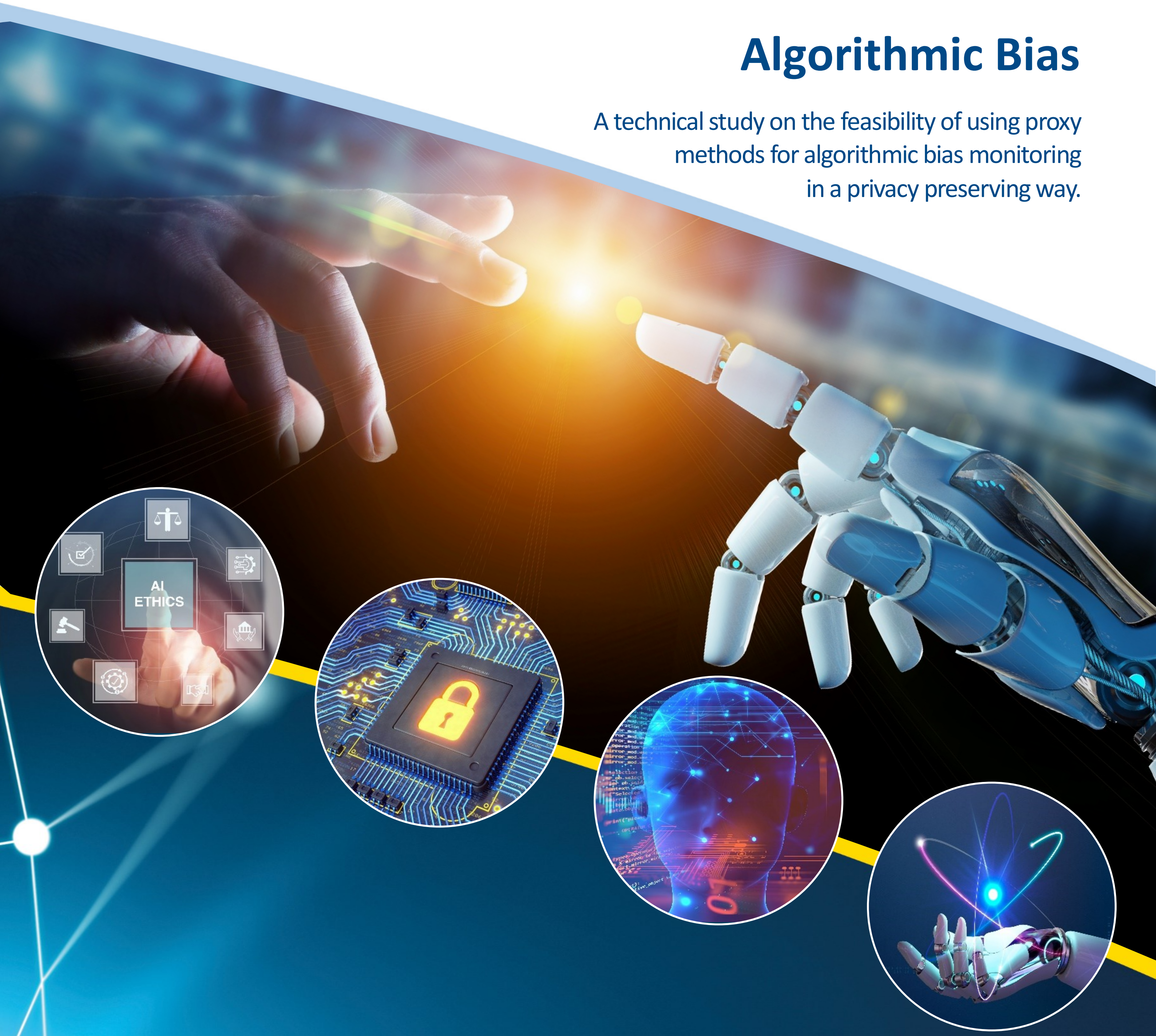
FRAZER-NASH
CONSULTANCY
A KBR COMPANY

Algorithmic Bias

A technical study on the feasibility of using proxy methods for algorithmic bias monitoring in a privacy preserving way.



