the structure for the report, which presents three case studies in each of the five "stages" of development of these systems which we set out above. We recognise and acknowledge that these stages are frequently not clearly segregated and distinct. However, our consultations with experts, the evidence we received and the analysis we undertook revealed that discrimination occurs in different ways at different points in the development and implementation process. Accordingly, in our view, organising the information in this way provides the best structure to examine the different dynamics at play.

Scope and Limitations

This report does not include primary or original research. Instead, we have sought to compile, consolidate, analyse and present a diverse range of compelling case studies from existing literature, such as journal articles, government reports, newspaper articles, and documents by civil society organisations which were sourced from a call for evidence and literature review. All case studies have been thoroughly researched, with relevant references and citations provided throughout. Care has been taken to present these sources and the information they provide in context and we have worked to verify the facts, to the extent possible.

[9] For details on this call for evidence, see Equal Rights Trust, *Equal Rights Trust launches call for evidence of AI and algorithmic discrimination*, 2022, available at: https://www.equalrightstrust.org/news/evidence-ai-and-algorithmic-discrimination.

We acknowledge and thank those who undertook the research and studies, developed and litigated the cases and investigated and published the stories which we refer to here.

While we have sought to ensure that the case studies included in this report reflect diverse forms of discrimination, arising in a range of different contexts, a report such as this cannot be exhaustive of the full range of ways in which algorithmic systems can result in discriminatory impacts. This said, we hope that by presenting case studies of discriminatory impacts arising at different points in the development and use of these systems and on a range of different grounds or characteristics, occurring in different parts of the world and in different areas of economic and social life, we can demonstrate the broad array of ways in which the use of algorithmic systems can result in discriminatory impacts.

This report does not provide an in-depth analysis of the ways in which algorithmic systems work — this is not our aim, and it is beyond our expertise. It is also important to stress that identifying evidence of discriminatory impacts does not require a detailed understanding of the way in which a system operates — discrimination is an impact or result, which can be established on the basis that a person has experienced disadvantage or harm connected to a protected characteristic. Our focus is on the impacts of these systems as they operate and are used — on identifying the discriminatory impacts arising from the use of these systems and explaining this in the most accessible way possible. Accordingly, our descriptions of the systems in question are as limited and non-technical as possible, with discussions limited to the facts required to understand the discrimination which occurred in each case.

Findings and conclusions

The examples collected and presented in this short report paint a compelling picture. This small but wide-ranging collection of case studies illustrates the scope and scale of the problem of algorithmic discrimination. Taken together, these case studies demonstrate that:

- Discrimination can and does occur at every stage in the development and deployment of algorithmic systems, from inception through to use.
- This discrimination can and does occur in all areas of life where algorithmic systems are deployed, from healthcare to employment, and from social security to social media.
- This discrimination has arisen on myriad grounds, from gender to nationality and from disability to race, affecting communities exposed to discrimination because of different aspects of their status, identity or beliefs.

The case studies demonstrate that algorithmic discrimination occurs both in situations where the impacts are clearly foreseeable — in particular by those exposed to discrimination — and in situations where these impacts are difficult to identify and predict. They also demonstrate that the use of these systems serves to both replicate and exacerbate existing patterns of discrimination and disadvantage and to create novel discriminatory dynamics and harms.

More broadly, an analysis of these examples demonstrates that because of the way in which algorithmic systems are developed and designed, trained and evaluated, deployed and used, they are frequently discriminatory by default. Systems which some still believe are inherently objective and fair in fact reinforce existing patterns of discrimination, reflect stereotypical assumptions and replicate bias.

In light of this evidence — the broad range of discriminatory impacts, the challenges in identifying and foreseeing these impacts and the potential for these systems to be discriminatory by default — it is essential that states and businesses adopt a new, pre-emptory and precautionary approach. They must adopt an Equality by Design approach to the development and deployment of algorithmic systems. This requires that they take a proactive approach to ensure that potential discriminatory impacts are identified and addressed before they occur and that equality considerations are intentionally incorporated into the design, development, and deployment of algorithmic systems.

CASE STUDIES

A INCEPTION

The inception stage can be considered the first step in the lifecycle of an algorithmic system. It is the stage when the need for the system is determined, the problem which it is intended to address is identified and its purpose is defined. The inception stage therefore involves defining the task for the system, determining the scope of the system and deciding any limits on its operation.

The following case studies — from Brazil, the Netherlands, and Paraguay — illustrate some of the ways in which discrimination can arise at this first stage in the development of the algorithmic system.

1 BRAZIL. Public Transport. Discrimination on the basis of gender, gender identity and expression.

Like many large cities, São Paulo, the most populous city in Brazil, has a mass transport system, popularly known as the 'metro'.

In April 2018, ViaQuatro (a metro operator) announced that it would be installing the **Digital Interactive Doors (DID)** system on its yellow metro line.[10] The DID system consisted of advertising panels with integrated cameras which were designed to "detect human faces and the emotion, gender, and age" of the person looking at the panels.[11] The system's developers claimed that DID was able to identify whether a person was male or female with an accuracy rate of 80 to 90%.[12]

The DID boards were installed at the entrances to metro stations. The aim of the system was, by detecting the characteristics of those passing the advertisements, to "tailor" advertisements to different individuals. ViaQuatro claimed that the system had been installed to "serve as a platform to share information".[13] In addition to sharing information, however, the system also collected and processed the data of the metro-users, without their consent.[14]

---

[10] Arroyo, V., Leufer, D., "Facial recognition on trial: emotion and gender 'detection' under scrutiny in a court case in Brazil", *Access Now*, 29 June 2020, available at: https://www.accessnow.org/facial-recognition-on-trial-emotion-and-gender-detection-under-scrutiny-in-a-court-case-in-brazil/#:~:text=The%20Brazilian%20Institute%20of%20Consumer,passengers%20without%20processing%20personal%20data/.

[11] Access Now, *Expert Opinion in IDEC vs. ViaQuatro*, 2020, available at: https://www.accessnow.org/press-release/data-for-sale-in-brazil/, p. 2.

[12] *Ibid*., p. 15.

[13] Arroyo, V., Leufer, D., "Facial recognition on trial: emotion and gender 'detection' under scrutiny in a court case in Brazil", *Access Now*, 29 June 2020, available at: https://www.accessnow.org/facial-recognition-on-trial-emotion-and-gender-detection-under-scrutiny-in-a-court-case-in-brazil/#:~:text=The%20Brazilian%20Institute%20of%20Consumer,passengers%20without%20processing%20personal%20data/.

[14] *Ibid*.

In August 2018, the Brazilian Institute of Consumer Protection (IDEC) brought a civil action case against ViaQuatro regarding its use of Digital Interactive Doors.[15] IDEC argued that ViaQuatro's use of the DID system violated the rights of the metro users and stated that it should be discontinued.[16]

The international non-governmental organisation, Access Now, in an expert opinion on the case, stated that several aspects of the DID system — as conceived and developed — discriminated against non-binary and trans persons. The system assumed a binary concept of gender, categorising people exclusively as either male or female.[17] In addition, it identified the gender of a person through facial analysis, with a presumption that gender and sex could be identified through the "physiological characteristics of a person's face".[18] In this way, the system not only mis-gendered non-binary and trans persons, but also risked mis-identification of cis-gender persons through the use of stereotypical and gendered perceptions of physical attributes. By assigning gender based on these assumptions, the system undermined the right of trans and non-binary people to self-identification and violated their dignity, helping to perpetuate the cycle of discrimination which individuals in these groups experience.[19]

In September 2018, a Brazilian court found that, by collecting and using the data of the metro users without their full and informed consent, the DID system violated users' right to information and freedom of choice. The court did not consider whether use of the system engaged with and violated the right to non-discrimination, Nevertheless, it ordered ViaQuatro to remove the cameras.[20]

While the claims of discrimination were not considered in court, the potential discriminatory impact of the DID system is clear. When the DID system was conceived, the developers defined sex and gender — whether intentionally or unintentionally — as binary: this was the case both in the output of the algorithmic systems and how it identified individuals. In doing so, the developers failed to consider the effect on transgender and non-binary individuals and on individuals who do not present as stereotypically male or female. Researchers have found that other automated gender recognition (AGR) technologies similarly deny people

---

[15] Canto, M., "Mind the Gap: the Privacy Void in Brazilian's Public Transport" *Oxford Human Rights Hub*, 26 October 2018, available at https://ohrh.law.ox.ac.uk/mind-the-gap-the-privacy-void-in-brazilians-public-transport/.

