HM Government

The Fairness Innovation Challenge: Key Findings

December 2025

Acknowledgements

This initiative was delivered in partnership with Innovate UK and with support from the Information Commissioners Office and the Equality and Human Rights Commission. We would like to extend our thanks for their support and contributions.

We also wish to acknowledge the project leads at the Alan Turning Institute, Coefficent Ltd, Open University and King's College London who generously shared their reflections and findings for this report.

Contents

The Fairness Innovation Challenge ____ 4

Open University - Promoting equity in higher education: A socio-technical approach to AI Fairness ____ 5

Alan Turing Institute - Proactive Monitoring of AI Fairness: Development of Design Patterns and a Code Analysis Tool ____ 6

Coefficient Systems Ltd - Inclusive Innovations: Transforming CV Screening for Equity ____ 7

King's College London - Assessing the feasibility of a neurosymbolic methodology for bias mitigation of cardiac arrest early warning in general wards ____ 8

Key Findings ____ 9

Data Access ____ 9

Detection ____ 10

Mitigation ____ 11

Lessons from Regulators ____ 12

Equality and Human Rights Commission ____ 12

Project Reflections ____ 12

Next Steps ____ 12

Information Commissioner's Office ____ 12

Project Reflections ____ 13

Next steps ____ 14

Conclusion ____ 14

The Fairness Innovation Challenge

In 2023, DSIT launched the Fairness Innovation Challenge in collaboration with Innovate UK and with support from UK regulators, the Information Commissioner's Office and the Equality and Human Rights Commission. This was a grant challenge that provided over £465,000 funding to drive the development of novel solutions to address bias and discrimination in AI systems.

Despite increased interest in the issue of bias and discrimination in Artificial Intelligence (AI) systems, organisations continue to face numerous challenges when trying to tackle these issues in practice:

- Demographic data is required to detect bias in AI systems, but organisations face practical, ethical and regulatory challenges when seeking to collect this data themselves.
- Determining whether AI systems are generating unfair outcomes requires organisations to define what fair outcomes look like, which can be challenging, and they lack guidance on how to select appropriate fairness metrics.
- Purely technical approaches to mitigate unfair bias face limitations and may risk breaching UK equalities law.

Applicants were asked to focus on a real-world AI use case and to adopt a socio-technical approach to addressing bias and discrimination, which seeks to address both statistical and structural biases associated with the use of AI systems.

Four projects were funded:

- **Higher Education:** The Open University explored ways to improve the fairness of AI systems in higher education.
- **Financial Services:** The Alan Turing Institute created a fairness toolkit for SMEs and developers to self-assess and monitor fairness in Large Language Models (LLMs) used in the financial sector.
- **Healthcare**: King's College London designed a neuro-symbolic approach to address bias and discrimination in early warning systems used to predict cardiac arrest in hospital wards, based on the CogStack Foresight model.
- **Recruitment**: Coefficient Systems Ltd.'s project focused on identifying bias in automated CV screening algorithms that are often used in the recruitment sector.

The Challenge concluded in March 2025. This report details key findings from across the four projects.

Open University - Promoting equity in higher education: A socio-technical approach to AI Fairness

UK universities are increasingly turning to AI applications to meet regulatory targets, improve retention, and enhance the student experience. Yet, as AI systems become embedded in Higher Education (HE), the question of fairness, both technical and social, remains underexplored. The FairAI4EDTech project addresses this urgent gap, advancing a socio-technical approach to AI fairness in HE, with a particular focus on tackling bias and discrimination in AI-driven learning analytics (LA) systems.

Building on the award-winning OUAnalyse[1] (OUA) system at The Open University which supports over 200,000 students through predictive analytics and tutor alerts, the project team developed a Framework of Responsible AI for Learning Analytics. This framework aims to provide a much-needed roadmap for UK HE institutions as they develop their digital strategies.

The first phase of the project rigorously investigated algorithmic bias: how to detect it, measure it, and mitigate it across diverse student cohorts. The team evaluated different fairness definitions and metrics and explored their application in a live production system. Findings show that responsible use of predictive models in education requires moving beyond traditional performance metrics toward a structured, context-sensitive approach to fairness. This means;

- establishing clear institutional values around equity,
- selecting fairness metrics aligned with those values
- ensuring fairness is monitored continuously, not assessed only once, and
- recognising that fairness is dynamic, varying across time and across different student groups.

When embedded in practice, fairness metrics can become a powerful instrument for closing awarding gaps, strengthening student support, and building trust in the use of AI systems.