March 2023

019178

54857R Issue: 1.0

Prepared for: Department for Digital, Culture, Media and Sport (DCMS)

SYSTEMS • ENGINEERING • TECHNOLOGY

Algorithmic Bias

A technical study on the feasibility of using proxy methods for algorithmic bias monitoring in a privacy preserving way

Client:	Department for Digital, Culture, Media and Sport (DCMS)
Client Ref.:	103344
Date:	March 2023
Classification:	OFFICIAL

Project No.:	019178	**Compiled By:**	Beckett LeClair, William Parker, Amanda Young
Document No.:	54857R	**Approved By:**	Gwen Palmer
Issue No.:	1.0	**Signed:**	

Distribution

Copy	Recipient	Organisation
1	James Scott	Department for Digital, Culture, Media & Sport (DCMS)
2	File	Frazer-Nash Consultancy Limited

DISCLAIMER

This document (including any annexes, appendices and associated computer files) is issued in confidence solely for the purpose for which it is supplied. Frazer-Nash Consultancy Limited accepts no liability for loss or damage suffered by any party resulting from use of this document, other than as provided for by the contract under which this document has been produced.

Originating Office: FRAZER-NASH CONSULTANCY LIMITED
Stonebridge House, Dorking Business Park, Dorking, Surrey, RH4 1HJ
Tel: +44 (0)1306 885050

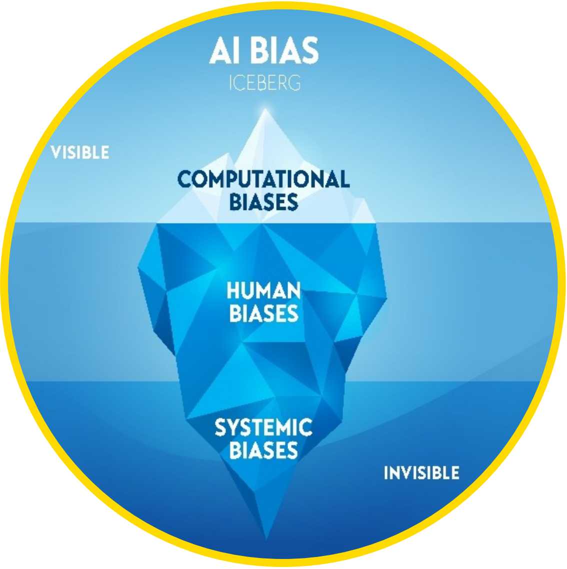
Executive Summary

It is becoming increasingly common for machine learning (ML) algorithms to be used for decision making, often behind the scenes and not regulated as clearly as human decision making. This creates a need for methods that can detect and hopefully mitigate bias in these systems. Data proxies have the potential to be used to infer important demographic data from available user information and this can be used to check for bias in otherwise opaque algorithms.

In this report, six methods have been identified that are assessed as likely to be viable for use in the UK as they have been trained on UK or international datasets. They cover a range of different proxies: **age, gender, and race/ethnicity** as well as maturity: **open-source, commercial tools, and academic papers**. A technical description has been provided for each method and a comparison of the performance of each method carried out against important metrics. The trades-off between these metrics and their implications for inferring the feasibility of methods is provided, to aid in future comparative studies.