[16] Institute for Research on Internet and Society, "Public Civil Action IDEC vs. ViaQuatro: IRIS' Opinion", 2019, available at: https://irisbh.com.br/en/publicacoes/public-civil-action-idec-vs-viaquatro-iris-opinion/.

[17] Access Now, *Expert Opinion in IDEC vs. ViaQuatro*, 2020, available at: https://www.accessnow.org/press-release/data-for-sale-in-brazil/, pp. 10-11.

[18] *Ibid*., p. 15.

[19] Arroyo, V., Leufer, D., "Facial recognition on trial: emotion and gender 'detection' under scrutiny in a court case in Brazil", *Access Now*, 29 June 2020, available at: https://www.accessnow.org/facial-recognition-on-trial-emotion-and-gender-detection-under-scrutiny-in-a-court-case-in-brazil/#:~:text=The%20Brazilian%20Institute%20of%20Consumer,passengers%20without%20processing%20personal%20data/.

[20] Canto, M., "Mind the Gap: the Privacy Void in Brazilian's Public Transport" *Oxford Human Rights Hub*, 26 October 2018, available at: https://ohrh.law.ox.ac.uk/mind-the-gap-the-privacy-void-in-brazilians-public-transport/.

the right to self-identify and can lead to misgendering.[21] This can be especially problematic beyond the use of these systems in contexts other than advertising.[22]

2. THE NETHERLANDS: Social Welfare. Discrimination on the basis of nationality.

In the Netherlands, the government provides a specific benefit scheme to support parents to meet the costs of childcare – the **kindersijslag**. Calculated on the basis of a parental income, the system reimburses part of the childcare costs incurred by parents each month.[23]

In 2013, the Dutch taxation authorities introduced an algorithmic decision-making system with the aim of calculating the risk of individuals making fraudulent claims for childcare cost reimbursement. The system created risk profiles based on criteria developed by the taxation authorities.[24] These criteria included not having Dutch citizenship, which was considered to be an indicator of high risk.[25] This resulted in the system disproportionately identifying non-Dutch nationals — and consequently persons from certain ethnic groups — as more likely to commit fraud than Dutch nationals.[26]

Relying upon the system, the taxation authorities wrongly accused thousands of people of fraud.[27] Those accused of benefit fraud were forced to repay *all* of the childcare benefits they had received — often equivalent to thousands of Euros.[28] The consequences were deep and far-ranging. Some of those accused took their own lives; tens of thousands of families were forced into poverty; and over a thousand children were placed in foster care.[29] The scandal was revealed in 2018, ultimately leading to the resignation of the

---

[21] Leufer, D., "How AI Systems Undermine LGBTQ Identity", *Access now*, 6 April 2021, available at: https://www.accessnow.org/how-ai-systems-undermine-lgbtq-identity/; See also, Keyes, O.S., "The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition" *Proceedings of the ACM on Human-Computer Interaction*, Vol. 2, No. 88, 2018, p. 1.

[22] Leufer, D., "Computers are binary, people are not: how AI systems undermine LGBTQ identity", *Access now*, 13 January 2023, available at: https://www.accessnow.org/how-ai-systems-undermine-lgbtq-identity/.

[23] Geiger, G., "How a Discriminatory Algorithm Wrongly Accused Thousands of Families of Fraud", *Vice*, 1 March 2021, available at: https://www.vice.com/en/article/jgq35d/how-a-discriminatory-algorithm-wrongly-accused-thousands-of-families-of-fraud.

[24] Heikkilä, M., "AI: Decoded: A Dutch algorithm scandal serves a warning to Europe — The AI Act won't save us", 30 March 2022, available at: https://www.politico.eu/newsletter/ai-decoded/a-dutch-algorithm-scandal-serves-a-warning-to-europe-the-ai-act-wont-save-us-2/.

[25] Amnesty International, *Xenophobic Machines, Discrimination Through Unregulated Use of Algorithms in The Dutch Childcare Benefits Scandal*, 2021, available at: https://www.amnesty.org/en/documents/eur35/4686/2021/en/, p. 16.

[26] *Ibid.*

[27] Geiger, G., "How a Discriminatory Algorithm Wrongly Accused Thousands of Families of Fraud", *Vice*, 1 March 2021, available at: https://www.vice.com/en/article/jgq35d/how-a-discriminatory-algorithm-wrongly-accused-thousands-of-families-of-fraud.

[28] *Ibid.*

[29] Heikkilä, M., "AI: Decoded: A Dutch algorithm scandal serves a warning to Europe — The AI Act won't save us", 30 March 2022, available at: https://www.politico.eu/newsletter/ai-decoded/a-dutch-algorithm-scandal-serves-a-warning-to-europe-the-ai-act-wont-save-us-2/.

Dutch Cabinet in 2021.[30] The Data Protection Authority imposed a fine of Euro 2.75 million on the Tax Administration for its unlawful and discriminatory processing of the childcare benefit applicants' data.[31]

In this case, nationality was identified as a key factor in determining likelihood of fraud at the conception of the system — it was a factor which was designed into the system intentionally, because of the perceived risk of fraud. Indeed, this criterion assumed such importance in the operation of the system that nationality became the defining factor in fraud assessment in many cases, leading to large scale misreporting, and ultimately to the misidentification of thousands of cases of "fraud".

3. PARAGUAY: Employment. Discrimination on the basis of language and ethnicity.

In 2018, the government of Paraguay launched an online platform — **ParaEmpleo** — to help people find jobs. The intention is that job-seekers use the platform to create a profile, listing their relevant qualifications, skills and specialisations.[32] The ParaEmpleo algorithm then matches the user with employment opportunities suited to their profile and recommends relevant courses to increase their chances of finding employment.[33]

ParaEmpleo is available to use only in Spanish and English. This is despite the fact that the country has two official languages — Spanish and Guaraní. Approximately 90% of the population speak Guaraní. Guaraní is spoken by the indigenous Guaraní people, many of whom are not bilingual;[34] a 2012 census found that 48% of the indigenous population speak Guarani as their main language.[35] Accordingly, the system's accessibility is not only limited on the basis of language, but also in a way which disproportionately impacts on the members of the Guarani ethnic group.[36]

---

[30] Amnesty International, *Xenophobic Machines, Discrimination Through Unregulated Use of Algorithms in The Dutch Childcare Benefits Scandal*, 2021, available at: https://www.amnesty.org/en/documents/eur35/4686/2021/en/, p. 14.

[31] Dutch Data Protection Authority (Autoriteit Persoonsgegevens), *Tax Administration Fined for Discriminatory and Unlawful Data Processing*, 7 December 2021, available at https://autoriteitpersoonsgegevens.nl/en/current/tax-administration-fined-for-discriminatory-and-unlawful-data-processing.

[32] Plata, G., "Artificial Intelligence: The New Way to Get a Job in Paraguay", *Inter-American Development Bank*, accessed 26 July 2023, available at: https://www.iadb.org/en/improvinglives/algorithms-get-you-job-paraguay.