The second phase investigated how human tutors interact with AI-driven predictions. Focus groups revealed that while tutors generally trusted the system's alerts, training, and their reflective professional judgement acted as an important safeguard against uncritical reliance on predictions. These findings underline the need for HE institutions to embed AI tools within a framework of ethical awareness, robust training, transparency, and professional autonomy.

To increase transparency and student agency, the team piloted a student-facing version of the OUA dashboard. This innovation emphasises giving students back their data and supporting them to act on it. Two key lessons emerged: students benefit from insights into their study patterns and LA systems that present data alone are insufficient; students need clear, prescriptive recommendations to translate insights into action. Moreover, dual access to predictive insights, by both tutors and students, also acts as a safeguard, reducing the risk of bias in how interventions are initiated and experienced.

The lessons learned from FairAI4EDTech have been consolidated into publicly available Framework for Learning Analytics in Higher Education.[2]

[1] https://analyse.kmi.open.ac.uk
[2] https://fairai4edtech.kmi.open.ac.uk

“Through FairAI4EDTech, we have learned that achieving fairness in AI for HE requires more than just technical fixes; it demands a socio-technical approach that brings together robust algorithms, institutional values, and reflective judgment. Our experience with OUAnalyse has shown the power of AI solutions to support students, but also the importance of continuous monitoring, transparency, and shared responsibility between institutions, tutors, and learners. Looking ahead, our priority is to engage with the wider education community and help shape the national conversation on Responsible AI in HE.”

Alan Turing Institute - Proactive Monitoring of AI Fairness: Development of Design Patterns and a Code Analysis Tool

In the financial services sector, AI is increasingly being used by investors and institutions to analyse the emotions and opinions expressed in news, reports or social media to inform their decision-making. One such example is FinBERT, a pretrained NLP model used to analyse the sentiment of financial text. The introduction or amplification of biases in systems of this kind could lead to financial decision-making that is skewed against certain groups in favour of others, contributing towards economic inequity across society.

The Alan Turing Institute team set out to evaluate the performance of FinBERT across different demographic groups. Throughout these experiments, the team also integrated the experiment methodologies and created a fairness monitoring tool to enable similar proactive evaluations in various AI development tasks.

In the experiments, they first generated a synthetic dataset made up of thousands of country-specific financial statements that explicitly reveal sources of bias. The synthetic dataset allowed the team to have controlled experiments to observe model behaviour when a text explicitly demonstrates biased statements. In the second stage, they used a real-world news dataset to explore how the model performed when the bias displays itself implicitly. The team created binary-sensitive characteristics based on financial definitions- in this case, Global North and Global South- and used several different metrics to assess the model’s performance across these demographics, including demographic parity, equalised odds, and equal opportunity.

Although the team found that FinBERT did not exhibit significant group disparity between Global South and Global North samples, they observed inconsistencies in the model’s performance on the real-world news datasets. They also found that adding a strong positive statement or incorporating large numerical figures could alter the model’s predictions, vulnerabilities that could be used to mislead financial market participants.

To enable a proactive approach in fairness evaluation and monitoring, the team developed their FAID (Fair AI Development) tool. The FAID tool is intended to support better fairness-related metadata management throughout the ML pipeline to improve information sharing among stakeholders involved in the development process. It provides a set of functions and templates to support development teams to share fairness-related metadata, risks, and transparency reports, encouraging all stakeholders involved in the design process to feel the responsibility to share design decisions with one another.

Alongside this, the team developed a number of design patterns encapsulating best practices in AI fairness, including traceability, responsibility, explainability, auditability and digestibility. These patterns are reusable solutions to common fairness-related business challenges to help

developers follow best practices. The patterns aim to bridge the gap between a high-level process for developing fair AI systems and the more granular, technical approaches to evaluation and mitigation that this requires. These are available alongside other project resources.[3]

> *"In our research, we addressed a critical challenge towards achieving trusted AI: the subtle but significant ways in which models can exhibit vulnerabilities even when biases are not immediately apparent. The inconsistencies we observed in model performance, such as susceptibility to manipulation from positive statements or numerical figures, underscore the need for a proactive and holistic approach to fairness. Identifying fairness issues only after deployment is often difficult and costly, which makes early traceability and auditability all the more essential to ensuring AI systems are not only fair in theory but also robust and trustworthy in practice. We have already started seeing early uptake of these approaches. We are proud to be part of this excellent cohort, supporting DSIT in advancing this important mission, and will further commit to promoting proactive approaches in trustworthy AI."*