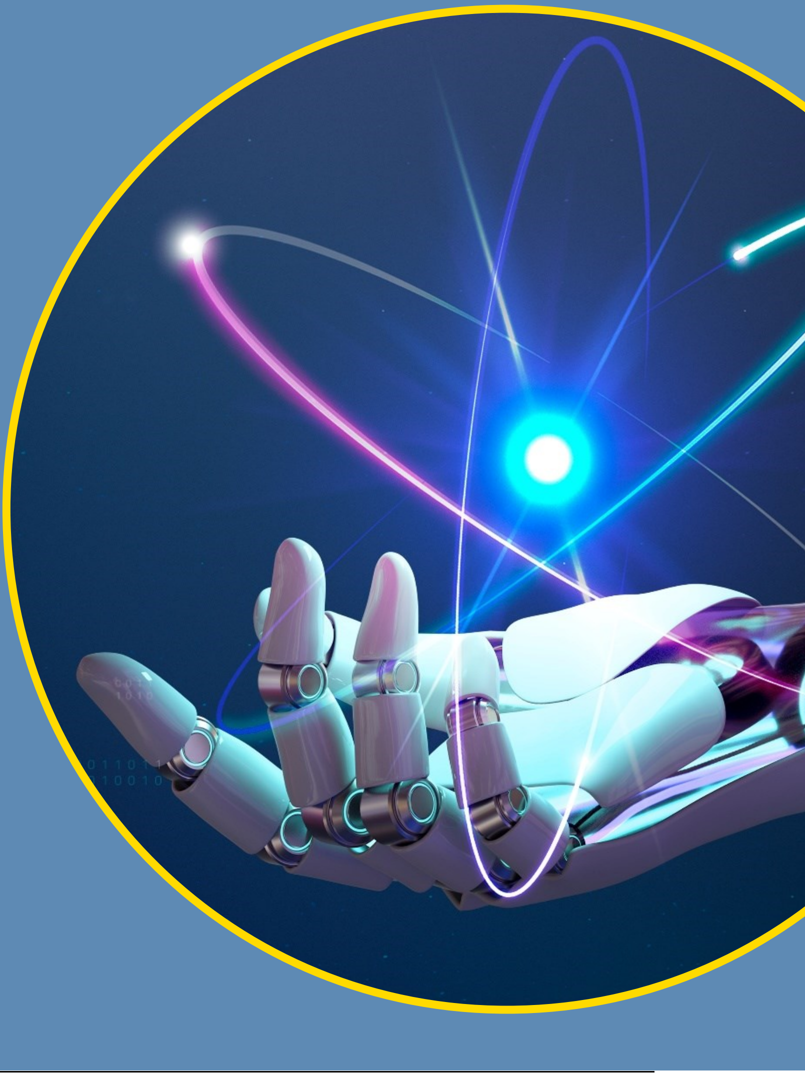
The three most important metrics for feasibility considered were the **accuracy**, **sustainability**, and **transparency** of a method of prediction:

- **Accuracy** is important to ensure bias can be detected to a reasonable level of confidence. However, variation in the availability and quality of accuracy metrics makes a calculation of minimum detectable bias using open literature very uncertain. A recommendation of this report is to perform independent reviews of the performance of proxy methods to gain a better understanding of their performance on UK specific datasets.
- It is important to note that due to the inherent changing nature of population demographics, proxies will change over time and hence the performance of models will change, this is referred to as a model's **sustainability**. Any method that is not continuously retrained on new data can become out of date and recommendations should be made on the frequency of retraining and revalidating methods. Another recommendation of this report is further work on assessing sustainability of proxies.
- Although **transparency** does not impact the performance of a model it is often needed to gain stakeholder support for use. There is an increasing demand for algorithms to be more transparent to ensure that there is not underlying bias that is masked by the 'black box' nature of machine learning techniques.



Privacy preservation techniques have been considered as the impact of privacy violations will have the most damaging effect. However, the techniques discussed in this report can be used to mitigate these risks and make the likelihood of a violation much lower.

Contents



1 Introduction

It is becoming increasingly common for machine learning (ML) algorithms to be used for decision making, often behind the scenes and not regulated as clearly as human decision making. A recent review into bias in algorithmic decision making (Reference 1) carried out by the Centre for Data Ethics and Innovation (CDEI) found that, to identify bias in a system, access is required to the demographic information about its users. However, due to current privacy laws, this data is not always available. To get around this, data proxies can sometimes be used to infer demographic data from the available user information. Bias detection is becoming more important as there are more reported instances of ingrained algorithmic bias in systems that impact people's lives (Reference 2). The aim of this report is to review some of the existing data proxy methods available to providers in the UK and to assess their feasibility for use in monitoring bias in algorithmic decision-making systems in a privacy-preserving way.