[33] *Ibid.*

[34] Equal Rights Trust, "Submission to the UN Special Rapporteur on Contemporary Forms of Racism, Racial Discrimination, Xenophobia and Related Intolerance on New Information Technologies, Racial Equality and Non-discrimination", December 2019, available at: https://www.equalrightstrust.org/news/equal-rights-trusts-submission-un-special-rapporteur-contemporary-forms-racism, para. 13; Katanich D., "Why Is the Guaraní Language Playing a Big Role in Paraguay's Election?', *Euronews*, last updated 29 April 2023, available at: https://www.euronews.com/culture/2023/04/29/discover-the-indigenous-language-at-the-heart-of-paraguays-upcoming-presidential-election.

[35] Costa, W., "'Culture is language": why an indigenous tongue is thriving in Paraguay", *The Guardian,* accessed 3 October 2023, available at: https://www.theguardian.com/world/2020/sep/03/paraguay-guarani-indigenous-language

[36] Equal Rights Trust, "Submission to the UN Special Rapporteur on Contemporary Forms of Racism, Racial Discrimination, Xenophobia and Related Intolerance on New Information Technologies, Racial Equality and Non-discrimination", December 2019, available at: https://www.equalrightstrust.org/news/equal-rights-trusts-submission-un-special-rapporteur-contemporary-forms-racism, para. 13.

As with the examples above, in this case, discrimination was built into the ParaEmpleo system from its inception, prior to any work to develop or design the system, when the choice was made to restrict the language of operation.

B DESIGN

Once the concept for an algorithmic system — the problem which it seeks to address and the means through which it aims to address it — has been developed, the system itself must be designed. Designing an algorithmic decision-making system can include, among other things: developing the algorithm; creating code; determining the data the algorithm will be trained on; and creating a user interface.[37] For example, in the case of a "predictive algorithm" — meaning an algorithm used to predict the likelihood of an event — the design stage could involve deciding what factors or variables the algorithm would consider, or not consider, when predicting likelihood of a particular outcome.

Three case studies — from the United States of America, Jordan, and New Zealand — involving the use of algorithmic systems in three very different public functions — access to healthcare, allocation of social welfare, and prevention of crime — demonstrate how discriminatory impacts can arise during the design stage. These case studies illuminate how the way in which an algorithm is designed, and the factors which the algorithm is designed to consider, can result in discriminatory outcomes on the basis of a range pf different protected characteristics.

4 UNITED STATES OF AMERICA: Healthcare. Discrimination on the basis of race.

In 2019, a study by a group of academics - Ziad Obermeyer, Brian Powers, Christine Vogeli and Sendhil Mullainathan – found that an algorithmic system used in healthcare management was having widespread discriminatory impacts on the basis of race. In the United States of America, a significant number of hospitals use an algorithmic system developed by **Optum**, to manage the delivery of healthcare.[38] The purported purpose of the system is to identify people with greater healthcare needs, in order provide them with relevant, targeted care and assistance.[39]

The system developed by Optum is designed to predict which patients are likely to require extra care and so enable hospitals to identify patients eligible for "high-risk care management" programs. Such programs, aimed at patients with particularly complex health needs, provide access to additional resources such as more primary-care visits and nursing staff with special training.[40] Optum provides services across the US,

---

[37] Marabelli, M., Newell, S., Handaunge, V., "The Lifecycle of Algorithmic Decision-Making Systems: Organizational Choices and Ethical Challenges", Forthcoming, last revised 27 April 2023, available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3891294, p. 6.

[38] *Ibid.* See also: *Optum*, "Care by State", accessed 3 October 2023, available at: https://www.optum.com/care/locations.html.

[39] Paul, K., "Healthcare algorithm used across America has dramatic racial biases", *The Guardian*, 25 October 2019, available at: https://www.theguardian.com/society/2019/oct/25/healthcare-algorithm-racial-biases-optum.

[40] Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S., "Dissecting racial bias in an algorithm used to manage the health of populations", *Science*, Vol. 366, No. 6464, 2019, p. 1.

and — while it is not known how many hospitals have been using Optum's algorithm — it is estimated that the algorithm is applied in the care management of approximately 200 million people annually.[41]

In their study, Obermeyer, Powers, Vogeli, and Mullainathan, found that the Optum algorithmic system was less likely to refer Black patients than white patients for high-risk care management programs.[42] The study found that even in cases where Black patients had the same or similar healthcare issues, they were less likely to be referred and that "at a given risk score, Black patients are considerably sicker than White patients, as evidenced by signs of uncontrolled illnesses".[43] Out of the patients referred by the algorithm to receive extra care, only 17.7% were Black.[44] According to the authors, addressing the racial disparity in the operation of the system would "increase the percentage of Black patients receiving additional help from 17.7 to 46.5%".[45]

The Optum algorithm was designed to use "health costs as a proxy for health needs".[46] This reflected an assumption that healthcare costs are indicative of medical needs — that those with greater medical need would experience higher cost.[47] Accordingly, the algorithm assigned patients a risk "score" based on their total annual healthcare expenditure.[48] In practice, this metric was not an appropriate proxy and did not accurately measure, indicate or predict actual healthcare needs.

On comparing the actual healthcare costs incurred by patients with the "same number of chronic health problems", the 2019 study found that, on average, the care paid for by Black patients cost approximately US $1,800 less, per annum than that provided to white patients.[49] The study posited a number of potential explanations for this difference, including "distrust of the health-care system" and "direct racial

---

[41] *Ibid.* See also: *Optum*, "Care by State", accessed 3 October 2023, available at: https://www.optum.com/care/locations.html.

[42] *Ibid*; Ledford, H., "Millions of Black People Affected by Racial Bias in Health-Care Algorithms", *Nature*, 2019, accessed 26 July 2023, available at: https://www.nature.com/articles/d41586-019-03228-6#:~:text=Study%20reveals%20rampant%20racism%20in,highlights%20ways%20to%20correct%20it.&text=An%20algorithm%20widely%20used%20in,a%20sweeping%20analysis%20has%20found.

[43] *Ibid*; Ledford, H., "Millions of Black People Affected by Racial Bias in Health-Care Algorithms", *Nature*, 2019, accessed 26 July 2023, available at: https://www.nature.com/articles/d41586-019-03228-6#:~:text=Study%20reveals%20rampant%20racism%20in,highlights%20ways%20to%20correct%20it.&text=An%20algorithm%20widely%20used%20in,a%20sweeping%20analysis%20has%20found.

[44] Obermeyer, Z., Powers, B., Voegli, C., Mullainathan, S., "Dissecting racial bias in an algorithm used to manage the health of populations", *Science*, Vol. 366, No. 6464, 2019, p. 3.

[45] *Ibid*.

[46] Obermeyer, Z., Powers, B., Voegli, C., Mullainathan, S., "Dissecting racial bias in an algorithm used to manage the health of populations", *Science*, Vol. 366, No. 6464, 2019.

[47] Ledford, H., "Millions of Black People Affected by Racial Bias in Health-Care Algorithms", *Nature*, 2019, accessed 26 July 2023, available at: https://www.nature.com/articles/d41586-019-03228-6#:~:text=Study%20reveals%20rampant%20racism%20in,highlights%20ways%20to%20correct%20it.&text=An%20algorithm%20widely%20used%20in,a%20sweeping%20analysis%20has%20found.

[48] *Ibid.*

[49] *Ibid.*, p 4.

discrimination by health-care providers". Irrespective of the reasons, the study found that Black patients were less likely to seek or be provided with treatment and so incurred lower costs.[50]

Through relying on costs as a proxy for need, the system replicated pre-existing patterns of discrimination and inequality in healthcare, exacerbating and embedding discrimination in access to services. The algorithm was designed to consider costs as a primary proxy indicator of health needs, without considering the various factors that contribute to the cost of healthcare which includes these systemic and structural inequalities. Those inequalities were then replicated and exacerbated by the algorithm.