Coefficient Systems Ltd - Inclusive Innovations: Transforming CV Screening for Equity

In the recruitment sector, AI is increasingly being used to support sourcing, screening, interviewing and selection of potential candidates. When used to screen CVs of job applicants, AI can help to drive efficiencies and reduce costs for firms. However, the use of AI systems that generate unfair outcomes risks disadvantaging individuals from certain demographic groups, potentially leading to discriminatory outcomes. Coefficient Systems Ltd worked with Bays Consulting Ltd to research the extent of this bias, develop tools to detect it, and create a pathway for commercial exploitation and knowledge dissemination.

A comprehensive market evaluation of AI-supported CV screening tools within Applicant Tracking Systems (ATS) was conducted. This report evaluated the capabilities, bias-mitigation features, and compliance of existing commercial tools with UK data protection and equality legislation. The investigation identified two primary functions of AI in CV screening: CV parsing (extracting information) and candidate shortlisting/ranking. The report concluded that while many tools claim to reduce human cognitive bias, they risk amplifying systemic biases if trained on historically prejudiced data. A significant challenge identified was the "black box" nature of many commercial systems, which prevents algorithmic scrutiny and makes it difficult to explain or challenge automated decisions.

The Coefficient-Bays Consulting project employed a unique socio-technical approach to identifying and remedying bias in AI-empowered recruitment by reviewing technological and human biases influencing recruitment decisions. First, the team synthetically generated 1268 demographically distinct names. They selected four roles they would use as test cases, including HR Business Partner, Senior Software Engineer, Retail Manager and Financial Analyst, and used an LLM to generate eight nearly identical resumes for each of the roles being tested, which differed only by assigned names. Using these synthetically generated CVs, the team prompted each LLM they were testing to rank the resumes for the most qualified candidate for each of the chosen roles, a process that was repeated 200 times per role for each LLM.

[3] https://github.com/alan-turing-institute/fairness-monitoring

The team found that the LLMs tested exhibited racial and gender bias when sorting CVs. This bias varied depending on the role being selected for; for example, one model favoured black candidates over other ethnicities for the HR Business Partner while disproportionately favouring Asian candidates for the Senior Software Engineer role. Similarly, this model tended to favour women over men for the HR Business Partner role, while favouring men over women for the Senior Software Engineer role. The team calculated average disparate impact ratios across different LLMs, enabling them to draw high-level comparisons about the relative fairness of different models when used for CV screening.

Building on this learning, the team developed 'Fairground[4],' an open-source technical toolkit to enable developers to test their automated CV software and applicant tracking systems for bias and discrimination. The intention is to now evolve this into a scalable Software-as-a-Service (SaaS) product, offering advanced monitoring, compliance, and auditing features to help organisations ensure their AI systems are fair, transparent, and compliant with UK law. The broader findings from this project have been shared publicly in the Fairness Tales[5] series.

"*This project is not just about providing a technological solution; it's about fostering an equitable recruitment environment that champions diversity and fairness where 86% of companies are incorporating AI into their hiring processes. Across the board, the UK's approach to algorithmic auditing is governmental rather than auditable, leaving much room for improvement and a pressing need for standardised oversight.*

We aim to set a new standard in the recruitment landscape, ensuring our AI tools serve to support unbiased decision-making. It's wonderful to have this opportunity to extend the knowledge we gained and build this into a solution that helps both companies involved grow." - Dr Sophie Carr, CEO of Bays Consulting

King's College London - Assessing the feasibility of a neurosymbolic methodology for bias mitigation of cardiac arrest early warning in general wards

AI is increasingly being used to better predict clinical outcomes in the healthcare sector. This includes the use of AI in early warning systems for in-hospital cardiac arrest. Bias in AI-driven early warning systems such as these can limit their effectiveness and perpetuate healthcare inequities, with serious and potentially life-threatening consequences for patients.

King's College London set out to identify and address bias in the CogStack Foresight model. This is a generative pretrained transformed model that can forecast diagnostic codes and any other standardised medical codes based on a source dataset. The model was being piloted for use in secondary care settings to predict cardiac arrest for in-hospital patients.

The King's College London team identified three types of bias present in the CogStack Foresight model:

- **Statistical bias**: After analysing the performance of the model on over 10,000 patient records, they found it performed worse for patients with shorter clinical documentation

[4] https://pypi.org/project/fairground/
[5] https://www.fairnesstales.com/

trails, including younger patients with fewer comorbidities and patients with shorter hospital stays (including emergency admissions).