In this report, six methods have been identified that are assessed as likely to be viable for use in the UK as they have been trained on UK or international datasets. They cover a range of different proxies: age, gender, and race/ethnicity as well as maturity: open-source, commercial tools, and academic papers. The remainder of this report will be structured as follows:

- Section 3 provides a technical description of each method.
- Section 4 presents a comparison of the methods and a discussion around metrics used to measure performance.
- Section 1 presents an assessment of the feasibility of each method for use in the UK.

2 Acronyms and Definitions

AHP	Analytical Hierarchy Process
AI	Artificial Intelligence
API	Application Programming Interface
CLI	Command Line Interface
CNN	Convolutional Neural Network
DCNN	Deep Convolutional Neural Network
DL	Deep Learning
GDPR	General Data Protection Regulation
GUI	Graphical User Interface
ICO	Information Commissioner's Office
IEEE	Institute of Electrical and Electronics Engineers
LSTM	Long Short Term Memory
ML	Machine Learning
RAG	Red / Amber / Green
SDK	Software Development Kit
SQL	Structured Query Language

3 Description of Selected Proxy Methods

With the field of AI and machine learning growing at a rapid pace, there are an ever-increasing number of proxy methods and tools available to UK organisations for identifying bias in a system. For this study, a selection of representative methods have been chosen. The selected methods cover a range of different proxies: age, gender, and race/ethnicity as well as types of approach, open-source, commercial tools, and academic papers. For the review to be useful, each of the models had to meet certain criteria, such as being trained on international or UK datasets and have enough publicly available performance data to assess their usefulness.

3.1 Discounted Methods

The methods presented in Section 0 were chosen as the best to review in the scope of this task, but they are by no means the only proxy methods that are available. There may be use cases where other methods are more suitable, or a wider assessment of available methods is required. For this reason, the other available methods originally considered for inclusion are presented in Appendix A.1. Models not trained on a UK dataset do not necessarily have to be discounted as they could be retrained on a more representative dataset. However, for race/ethnicity prediction methods, like BISG, designed for use in the US, the outputs are less beneficial where the categories don't match the UK population demographic.

3.2 Investigated Methods

The methods described in this section, were down selected from those presented in Appendix A.1 as the best methods for a more detailed investigation into feasibility of use in the UK. A summary of the methods is provided in Table 1 later in this section.

1. **Namsor**

   Namsor is a commercial tool that makes use of artificial intelligence (AI) to classify names by country of origin, ethnicity, country of residence and gender. It uses a specialised data mining software to recognise the linguistic or cultural origin of personal names in any supported alphabet/language (Reference 3). As it is a commercial tool, it is continually maintained with new training data being periodically added. Use of the service incurs a fee.

2. **Demographic Inference and Representative Population Estimates from Multilingual Social Media Data (Social Media)**

   This is a model from an academic paper which describes a deep learning (DL) system for demographic inference. It was trained on a Twitter dataset using profile images, screen names, names, and biographies (Reference 4). Automatic machine translation is used when biographies are written in languages other than English, this is done per word and so some reduced performance could reasonably be expected. The model is open source and trained on an international dataset but is currently not maintained.

3. **Ethnicolr**

   Ethnicolr is an open-source model that uses United States census data, Florida voting registration data and Wikipedia data to predict race and ethnicity based on first and last names or just last name. The method uses a Long Short Term Memory (LSTM) Network to model the relationship between the sequence of characters in a name and the race/ethnicity of a person (Reference 5).

4. **Wiki-Gendersort**

   Wiki-Gendersort is an open-source tool which uses first names to assign gender based on the Wikipedia database. The method reads Wikipedia pages and assigns gender to a name by instances of gendered words e.g., 'he'/'she' (Reference 6).