5 JORDAN: Social Welfare. Discrimination on the basis of gender and nationality.

The Government of Jordan operates an automated cash transfer program called **Takaful**, financed by the World Bank, which is aimed at alleviating poverty.[51] Takaful uses an algorithm to identify families which are eligible for financial assistance, by ranking them in order of need based on a set of 57 socio-economic indicators including "household size".[52] Household size is considered to be a reasonable indicator of financial need as "the more people a household must feed, the higher the need".[53]

The international non-governmental organisation Human Rights Watch has reported that, when calculating household size, the Takaful algorithm counts only those household members with Jordanian citizenship. This is directly discriminatory on the basis of nationality —households including non-nationals receive less financial assistance.

In addition, however, the system also has a gender discriminatory impact. Under Jordanian law, only men can pass on citizenship to their spouse and children; Jordanian women only have the right to pass on their citizenship in certain specific cases.[54] Jordanian women married to non-Jordanian men cannot pass on Jordanian citizenship to their spouse or other family members. As a result, a household that consists of a Jordanian woman with a non-Jordanian spouse is counted by the algorithm as a one-person household.[55] In many cases, these households are not eligible for assistance; in the few instances where such a household is eligible for financial aid, it is only entitled to the minimum amount.[56]

---

[50] Ledford, H., "Millions of Black People Affected by Racial Bias in Health-Care Algorithms", *Nature*, 2019, accessed 26 July 2023, available at: https://www.nature.com/articles/d41586-019-03228-6#:~:text=Study%20reveals%20rampant%20racism%20in,highlights%20ways%20to%20correct%20it.&text=An%20algorithm%20widely%20used%20in,a%20sweeping%20analysis%20has%20found.

[51] Ryan-Mosley, T., "An Algorithm Intended to Reduce Poverty in Jordan Disqualifies People in Need", *MIT Technology Review*, 13 June 2023, available at: https://www.technologyreview.com/2023/06/13/1074551/an-algorithm-intended-to-reduce-poverty-in-jordan-disqualifies-people-in-need/.

[52] Human Rights Watch, "Automated Neglect: How The World Bank's Push to Allocate Cash Assistance Using Algorithms Threatens Rights", 13 June 2023, available at: https://www.hrw.org/report/2023/06/13/automated-neglect/how-world-banks-push-allocate-cash-assistance-using-algorithms, pp. 39-40.

[53] *Ibid*., p. 45.

[54] *Ibid*., p. 46.

[55] *Ibid*., pp. 40, 46.

[56] *Ibid*., p. 46.

As this demonstrates, while the decision to consider household size as an indicator of poverty appears reasonable and justifiable, the specific design choice regarding the definition of household size has severe discriminatory impacts. In this instance, the algorithm discriminated against those with non-Jordanian nationality and women, by replicating and reinforcing existing inequalities, which are themselves the result of discriminatory legal provisions.

6 NEW ZEALAND: Criminal Justice. Discrimination on the basis of ethnicity.

As part of its offender management system, the New Zealand Department of Corrections uses an algorithmic risk assessment tool called **ROC*ROI** (Risk of Re-Conviction multiplied by Risk of Imprisonment).[57] The algorithm is used to predict recidivism and plays a central role in decisions related to sentencing and parole.[58]

The algorithm calculates the probability of an offender being re-convicted and re-imprisoned within five years from the date of assessment.[59] It conducts this calculation with reference to approximately 30 different factors. The focus is on an offender's prior conviction history: data such as the number of past convictions for a crime, the number of prior imprisonments, and the total time spent in prison are all considered.[60] The algorithm assigns each of these risk factors a numerical score and uses that to calculate the chances of offenders re-offending.

Because the algorithm focuses on risk factors associated with past criminal conviction, it replicates existing systemic and structural discrimination against Māori persons in the criminal justice system. The Māori are New Zealand's indigenous ethnic community and constitute 17.4% of the population (as of June 2022).[61] Due to a variety of systemic and structural inequalities, Māori persons are more likely to have been arrested, charged, and convicted of crimes than members of the non- Māori population.[62] In fact, when factors such as gender, socio-economic status, and educational qualifications are controlled for, Māori people have been found to be between 1.6 and 2.4 times more likely to have a criminal conviction than non-Māori people.[63]

---

[57] Bakker, L., Riley, D., O'Malley, J., "Risk of Re-Conviction: Statistical Models which predict four types of re-offending", 1999, available at: https://www.corrections.govt.nz/resources/research/risk-of-reconviction, p. 15; New Zealand Government, Stats New Zealand, *Algorithm Assessment Report*, 2018, available at: https://data.govt.nz/toolkit/data-ethics/government-algorithm-transparency-and-accountability/algorithm-assessment-report/, p. 21.

[58] Keddell, M., "How Fair is an Algorithm? A Comment on the Algorithm Assessment Report", 7 December 2018, available at: https://reimaginingsocialwork.nz/2018/12/07/how-fair-is-an-algorithm-a-comment-on-the-algorithm-assessment-report/.

[59] New Zealand Government, Stats New Zealand, *Algorithm Assessment Report*, 2018, available at: https://data.govt.nz/toolkit/data-ethics/government-algorithm-transparency-and-accountability/algorithm-assessment-report/, p. 21.

[60] *Ibid.*

[61] Stats NZ*, Māori Population Estimates: At 30 June 2022*, 17 November 2022, available at: https://www.stats.govt.nz/information-releases/maori-population-estimates-at-30-june-2022/

[62] Keddell, M., "How Fair is an Algorithm? A Comment on the Algorithm Assessment Report", 7 December 2018, available at: https://reimaginingsocialwork.nz/2018/12/07/how-fair-is-an-algorithm-a-comment-on-the-algorithm-assessment-report/.

[63] Fergusson, D., Horwood, L., Swain-Campbell, N., "Ethnicity and Criminal Convictions: Results of a 21-year Longitudinal Study", *Journal of Criminology*, Vol. 36, No. 3, December 2003, pp. 354, 364.

As the ROC*ROI system has been designed to use past conviction as an indicator of re-offending risk, it replicates this bias in the data and disproportionately classifies Māori persons as more likely to re-offend.[64]

The ROC*ROI measure, therefore, replicates and exacerbates the patterns of racial discrimination and inequality which already exist in New Zealand's justice system. While prior convictions are clearly a relevant factor in calculating the risk of re-offending, unless the potential equality impacts of such an approach are fully assessed and addressed, the system will have indirectly discriminatory impacts.

C TRAINING

As each of the three examples in the preceding section illustrate, algorithmic systems work on the basis of data. Algorithms are "trained" using existing data sets to identify patterns and calculate needs, risks, likelihoods or other outputs or outcomes. Once an algorithmic system has been designed, it is trained so that it "learns" how to respond when presented with new data. This is an essential stage in the development of algorithmic decision-making systems, as the training and learning process determines how the system will behave and how it will reach its conclusions.

However, as some of the case studies already examined indicate, many pre-existing datasets are not properly representative of whole populations or reflect past or current patterns of discrimination and inequality. Unless care is taken in the selection, use and management of data and in the way in which the algorithm is developed to understand and use this data, algorithmic systems will simply replicate — or even amplify and exacerbate — existing inequalities. The three case studies below examine situations in which algorithmic systems have been trained using very different datasets — photographs, speech, and statistics — and demonstrate how the use of data and the training of these systems can result in discriminatory impacts.

7 GLOBAL: Competition. Discrimination on the basis of race.

Beauty.AI is a website developed by Youth Laboratories, a company based in Russia and Hong Kong.[65] In 2016, Beauty.AI ran an international beauty contest, in which an estimated 6,000 people from over 100 countries participated, submitting photos of themselves to be judged.[66] This contest was promoted as the first of its kind because it was judged by a set of three algorithmic decision-making systems, rather than by

[64] Fredrickson, O., "Risk Assessment Algorithms in the New Zealand Criminal Justice System", *New Zealand Law Journal*, Vol. 328, No. 330, 2020, p. 6.