- **Model bias**: The model had an over-reliance on coded data related to symptoms, observations, diagnoses and procedures, while neglecting quantitative clinical measurements such as laboratory values and vital signs.
- **Structural bias**: The hospital data used to train the model contained biases related to gender disparities in the identification of cardiac arrest.

They set out to improve the model's ability to be a fair and effective tool for early recognition of in-hospital cardiac arrest through the development of a neurosymbolic methodology for bias mitigation 'METHOD' (Modular Efficient Transformer for Healthcare Outcome Discovery).

METHOD first involved developing architectural innovations, including adaptive sliding window attention and u-net inspired skip connections, to address statistical and model bias. After implementing these innovations, the team found the model had stable performance for patients with varying lengths of documentation and maintained accuracy for patients with limited clinical history.

Next, the team sought to address structural bias by integrating clinical guidelines into the model. This was achieved by reformulating the relevant clinical guidelines into a knowledge graph and devising an interaction protocol between the graph and the METHOD neural architecture.

The team then held an evaluation workshop with clinicians. They found clinicians had an ~90% agreement rate with the improved model across test cases and confirmed the effectiveness of the solution in identifying bias within clinical decision pathways.

Key Findings

Data Access

Many approaches to detecting and mitigating bias in AI systems require access to demographic data about service users. This includes characteristics that are protected under the Equality Act 2010, such as age, sex, and race, as well as other socioeconomic attributes.

However, organisations building or deploying AI systems struggle to access the demographic data they need. As set out in DSIT's report - 'Enabling responsible access to demographic data to make AI systems fairer' (2023) – they face a range of legal, ethical and practical challenges, including concerns around public trust, navigating regulatory compliance, and data quality. In recent years, there has been growing interest in the potential of novel approaches to overcome some of these challenges.

Synthetic data refers to data created to mimic the properties and patterns of real-world data.[6] It can be useful in addressing bias in AI systems by eliminating the need to collect real-world demographic data for bias detection and can be used to rebalance training datasets to create fairer AI systems. Synthetic data can also help to address privacy concerns by creating

[6] https://www.gov.uk/government/publications/ai-insights/ai-insights-synthetic-data-html

artificial datasets that bear no correlation to known real-world datasets but still reflect target distributions.

Both Coefficient Systems Ltd and the Alan Turing Institute used synthetic data as part of the solutions they developed in the Fairness Innovation Challenge. Coefficient used Chat GPT-4 to create 768 synthetic CVs with near identical educational achievements and work experience, each of which was assigned a distinct name from one of four ethnic groups. This enabled them to test how nine different large-language models ranked these CVs for different roles, uncovering racial and gender bias across these models. The Alan Turing Institute also used an LLM to generate thousands of synthetic financial statements that explicitly reveal sources of bias to evaluate the performance of FinBERT.

These projects demonstrate the utility of novel data generation techniques like synthetic data in aiding bias detection and mitigation where real-world data is unavailable. Other solutions, such as data intermediaries and the use of proxies, could also provide useful. By supporting project teams to trial the use of synthetic data, the Fairness Innovation Challenge has contributed towards the emergence of a future ecosystem characterised by the use of a variety of data access solutions that best meet the needs of organisations evaluating their AI systems for bias and their users.

Detection

Bias in AI systems can manifest in multiple and complex ways. In order to detect it, organisations can assess the inputs and outputs of the system to determine if there is unfair bias in the training data, the model itself, or the outcome of a decision or classification made by the system.

There are numerous challenges associated with detecting bias in AI systems. Training data may exhibit biases due to historical inequities and the conscious and unconscious decisions made by those who collect and process this data, which can be pervasive and hard to address in practice. Detection requires the selection of fairness metrics, mathematical definitions that encode simplified notions of fairness. There are different kinds of fairness metrics and selecting the right ones can prove challenging, as they are relevant and applicable in different contexts. Human interpretation of the outputs of AI systems can introduce additional bias into the system, which cannot be detected through the use of solely technical approaches to bias detection.

King’s College London uncovered multiple kinds of bias in the Foresight model. This included structural bias in the training data due to gender disparities in the identification of cardiac arrest, which sees women disproportionately underrepresented in the training dataset.