5. **Age and Gender Classification using Convolutional Neural Networks (Neural Networks)**

This academic paper's model uses facial recognition to classify the age and gender of a person in a picture. The method uses a deep convolutional neural network and trains on open-source real world images (Reference 7). The model itself is open source and trained on an international dataset but is currently not maintained.

6. **Gender API**

Gender API is a commercial tool that predicts gender using a lookup database when provided with an email address or name. The tool can either use country as an input to improve gender prediction or make a prediction about country of origin from the name (Reference 8). As it is a commercial tool it is continually maintained with new training data being periodically added, but usage does incur a fee.

3.2.1 Summary of Investigated Methods

Table 1 presents a summary of key properties of the investigated methods.

ID	Method	Type	Proxy	Output
1	Namsor	Commercial tool	Names	Gender, ethnicity, country of origin
2	Demographic Inference and Representative Population Estimates from Multilingual Social Media Data	Academic paper	Twitter bio	Age, gender
3	Ethnicolr	Open source	Names	Race or ethnicity
4	Wiki-Gendersort	Open source	Names	Gender
5	Age and Gender Classification using Convolutional Neural Networks	Academic paper	Images	Age, gender
6	Gender API	Commercial tool	Names & Email address	Gender, race or ethnicity

Table 1 - Summary of methods assessed in this report.

4 Evaluation of Methods

In order to highlight the strengths and weaknesses of each method, a qualitative evaluation was conducted. This has provided discussion points around all the key trade-offs that will need to be made when using any proxy method. Accuracy, granularity, and privacy are complicated metrics that are discussed in further detail in Section 1. The requirements for the method will change depending on the proxy available, the required output, the area of use and the scope of the organisation wishing to implement it. This means there is no all-round best method, but this section aims to guide the reader in a methodology for choosing the most appropriate method and identifying the shortfalls of using various proxy methods.

4.1 Binary Comparison

To begin comparing the methods listed in Table 1, a series of binary (yes/no) questions were asked. The aim of this initial 'first-pass' was to quickly review key features of the methods. This is a good way to identify methods that are not fit for purpose and should be excluded from the organisation's investigation. The results of the initial comparison are presented in Table 2.

The aim of this report is to investigate the potential use of proxy methods in the UK so the first question ('Was the model trained with UK/international data?') was considered the most important. As can be seen in Appendix 1, a lot of methods were rejected from further assessment because they were too US-specific. In other scenarios the requirements may be different, e.g., for a small company the cost of implementing and maintaining a model may be too high and therefore only currently maintained models should be considered. This approach cannot be used in isolation as it does not consider a lot of important metrics; we will discuss this further in Section 4.2.

Question	Namsor	Social media	Ethnicolr	Wiki-Gendersort	Neural networks	Gender API
Was the model trained with UK/international data?	1	1	1	1	1	1
Has the model been tested on UK data?	1	1	0	1	0	1
Could the model be retrained on a new dataset?	0	1	0	0	1	0
Does the model have a user guide?	1	1	1	0	1	1
Is the model maintained?	1	0	1	0	0	1
Is the code open source?	0	1	1	1	1	0
Does the model have a GUI/API?	1	0	1	0	0	1

Table 2 – Initial comparison of metrics using binary questions, highlighted cells indicate a positive result.

4.2 Analytical Hierarchy Process (AHP)

AHP is a technique used to provide a more quantitative approach to deciding the best option by using a set of qualitative and quantitative criteria to judge all options. The method uses pairwise comparison to assign weightings to each criterion (Reference 9). The options are then given a score for each metric, and this combined with the weightings provides a score which can be used to rank them for comparison. The metrics considered in this analysis are summarised in Table 3.