[65] Pearson, J., "Why an AI-Judged Beauty Contest Picked Nearly all White Winners", *VICE*, 5 September 2016, available at: https://www.vice.com/en/article/78k7de/why-an-ai-judged-beauty-contest-picked-nearly-all-white-winners.

[66] Levin, S., "A Beauty Contest was Judged by AI and the Robots didn't like Dark Skin", *The Guardian*, 8 September 2016, available at: https://www.theguardian.com/technology/2016/sep/08/artificial-intelligence-beauty-contest-doesnt-like-black-people.

human judges.[67] The algorithmic systems purportedly assessed the contestants based on their facial symmetry, how youthful their appearance was relative to their age, and several other factors.[68]

As reported in the Guardian newspaper, among the 44 contestants identified by the algorithmic systems as the most attractive, almost all were white: six winners were Asian and only one had visibly darker skin.[69] These outcomes were a direct result of the dataset on which these algorithms had been trained. Beauty.AI used a dataset of photos that were pre-labelled as attractive. The majority of these photos were of white, European individuals.[70] As a result of the use of this data set, the system "learnt" that facial features which are common to white people are more attractive than others.

While there is no evidence that the developers of this system created the algorithms to consider white skin and features "beautiful", the algorithm inferred this conclusion from the training data which it was provided with.[71] The under-representation of other races and ethnicities in the training data set led the algorithm to "learn" that lighter complexions and "white" facial features were more attractive than others, and thus discriminate against those from other racial groups.

8 UNITED STATES OF AMERICA: Voice Activated Tools. Discrimination on the basis of ethnicity.

Speech recognition tools — algorithmic systems which can identify and respond to words spoken aloud — such as **Amazon's Alexa, Apple's Siri** and **Google's Assistant** — are used by millions of people across the globe. According to Dr. Daniela Braga (founder and CEO of DefinedCrowd, a platform that collects training data for AI systems), research has shown that these tools are not equally accurate for all accents or languages. According to Dr Braga; in some cases, these systems are only able to reliably understand "white, non-immigrant, upper-middle-class Americans".[72]

In 2018, the Washington Post published the findings of a study, which it had conducted in collaboration with Globalme and Pulse Labs (two testing companies). The Post collaborated with the testing companies to design the study, which was then administered using "recently updated Amazon Echo Dot and Google Home devices with volunteers in the U.S. and Canada".[73] The study, which focused on the ability these two

[67] Pearson, J., "Why an AI-Judged Beauty Contest Picked Nearly all White Winners", *VICE*, 5 September 2016, available at: https://www.vice.com/en/article/78k7de/why-an-ai-judged-beauty-contest-picked-nearly-all-white-winners.

[68] *Ibid.*

[69] *Ibid.*

[70] *Ibid.*

[71] Levin, S., "A Beauty Contest was Judged by AI and the Robots didn't like Dark Skin", *The Guardian*, 8 September 2016, available at: https://www.theguardian.com/technology/2016/sep/08/artificial-intelligence-beauty-contest-doesnt-like-black-people.

[72] DefinedCrowd Corp., "Mind the (Accent) Gap: DefinedCrowd Contributing to More Inclusive Speech Technology", *PR Newswire*, 29 July 2021, available at: https://www.prnewswire.com/news-releases/mind-the-accent-gap-definedcrowd-contributing-to-more-inclusive-speech-technology-301344593.html.

[73] Harwell, D., "Why Some Accents Don't Work on Alexa or Google Home", *The Washington Post*, 19 July 2018, available at: https://www.washingtonpost.com/graphics/2018/business/alexa-does-not-understand-your-accent/.

smart speaker systems to recognise different accents, found that there was a notable disparity among those whom the systems could understand.[74] For example, the study found that when a person with a typical mid-West American accent asked Alexa a question, the smart speaker understood the speaker well, while when someone with a Spanish accent asked the same question, it struggled to understand.[75] More specifically, people whose first language was Spanish were understood less than those who grew up near California or Washington states, where many tech companies are based.[76]

The inaccuracy in understanding by voice recognition systems is attributed by the study authors to the fact that the systems are not exposed to diverse voices during the training phase: "too many of the people training, testing and working with the systems all sound the same".[77] Under-representation of persons from certain ethnic groups within the training dataset results in a system which performs differently for different service users, depending on their ethnicity. To address this problem, some have suggested using more training data and speech datasets that include "Spanish-accented English data from the Americas".[78]

This phenomenon is particularly problematic in the United States of America, where more than 35 million people speak languages other than English as their first language, with about 60% of this group speaking Spanish at home.[79] The result is that persons from minority ethnic communities who do not speak English as their first language are disproportionately exposed to a system which does not function as intended – or indeed as presented and marketed.

Other experts have suggested that the performance problems of Alexa, Siri and Assistant could also arise for persons with disabilities which affect their speech, such as cerebral palsy and amyotrophic lateral sclerosis.[80] While speech recognition tools have the potential to act as assistive devices for persons with certain forms of disability, thus removing barriers, accommodating difference and increasing accessibility, without care in the selection and use of data to train these systems, persons with certain disabilities will experience a poorer quality service than other users.

---

[74] *Ibid.*

[75] *Ibid.*

[76] *Ibid.*

[77] Harwell, D., "Why Some Accents Don't Work on Alexa or Google Home", *The Washington Post*, 19 July 2018, available at: https://www.washingtonpost.com/graphics/2018/business/alexa-does-not-understand-your-accent/.

[78] DefinedCrowd Corp., "Mind the (Accent) Gap: DefinedCrowd Contributing to More Inclusive Speech Technology", *PR Newswire*, 29 July 2021, available at: https://www.prnewswire.com/news-releases/mind-the-accent-gap-definedcrowd-contributing-to-more-inclusive-speech-technology-301344593.html.

[79] DefinedCrowd Corp., "Mind the (Accent) Gap: DefinedCrowd Contributing to More Inclusive Speech Technology", *PR Newswire*, 29 July 2021, available at: https://www.prnewswire.com/news-releases/mind-the-accent-gap-definedcrowd-contributing-to-more-inclusive-speech-technology-301344593.html.

[80] Rangnekar, P., "Decoding the "Encoding" of Ableism in Technology and Artificial Intelligence", *ScienceConnect,* 26 July 2021, available at: https://www.science-connect.com/post/decoding-the-encoding-of-ableism-in-technology-and-artificial-intelligence; see also AI Now Institute, *Disability, Bias, and AI*, November 2019, available at: https://ainowinstitute.org/publication/disabilitybiasai-2019, p. 14.

9 UNITED STATES OF AMERICA: Education. Discrimination on the basis of race and gender.

In 2013, the computer science department of the University of Texas at Austin started using a machine-learning system known as **GRADE** to evaluate the applicants to the department's PhD program. GRADE — short for GRaduate ADmissions Evaluator — was developed by a faculty member and a graduate student in computer science at the University; the system had been created to help the department's graduate admissions committee save time.[81] The system predicted the likelihood of the committee approving an applicant and expressed that likelihood as a numerical score.[82] In 2020, the department discontinued the use of the system.[83] In an official tweet, the department admitted that its decision to stop using the system was based on the fact that the system could be biased.[84]

The developers of the GRADE system trained it using a database of previous admissions decisions; while making predictions, the system identified and then used patterns from this database.[85] GRADE ranked applicants using several criteria, including: grade point average (GPA); universities previously attended; letters of recommendation; research interests; and preferred faculty advisors.[86] To predict whether an applicant was likely to be successful, the system compared the applicant's data with that of the students who had been previously accepted into the PhD program.[87] Using this method, the system filtered out the applicants who were not likely to be successful. This process of screening was meant to help the department focus on the more promising candidates out of the applicant pool.[88]

It is not clear how many students were adversely impacted by the use of the GRADE system. However, the fact that the system was designed to replicate the decisions of the admissions committee prior to 2013 raised serious concerns regarding the automation of bias in graduate admissions.[89] Certain groups, such as

---

[81] Burke, L., "The Death and Life of an Admissions Algorithm", *Inside Higher Ed*, 13 December 2020, available at: https://www.insidehighered.com/admissions/article/2020/12/14/u-texas-will-stop-using-controversial-algorithm-evaluate-phd.