The Open University undertook a thorough bias metric selection process to identify bias in their learning analytics model. This involved assessing the types and severity of errors in their predictive learning analytics model alongside considering the wider societal context in which the model is being deployed, including the awarding gap that sees worse educational outcomes for students from traditionally underrepresented groups. The team selected two metrics that best capture their main concerns around fairness- equal opportunity and statistical parity difference- which they used to identify varying levels of gender and disability bias across the modules to which the model is applied. The team also held focus groups to understand and help to address potential bias in tutors’ interpretation of the outputs of the learning analytics model.

The work conducted by project teams demonstrates the myriad ways in which bias can present itself in AI systems and the importance of efforts to detect bias across different components of an AI system. They also demonstrate the need to select context-specific fairness metrics and the level of consideration that is required to align technical approaches to bias detection with the complexity of societal notions of fairness. Notably, all projects detected some level of bias in the AI systems they evaluated, demonstrating the pervasiveness of this issue across different sectors.

Mitigation

Mitigation and continued monitoring of identified biases is critical to ensure the fairness of AI systems. Given the complex and myriad ways in which bias can manifest in AI systems, this can prove challenging. Different types of bias will require the use of different bias mitigation techniques. These may include technical interventions- such as rebalancing or reweighting input data or making modifications to the model and its outputs - or non-technical interventions, like ensuring diverse development and deployment teams and addressing human biases in the interpretation of model outputs. In addition, some mitigation techniques may introduce additional risks; for example, approaches like reweighing - which actively seek to favour disadvantaged groups - may be interpreted as positive discrimination.

Project teams adopted a diverse range of strategies and techniques to make their AI systems fairer. After detecting three distinct types of bias in the Foresight model, the Kings College London team reimagined the transformer architecture of the model through an approach they entitled METHOD (Modular Efficient Transformed for Healthcare Outcome Delivery), leading to improved model performance on patients with shorter paper trails. To mitigate historical gender bias in the training data, the team also developed a ‘neuro-symbolic framework’ by incorporating clinical guidelines into the METHOD neural architecture, helping to address historical gender biases in the training data. Alongside these technical innovations, the team validated this methodology through evaluation workshops with clinical experts, demonstrating a roughly 90% agreement rate between clinicians and the Foresight model.

The Open University adopted a different approach to making their system fairer. This included co-creating and piloting a student dashboard with course students, whose insights and feedback were collected via focus groups. This dashboard makes predictions visible to students, helping them become more aware of their study patterns and empowering them to take proactive measures in collaboration with tutors. Alongside this, the team used their findings to develop a Responsible AI Framework for Learning Analytics in Higher Education Institutions to support other higher education institutions to implement similar bias mitigation efforts in learning analytics.

Findings from across the projects demonstrate that effective strategies for improving the fairness of AI systems require a diverse, multi-pronged approach that utilises both technical and non-technical interventions to address differing forms of bias. They also highlight how carefully considered bias mitigation strategies are not necessarily in conflict with model performance and can lead both improved performance and fairness. Findings from the Fairness Innovation Challenge can therefore help to inform the effectiveness of future approaches to making AI systems fairer.

Lessons from Regulators

Equality and Human Rights Commission

Artificial intelligence (AI) was a priority in the EHRC's strategic plan for 2022 to 2025. It is widely recognised that, while AI has potential to benefit society, it also comes with risks of bias and discrimination as well as risks to human rights. The EHRC was keen to be involved in the Fairness Innovation Challenge, working alongside the Information Commissioner's Office, to support responsible innovation. We saw the challenge as an opportunity to develop our understanding of bias and discrimination in AI systems, to inform our approach to regulation and to educate a new audience of tech developers on equality and human rights.

Project Reflections

It was a privilege to work with the four Challenge teams and to see their careful and expert work to identify and mitigate bias in their systems. We explored the risk that bias, including human bias, could lead to unlawful discrimination under the Equality Act. We also focussed on Article 8 of the Human Rights Act which protects the right to respect for private and family life. We were keen to develop our understanding of the ways in which an AI system could be biased, how that presents differently depending on the use case and the system, and what that means in terms of outcomes - the potential impact on people.

Being clear on the aim and objective of the AI system is an important starting point for using AI and one we explored with the teams. Is there a clear problem or need to address, is it only an AI system that can meet that? This is an important fairness consideration and informs the justification for using personal and sensitive data.

Accessibility, explainability and transparency of information and ways of working are also key to human involvement and scrutiny of AI systems, to ensure they are working fairly. The importance of being able to explain things clearly, particularly for a non-technical audience, was looked at with all the teams throughout the course of the Challenge.