Metric	Description	Scoring criteria
Direct cost	The cost associated with any licences, subscriptions, or one-off payments necessary to use the model.	RAG (red, amber, green) ranking. ▪ Red is the most expensive upfront cost for a model. ▪ Amber is a model with a smaller cost. ▪ Green is a free to use model with no restrictions.
Training set size	This metric ranks the overall size of the training dataset, along with the reliability/quality of the training data.	RAG (red, amber, green) ranking. ▪ Red is either a small dataset (relative to other tested models), a dataset that is unrepresentative of the UK, or a case where no detailed information is provided. ▪ Amber is large dataset which can be applied to the UK. ▪ Green is a very large dataset, with specific data for the UK.
Accuracy	This metric ranks the overall accuracy of a method and uses the Fs1 score as this measurement is more sensitive to smaller groups within a sample, which is often the case for protected characteristics. There are many difficulties associated with trying to gain a good measure of model accuracy, this is further discussed in section 5.1.	F1 score as defined in Section 4.2.2. (Reference 10)
Sustainability	This metric considers whether a model's methodology will become outdated in the future, and if the model would be likely to maintain its accuracy for shifting demographics.	RAG (red, amber, green) ranking. ▪ Red is for a model that will become outdated very quickly. ▪ Amber is a model which can potentially last in the long term but may need some modifications to account for certain demographics. ▪ Green is when the model is considered fully capable of adapting to reasonably expected demographic shifts.

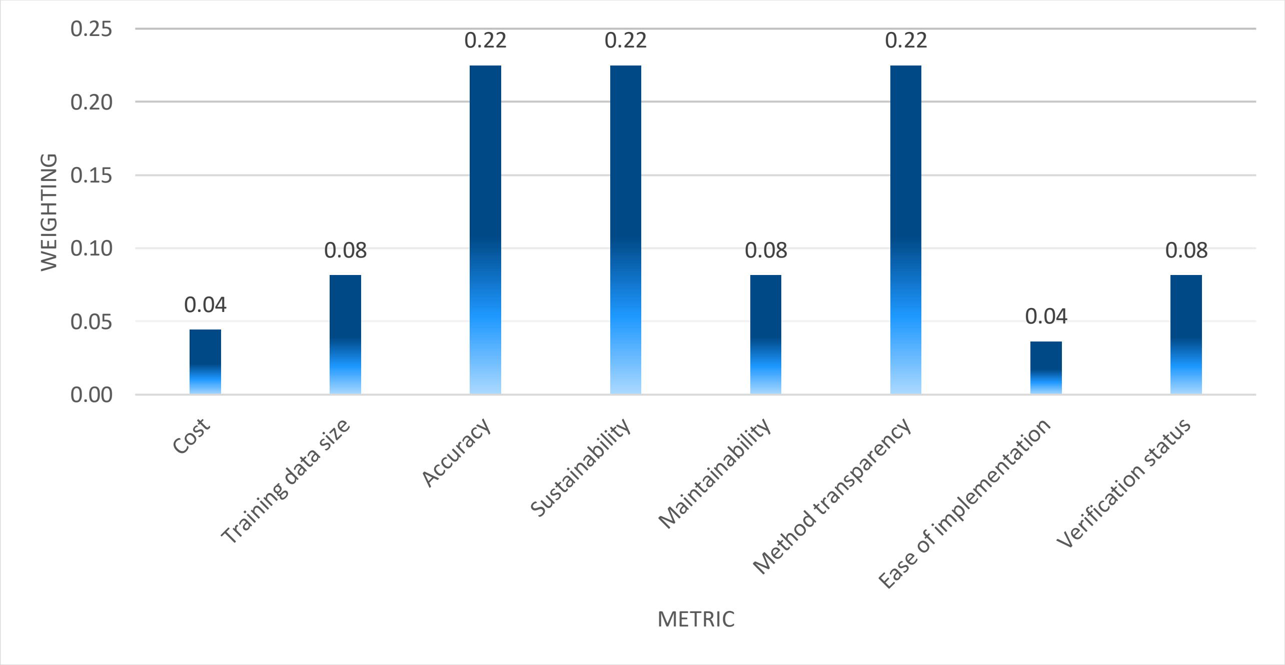
Metric	Description	Scoring criteria
Maintainability	How much effort is required to keep the model running into the future, or to train on more relevant datasets for a specific region.	RAG (red, amber, green) ranking. ▪ Red is for a model that is not maintained by any group. ▪ Amber is for a model that is updated infrequently. ▪ Green is for a model which receives consistent updates.
Transparency	Measures how much of a model's methodology and training data is available to the public.	RAG (red, amber, green) ranking. ▪ Red is for a model which gives no/very little information on the methodology. ▪ Amber is for a model which gives a detailed description of the methodology without providing access. ▪ Green is for a model where the code is available to analyse.
Ease of implementation	A measure of how much effort is required to set up a model and integrate it into existing systems.	RAG (red, amber, green) ranking. ▪ Red is for a model with no user guide or contactable support. ▪ Amber is for a model with a user guide. ▪ Green is for a model with a user guide and a contactable support team.
Verification status	A metric to show the level of verification the model has already been subjected to.	RAG (red, amber, green) ranking. ▪ Red is where no verification can be found. ▪ Amber is when the model has been widely referenced, but without the guarantee of endorsement. ▪ Green is for a model that has been audited and endorsed by multiple different organisations/groups.