[82] *Ibid.*

[83] *Ibid.*

[84] For the link to the official tweet, see Burke, L., "The Death and Life of an Admissions Algorithm", *Inside Higher Ed*, 13 December 2020, available at: https://www.insidehighered.com/admissions/article/2020/12/14/u-texas-will-stop-using-controversial-algorithm-evaluate-phd.

[85] Burke, L., "The Death and Life of an Admissions Algorithm", *Inside Higher Ed*, 13 December 2020, available at: https://www.insidehighered.com/admissions/article/2020/12/14/u-texas-will-stop-using-controversial-algorithm-evaluate-phd.

[86] Quach, K., "Uni Revealed it Killed off its PhD-applicant screening AI — Just as its Inventors Gave a Lecture about the Tech", *The Register*, 8 December 2020, available at: https://www.theregister.com/2020/12/08/texas_compsci_phd_ai/.

[87] *Ibid.*

[88] *Ibid.*

[89] Burke, L., "The Death and Life of an Admissions Algorithm", *Inside Higher Ed*, 13 December 2020, available at: https://www.insidehighered.com/admissions/article/2020/12/14/u-texas-will-stop-using-controversial-algorithm-evaluate-phd.

women, Black, and Latinx, are underrepresented in the field of computer science; in 2017, for example, almost 80% of undergraduates in computer science at the University of Texas were men.[90] The 2014 Taulbee Survey, conducted annually by the Computing Research Association, found that African-Americans constituted only approximately 2.4% of Master's degree recipients and 1.5% of PhD degree recipients in computer science and related fields in the US and Canada.[91] Accordingly, a system which looks at the characteristics of past admissions is likely to replicate this pattern, reinforcing patterns of discrimination and inequality,

D LEARNING

Algorithmic systems do not only learn during the training stage; systems continue to learn after deployment, through use. Learning can take place through the environment and in particular through interaction with users. For example, a system might learn how to respond to certain inputs based upon the way that human users respond to the same information.

In common with the discriminatory patterns which occur during the design and training stages, discrimination risks frequently arise due to the exposure of algorithmic systems to unrepresentative data. The following case studies illustrate how the learning stage of an algorithmic systems can lead to discrimination, with examples of such discrimination arising on the basis of gender, sexual orientation, and disability.

10 GLOBAL: Advertising. Discrimination on the basis of sex and gender.

Google uses an advertising algorithm is to target advertisements to users, in order to optimise engagement and uptake. A 2014 study by researchers at Carnegie Mellon University and the International Computer Science Institute at Berkeley investigated the Google advertising algorithm, examining whether and how the system presented advertisements differently to men and women.[92]

The Google algorithm used a variety of factors to assess users and target advertising. One of the factors used was data on the ways in which different users engaged with the system itself. For example, if users in a certain age group, or in a particular location, were more likely to engage with certain websites and adverts, then the algorithm would show similar adverts to users in these age groups or locations.

The researchers who conducted the 2014 study built an automated tool called 'AdFisher', which presented as if it were male and female job seekers. They then used the tool to test how the Google advertisement algorithm determined the types of adverts it displayed to different users.[93] The study found that male job seekers were targeted with advertisements for high-paying jobs approximately six times more frequently than the female

---

[90] *Ibid.*

[91] Zweben, S., Bizot, B., "2014 Taulbee Survey", *Computing Research News*, Vol. 25, No. 5, May 2015, pp. 4, 15.

[92] Datta, A., Tschantz, M.C., Datta, A., "Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination", *arXiv:* 1408.6491, 2014, available at: https://arxiv.org/abs/1408.6491, p. 13.

[93] Gibbs, S., "Women Less Likely to be Shown Ads for High-Paid Jobs on Google", *The Guardian*, 8 July 2015, available at: https://www.theguardian.com/technology/2015/jul/08/women-less-likely-ads-high-paid-jobs-google-study.

job seekers. The study also found that the system was more likely to show advertisements for coaching services for high-paying jobs (executive positions with salaries at $200k+) to men: Google showed such advertisements 1,852 times to men and just 318 times to women.[94] This targeting occurred despite the job seekers' search history, online behaviour and other factors being identical or similar.[95]

The research team found that the algorithm had learnt from the behaviour of its users that men are more likely to conduct searches for executive positions and higher paying jobs. Given the design of the algorithm and its reliance on user data, with a higher number of men engaging with advertisements for high-paying jobs, the algorithm learnt to display such advertisements more frequently to men.[96] In so doing, the advertising algorithm reinforced stereotypical social norms, structural discrimination and inequalities within the world of work.

11 CANADA: Transport. Discrimination on the basis of disability.

In 2015, Jutta Treviranus – an academic working on a project for the Ontario Ministry of Transportation – studied an artificial intelligence system designed to guide **automated vehicles** (driverless cars) through intersections. The system would instruct the vehicle to "proceed, stop or adjust to avoid obstacles".[97] Treviranus tested the system to see what it would do when it encountered people who moved in unexpected ways. With the help of a friend – a wheelchair user – she identified a fundamental flaw in the system:

> *I brought a capture of a friend of mine that propels herself backwards in her wheelchair. She has strong legs, but her movements are poorly controlled and while she can't stand, she can move very fast by pushing her wheelchair with her legs and feet. […] When I presented a capture of my friend to the learning models, they all chose to run her over. I was told to try again once the learning models had been exposed to more data about people using wheelchairs in intersections. I was told that the learning models were immature models that were not yet smart enough to recognize people in wheelchairs, they would expose them to learning data that included many people using wheelchairs in intersections. When I came back to test out the smarter models they ran her over with greater confidence. I can only presume that they decided, based on the average behavior of wheelchairs, that wheelchairs go in the opposite direction.*[98]

According to Treviranus, the explanation for the system's decision making lay in the way in which it interpreted and learnt from the data presented to it. The system in question based its decision on the behaviour of the

---

[94] *Ibid.*

[95] Datta, A., Tschantz, M.C., Datta, A., "Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination", *arXiv:* 1408.6491, 2014, available at: https://arxiv.org/abs/1408.6491, p. 13.

[96] Cossins, D., "Discriminating Algorithms: 5 Times AI Showed Prejudice", *New Scientist*, 12 April 2018, available at: https://www.newscientist.com/article/2166207-discriminating-algorithms-5-times-ai-showed-prejudice/.

[97] Treviranus, J., "Sidewalk Toronto and Why Smarter is Not Better*", *Medium*, 30 October 2018, available at: https://medium.datadriveninvestor.com/sidewalk-toronto-and-why-smarter-is-not-better-b233058d01c8.

[98] Treviranus, J., "Sidewalk Toronto and Why Smarter is Not Better*", *Medium*, 30 October 2018, available at: https://medium.datadriveninvestor.com/sidewalk-toronto-and-why-smarter-is-not-better-b233058d01c8.

majority of persons,[99] while failing to acknowledge difference or diversity. In essence, the system established "norms" and was not able to accommodate "outliers", irrespective of how much additional data was provided.[100] This kind of "learning" model poses particular discrimination risks for persons with disabilities, as Treviranus' study illustrates.