Next Steps

The EHRC's work on AI continues. As a strategic regulator with a broad remit our focus is on the areas where the risk with AI is greatest and where we are particularly well-equipped to make an impact. Transparency of AI usage and challenging where we feel there is an incompatibility with equality and or human rights law are key priorities for our work.

Information Commissioner's Office

AI is a priority area for the ICO: as a regulator we are aware of the potential for public benefit and the risk of individual detriment posed by AI. Our collaboration with EHRC on the DSIT Fairness Innovation Challenge directly supports and influences the compliant use of AI-enabled technologies.

Our message is that data protection compliance starts at the design stage, not at the privacy notice. The way a technology functions must be designed with privacy in mind so that adopters of these technologies can meet their obligations under data protection law. Designing in functionality to support privacy builds greater public trust in data-driven technologies, which benefits innovation and the UK's economic growth.

On this Challenge we wanted to hear firsthand from developers about how they would incorporate data protection by design into their innovations when designing AI for fairness. We wanted to support innovators to use data compliantly and confidently to produce fairer outcomes for individuals. We were particularly interested in why design decisions were made and communicating this reasoning to the public.

Project Reflections

The use of AI can polarise public opinions: it can be seen as a great opportunity to produce more accurate predictions or results, or it can be seen as a risk of compounding existing biases and inequities in society.

This Challenge was a wonderful opportunity to discuss in depth the decision-making processes being made by innovators to produce AI models of public benefit.

King's College London's project addressed how to produce accurate predictions in an environment where data is patchy and access to more training data is difficult. Their approach to this issue has relevance for other innovators seeking greater accuracy in AI, for clinicians wanting to check AI tools are fit for purpose, and for members of the public concerned about how AI works. Instead of gathering more data to solve the issue, King's College London made best use of the data they had by structuring how the layers of processing within their model communicated with one another. To address historic biases, which had resulted in underdiagnosis of women and younger patients at risk of in-hospital cardiac arrest (IHAC), they added clinical best practice guidance into their processing. Our work with King's College London stressed the importance of communicating not only what their model does but how and why to clinicians, patients, and members of the public as this is crucial for public trust and wider adoption of AI.

The processing of special category data requires an additional processing condition under data protection law – but that shouldn't be interpreted by innovators as a reason not to process this data if it is essential to the purpose of processing.

The Open University's work demonstrated how collecting and including special category data in Predictive Learning Analytics (PLA) can improve the learning outcomes for students with intersectional disadvantages. Our discussions focussed on role-based access to personal data to protect student privacy, as there are key differences between the data tutors and administrators need to see. Having these access controls gives students greater confidence in sharing special category data about themselves. This matters because the students were involved in the co-design of a dashboard, informed by PLA, so that they could see their own learning trajectories and take the recommended steps to improve their chances of success.

Bias in the training data for AI can affect a person's chances of recruitment, as explored by Coefficient's work on CV screening tools, or the growth of whole economies, as researched by the Alan Turing Institute. When processing people's data, at whatever scale AI tools are used, data protection by design is essential to protect the rights of individuals and benefit society as a whole.

Next steps

AI remains a priority area for the ICO due to its potential to deliver significant public benefits when deployed compliantly and responsibly. The ICO is continuing to work on a range of AI-related topics, including automated decision-making and other applications of AI.

Conclusion

The Fairness Innovation Challenge provided a unique opportunity to explore the practical challenges and potential solutions for addressing bias and discrimination in AI systems across a range of sectors. Through collaborative efforts across project teams and regulators, innovative approaches were successfully developed to enable access to demographic data, identify sources of unfairness, and mitigate biased outcomes. The findings demonstrate that fairer AI systems can be achieved by implementing novel socio-technical approaches.

The Challenge highlighted that bias can be a real issue in many AI systems currently in use, but also that meaningful progress is possible when technical innovation is combined with ethical and regulatory considerations. The lessons learned provide a strong foundation for future work to ensure AI systems are fair, trustworthy, and beneficial for all. They offer practical insights for organisations seeking to understand and improve fairness in AI, and should inform ongoing efforts to safeguard UK consumers and businesses while unlocking the opportunities AI presents.

This report is available from: www.gov.uk/government/organisations/department-for-science-innovation-and-technology

If you need a version of this document in a more accessible format, please email alt.formats@dsit.gov.uk. Please tell us what format you need. It will help us if you say what assistive technology you use.