Table 3 - Performance metrics.

4.2.1 Ranking Metrics

The first step was to use pairwise comparison to rank each of the determined metrics against each other (Reference 11). This involved comparing each metric one-on-one with all the other metrics and giving a score which signifies which of the two metrics is considered more important. These could then be combined to provide a total weighting for each metric type. The weightings themselves can be varied to fit with the required use case for a proxy method. The proposed metrics for this assessment are presented in Figure 1 and the justifications for each comparison based on the scope of this task are given in Appendix A.2. The weightings have been normalised, so that they sum to 1. It can be seen that for this report **accuracy, sustainability and transparency** are the key metrics with regards to feasibility of use. Accuracy ensures the model is able to adequately detect bias and sustainability ensures the model will continue to do so for a useful amount of time. While transparency does not impact the performance of model it is often needed to gain stakeholder support for use. Cost and ease of implementation are included as factors but given the purpose of this work is to understand feasibility, they are given a lower weighting.

Figure 1 - Proposed weightings for performance metrics.



4.2.2 Ranking of Methods

When using AHP, each method is ranked against a metric and is given a score from 0 to 100. Some metrics were ranked using a traffic light system, where green was equivalent to 100, amber to 50 and red to 0. Where available, F1 score is used as the accuracy metric for each model. F1 score is an accuracy metric for classification of machine learning models that combines the precision and recall scores of a model for each class in the model (Reference 10). The F1 scores are often then averaged to get a single score for the model. This metric is more appropriate than classification accuracy when classes are imbalanced or when false results have a disproportionately high impact.

4.2.2.1 Namsor

Upfront Cost	Training data	Accuracy	Sustainability	Maintainability	Method transparency	Ease of implementation	Verification status	Total
0	8.2	22	11.2	8.2	0	3.6	8.2	**61.4**

Advantages: Namsor is a highly accurate commercial tool, achieving an F1 score of 97.9 (Reference 12), which comes with many benefits. The tool is maintained and consistently updated by a dedicated team, constantly improving the accuracy of any results and increasing the size of the training dataset, with Namsor claiming the tool has processed 7.5 billion names (Reference 3). Other advantages include that the tool has been externally verified by multiple different sources (Reference 13, 14, 15) and the tool is simple to implement into systems, with fully documented APIs provided, along with support staff to contact.

Disadvantages: Though the tool has been externally verified, the methodology behind the tool is proprietary software and not available for the public to see, making it harder for the public to understand how the tool works and whether or not to trust its results. There is also an upfront cost because of the tool's subscription model. This is a constant cost, though this could be mitigated by savings due to the Namsor team completing the backend maintenance themselves, along with the help they can provide for implementing Namsor into existing systems. The sustainability of the tool will depend on how Namsor adapts to changing demographics, such as the increase in people identifying as non-binary and if the tool will be changed to make non-binary predictions. There can also be changes to what gender or ethnicity a name is associated with as time progresses, though these changes should be accounted for with the continued training.