12 REPUBLIC OF KOREA: Chatbot. Hate speech, harassment and discriminatory language.

In December 2020, Scatter Lab, a technology company in the Republic of Korea, launched an AI chatbot called **Lee Luda**. The chatbot assumed the persona of a female university student who could interact with users through an existing messenger app.[101]

In the first weeks of its launch, Lee Luda attracted over 7.5 million users who were impressed by the bot's natural-seeming responses.[102] It had initially learnt these responses by analysing 10 billion actual conversations on a messaging app, KakaoTalk.[103] Once deployed, the chat bot continued to learn from the way in which users interacted with it.

Lee Luda quickly began to make homophobic, racist, and ableist remarks. In various instances, the bot claimed to hate lesbians, Black people and persons with disabilities.[104] One of the main drivers of Lee Luda's responses was the way in which users interacted with it. Lee Luda learnt homophobic, racist, and ableist remarks both from the training data on KakaoTalk but also from the responses that users provided during their chats. Similarly, Lee Luda learnt from users who intentionally took advantage of its learning model to manipulate the bot into making sexually offensive comments and otherwise unwanted remarks of a sexual nature.[105] Following numerous complaints, the bot was eventually taken down.

A similar situation occurred with Microsoft's 'Tay' chatbot which was designed to learn from its interactions with Twitter users but was discontinued after it learnt from users to make homophobic and racist

---

[99] AI Now Institute, "Disability, Bias, and AI — Report", 20 November 2019, available at: https://ainowinstitute.org/publication/disabilitybiasai-2019, p. 13.

[100] *Ibid.*

[101] Wakefield, L., "Lee Luda: AI Chatbot Pulled After it Started "Hating" Lesbians and Black People", *PinkNews*, 14 January 2021, available at: https://www.thepinknews.com/2021/01/14/lee-luda-ai-chatbot-facebook-messenger-lesbians-racism-homophobia-discrimination/.

[102] McCurry, J., "South Korean AI Chatbot Pulled from Facebook after Hate Speech towards Minorities", *The Guardian*, 14 January 2021, available at: https://www.theguardian.com/world/2021/jan/14/time-to-properly-socialise-hate-speech-ai-chatbot-pulled-from-facebook.

[103] Hyun-ju, O., "Civic Groups File Petition over Human Rights Violations by Chatbot Luda", *The Korea Herald*, 3 February 2021, available at: http://www.koreaherald.com/view.php?ud=20210203001136.

[104] Wakefield, L., "Lee Luda: AI Chatbot Pulled After it Started "Hating" Lesbians and Black People", *PinkNews* 14 January 2021, available at: https://www.thepinknews.com/2021/01/14/lee-luda-ai-chatbot-facebook-messenger-lesbians-racism-homophobia-discrimination/; Kim, D., "Chatbot Gone Awry Starts Conversations about AI Ethics in South Korea", *The Diplomat*, 16 January 2021, available at: https://thediplomat.com/2021/01/chatbot-gone-awry-starts-conversations-about-ai-ethics-in-south-korea/.

[105] *Ibid.*

comments.[106] In both of these cases, system developers failed to anticipate the possibility that the chatbot would learn harmful, stereotypical and prejudiced language from users. Both cases demonstrate how algorithmic systems can "learn" discriminatory behaviour during use.

E DEPLOYMENT AND USE

Once an algorithmic system has been developed, it will become available for deployment and use. As the cases above illustrate, the majority of discriminatory impacts of these systems become evident only at this stage, as this is when the systems begin to interact with human users. Nevertheless, as set out above, many of the discriminatory impacts are "built into" the systems at an earlier point in the inception, design, training or learning stages of the system.

With this said, there are also examples of algorithmic discrimination which arise because of the way in which the system is deployed and used. In this case, the discriminatory behaviour of the system is not necessarily intrinsic to the system design or operation but is a contingent effect of the way in which it is used. The potential for discriminatory impacts to arise at this stage in the process depends upon the deployment context: an algorithmic system that works well in one context might result in a discriminatory impact when used in another. The following three case studies exemplify some of the ways in which an algorithm can create and replicate existing patterns of discrimination if they are deployed in certain contexts or used in certain ways.

13 GLOBAL: Content Moderation. Discriminatory denial of freedom of expression

Social media platforms such as **Twitter** and **Facebook** frequently use algorithmic systems in their content moderation programmes to identify potential hate speech by analysing the language used in posts.

However, when moderating content online, what is considered offensive depends greatly on social context, something that these algorithms may be unable to account for.[107] For example, algorithms may flag terms such as "queer" or "gay" as potential hate speech when they are being used in ways that affirm the individual or group identity of LGBT+ persons.[108] This is particularly problematic where words have been reclaimed by marginalised communities.[109] The practical effect is that the legitimate speech of

[106] Vincent, J., "Twitter Taught Microsoft's AI Chatbot to be a Racist Asshole in Less than a Day", *The Verge*, 24 March 2016, available at: https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist.

[107] Ghaffary, S., "The Algorithms that Detect Hate Speech Online are Biased against Black People", *Vox*, 15 August 2019, https://www.vox.com/recode/2019/8/15/20806384/social-media-hate-speech-bias-black-african-american-facebook-twitter.

[108] Kennedy, B., Jin, X., Davani, A.M., Dehghani, M., Ren, X., "Contextualizing Hate Speech Classifiers with Post-Hoc Explanation", *arXiv*, 2005.02439, 2020, available at: http://arxiv.org/abs/2005.02439.

[109] See, for example, Rahman, J., "The N Word: Its History and Use in the African American Community", *Journal of English Linguistics*, Vol. 40, No. 2, 2011.

certain groups is more likely to be censured, with indirectly discriminatory impacts.[110] According to one study, tweets by "African American authors are 1.5 times more likely to be flagged as being 'offensive.'"[111]

The way in which algorithms are currently designed, developed, and deployed means that it can be difficult for such systems to appreciate, understand and adapt for context. Yet this context is critical: language can both inform an individual's experience of prejudice and provide a means to overcoming and addressing it.

14 INDIA: Facial Recognition Technology. Discrimination on the basis of religion.

Facial recognition technology is a means of identifying people by mapping a person's facial features from a photograph or video and then comparing the mapped information with a catalogue of previously mapped faces to identify a match.[112] The technology is being used increasingly by law enforcement agencies in many parts of the world, despite the fact that it has been found to misidentify people on a regular basis.[113] In New Delhi, the capital of India, for example, the police force uses data from closed-circuit television (CCTV) cameras installed across the city to identify and track individuals.[114]

New Delhi's Muslim population is over-policed: there are more police stations in Muslim-dominated areas, meaning such areas have a larger police presence and more CCTV cameras than other areas.[115] According to the findings of a study published by the Vidhi Centre for Legal Policy in 2021, since Muslim-dominated areas are already targeted by law enforcement, the use of facial recognition technology means that Muslim-dominated areas are likely to be further "over-surveilled, over-policed and thus subject to more errors".[116]

As the accuracy of the facial recognition technology used by the police is reportedly just 2%,[117] the use of the algorithmic system replicates, reinforces and exacerbates existing patterns of systemic discrimination and inequality. This both arises from – and further contributes to – over-policing. This is because when technology that is prone to errors is applied disproportionately to a certain group of people (in this case,

---

[110] *Ibid.*

[111] Sap M., et al, "The Risk of Racial Bias in Hate Speech Detection", *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, p. 1671.

[112] For a thorough explanation of how facial recognition technology works, see Introna, L., Nissenbaum, H., "Facial Recognition technology: A Survey of Policy and Implementation Issues", Working Paper, *Lancaster University Management School*, 2010, p. 11.

[113] Fung, B., Metz, R., "This May Be America's First Known Wrongful Arrest Involving Facial Recognition", *CNN*, 24 June 2020, available at: https://www.cnn.com/2020/06/24/tech/aclu-mistaken-facial-recognition/index.html.

[114] Vipra, J., "The Use of Facial Recognition Technology for Policing in Delhi: An Empirical Study of potential Discrimination", Working Paper, *Vidhi Centre for Legal Policy*, 2021, p.7.

[115] *Ibid.* p. 14.

[116] *Ibid*. p. 6.

[117] Business Standard, "Delhi Police Facial Recognition Software Has Only 2 Per Cent Accuracy: HC Told", 23 August 2018, available at: https://www.business-standard.com/article/pti-stories/delhi-police-facial-recognition-software-has-only-2-per-cent-accuracy-hc-told-118082301289_1.html.

the Muslims in New Delhi), the members of this group experience misidentification on a correspondingly higher level.[118]

While the failure rate of the system may be uniform across the population as a whole, the deployment of the facial recognition in one geographical area, with a large population of a minority or marginalised religious or ethnic group leads to a disproportionate impact of the failure on this group.

15 UNITED STATES OF AMERICA: Facial Recognition Technology. Discrimination on the basis of race.

Facial recognition technology has also been deployed by police forces in the United States of America. In 2020, police officers in Detroit arrested Robert Julian-Borchak Williams (a Black man) because a facial recognition algorithm incorrectly matched his driver's license photo with that of a suspected thief.[119] The incident is believed to be the first known wrongful arrest on account of the use of facial recognition technology in the country.[120] In an administrative complaint filed with the police department, the American Civil Liberties Union alleged that the police had relied on "flawed and racist facial recognition technology" to arrest Williams.[121]

Researchers attribute the lack of accuracy in facial recognition technology to the "demographic imbalances in the training data" used to train the system.[122] Since the systems are trained mainly using images of white faces, the algorithms will be more accurate in matching faces from this group, when compared to others.[123] This means that the algorithms have "different accuracy rates for different demographic groups".[124] A 2019 study on the accuracy of facial recognition systems, published by the National Institute of Standards and Technology, found that the rate of false positives — meaning the "incorrect association of two subjects" — is the "highest in West and East African and East Asian people, and lowest in East European individuals".[125]

---

[118] Vipra, J., "The Use of Facial Recognition Technology for Policing in Delhi: An Empirical Study of potential Discrimination", Working Paper, *Vidhi Centre for Legal Policy*, 2021, p.13.

[119] Allyn, B., "The Computer Got It Wrong': How Facial Recognition Led To False Arrest Of Black Man", *NPR*, 24 June 2020, available at: https://www.npr.org/2020/06/24/882683463/the-computer-got-it-wrong-how-facial-recognition-led-to-a-false-arrest-in-michig.

[120] Fung, B., Metz, R., "This May Be America's First Known Wrongful Arrest Involving Facial Recognition", *CNN*, 24 June 2020, available at: https://www.cnn.com/2020/06/24/tech/aclu-mistaken-facial-recognition/index.html.

[121] American Civil Liberties Union, "ACLU of Michigan Complaint re Use of Facial Recognition", 2020, available at: https://www.aclu.org/sites/default/files/field_document/dpd_complaint_v_final.pdf.

[122] Bruveris, M., Mortazavian, P., Gietema, J., Mahadevan, M., "Reducing Geographic Performance Differentials for Face Recognition", *2020 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, 2020, p. 98.

[123] Musa, S., "Using Machine Learning to Overcome Facial Recognition Bias in Africa." *Global Scientific Journal*, Vol. 10, No. 11, 2022, p. 2.

[124] Crumpler, W., "The Problem of Bias in Facial Recognition", *Center for Strategic and International Studies*, 1 May 2020, available at: https://www.csis.org/blogs/strategic-technologies-blog/problem-bias-facial-recognition.

[125] Grother, P., Ngan, M., Hanaoka K., "Face Recognition Vendor Test Part 3: Demographic Effects", *National Institute of Standards and Technology*, 2019, p. 2.

Here, the discriminatory impact of the algorithmic system stems from the lack of representation and diversity in the ethnicities on which it was trained.[126] However, the context in which it was deployed still plays an essential role in the algorithm's discriminatory impact. In this case, the algorithm was deployed in a system that has already been found to discriminate against certain groups, particularly "Africans and those of African descent".[127] This was neither identified nor addressed by the developers of the algorithm in their training data nor by those who adopted and deployed the system. As a result, the system replicated and exacerbated existing patterns of discrimination within the justice system.

[126] See the discussion on training algorithms in Chapter 2(C) of this report.

[127] United Nations High Commissioner for Human Rights, "Promotion and protection of the human rights and fundamental freedoms of Africans and of people of African descent against excessive use of force and other human rights violations by law enforcement officers through transformative change for racial justice and equality", 2022, A/HRC/51/53, paras. 19 and 26; Minnesota Department of Human Rights, "Investigation into the City of Minneapolis and the Minneapolis Police Department", 2022, available at: https://mn.gov/mdhr/mpd/findings/.

CONCLUSIONS

Algorithmic decision-making systems can provide innovative and effective solutions to a wide array of different problems and challenges in business, government and society more broadly. As the case studies examined here demonstrate, algorithmic systems have been developed to address problems and meet societal needs in areas such as education, social security and healthcare, as well as to provide goods, services and experiences in response to consumer demand. Indeed, there is significant – largely unexploited – potential for algorithmic systems to be developed and deployed in ways which advance equality, through removing barriers, eliminating bias and increasing equality of access, opportunity and participation.

Yet, as these case studies illustrate, in practice, the use of these systems frequently result in discriminatory impacts. In some of these cases, these impacts are foreseeable and may indeed be the result of deliberate policy decisions: the system may have a discriminatory impact because of a choice to treat a particular group sharing an identity, status or belief differently. Frequently, however, these impacts are unintended or unforeseen — the result of those involved in the development of the systems not understanding patterns of systemic inequality and the dynamics of discrimination and so not appreciating how the use of certain data or the adoption of particular decision-making protocols will result in discrimination.

This small selection of case studies from around the world demonstrate that algorithmic systems can and do result in discriminatory impacts on any ground of discrimination and in all sectors and areas of life. Indeed, some of these cases illustrate how the use of these systems can result in novel patterns of discrimination, occurring on the basis of characteristics — or combinations of characteristics — which are not yet well-recognised in law, or in new and emerging sectors of the economy. Taken together, the cases show how algorithmic discrimination can arise at any point in the life cycle of the technology, on the basis of any ground, in any area of life and in any part of the world.

More broadly, the case studies show that, because of the way in which algorithmic systems are developed and designed, trained and evaluated, deployed and used, they are frequently discriminatory by default. Through reliance on stereotypical assumptions, the use of data which is not representative or which reflects existing patterns of inequality and disadvantage, the, exposure of systems to the prejudice of human users, and a host of other factors, discriminatory behaviours are frequently built into these systems, sometimes deliberately, more frequently as a result of ignorance.

While the discriminatory impacts of these systems may be unintended, unforeseen, or challenging to identify or understand, this does not limit the obligations of States and responsibilities of business to prevent discrimination.

States and businesses have obligations to prevent the discriminatory impacts of algorithmic systems. International law requires that States do not discriminate in law, policy and practice and that they take effective measures to prevent discrimination by business and other actors, including through the adoption of comprehensive anti-discrimination legislation. Even where a State falls short of these duties, businesses have responsibilities to respect the right to non-discrimination which exists independently of States' obligations. All of these duties apply to the use of algorithmic systems.

These case studies demonstrate that States and businesses are, all too often, failing to meet these obligations in the regulation, development, and use of algorithmic systems. In order to fulfil these obligations, it is essential that they adopt a new, pre-emptive and precautionary approach.

States and businesses will only be able to meet their obligations to eliminate discrimination and advance equality if they take a proactive approach to identifying, assessing and addressing the equality impacts of algorithmic systems during their development. In light of the evidence that these systems are frequently discriminatory by default, it is essential that States and businesses adopt a proactive, intentional **equality by design** approach to the development, use and regulation of algorithmic decision making systems.