4.2.2.2 Demographic Inference and Representative Population Estimates from Multilingual Social Media Data

Upfront Cost	Training data	Accuracy	Sustainability	Maintainability	Method transparency	Ease of implementation	Verification status	Total
4.4	8.2	20.6	11.2	4.1	22.5	1.8	4.1	**76.9**

Advantages: Some of the main advantages of this tool stem from the open-source nature of the tool. This gives the tool no upfront costs, as it can be freely downloaded from GitHub, so there are also no subscription costs. The tool's methodology is also freely available to look over and scrutinise. The openness of the tool should help increase public trust and confidence in any results produced. The tool had a very large initial training dataset, using 14.53M twitter accounts for gender recognition training and 2.61M for age recognition training (Reference 4). For the gender detection aspect of the tool, the F1 score was 91.8.

Disadvantages: Despite there being no upfront costs, there would still be running costs associated with maintaining and implementing the tool. The tool still has sporadic updates, but there is no dedicated team. As the tool is open source, the maintenance can be performed by any group with the technical knowledge to do so. The tool does have a user guide to help with implementation, though there is no contact team to help with any queries, and the tool is only written in one language. The paper for the tool has been widely cited, but it is unknown if these are papers verifying the validity of the tool or scrutinising it. The tool's sustainability will be affected by the rising number of people identifying as non-binary, who cannot be correctly categorised by this tool in its current state. Further development of the tool would be required in the future for it to be able to accurately identify all members of the population.

4.2.2.3 Ethnicolr

Upfront Cost	Training data	Accuracy	Sustainability	Maintainability	Method transparency	Ease of implementation	Verification status	Total
4.4	4.1	16.4	22.5	4.1	22.5	1.8	4.1	**79.9**

Advantages: Ethnicolr is an open-source tool, which comes with many advantages. The tool is freely available to download from GitHub, meaning there are no initial upfront costs or recurring fees for using the tool. The tool itself is also fully transparent, with the methodology publicly available to be verified and scrutinised. The paper detailing how the tool works (Reference 5) has also been cited 132 times, though it is not possible to know if more sources were praising the methodology or scrutinise it. There are two versions of the tool, trained on different datasets: one set is Florida voter registration data, the other is Wikipedia data. Since the Wikipedia data uses international data, its results are more applicable to the UK. The dataset had over 140,000 names and gave an F1 score of 73. Although this is a lower score than other tools, as there are more categories for predicting ethnicity, it's difficult to directly compare to other tools, which often have a binary choice for predicting gender. The tool is very sustainable, with a deep range of ethnicities that can be predicted. As time goes on, different ethnic groups can become more integrated, but the tool will be able to handle these demographic changes using new training data, so the tool itself would not need fundamental changes to categorising people.

Disadvantages: The main disadvantage of this tool is the maintainability and its ease of implementation. The tool is currently maintained online by an active community, but there is no guarantee this support will continue in the future. This unreliability of the maintenance could incur unforeseen costs. Whilst the tool does have a user guide, if there was trouble with implementing the tool into an already existing system, unlike for most commercial tools, there would be no help team to contact.

4.2.2.4 Wiki-Gendersort

Upfront Cost	Training data	Accuracy	Sustainability	Maintainability	Method transparency	Ease of implementation	Verification status	Total
4.4	4.1	20	11.2	4.1	11.2	0	0	**55.1**

Advantages: Wiki-Gendersort is an open-source tool which comes with a few key advantages. The tool has no upfront or recurring fees to use the tool as it can be freely downloaded from GitHub. As the tool is open source, the method is completely transparent and can be fully scrutinised and verified by the public. A fairly large training dataset consisting of 694,376 names was used to train the model (Reference 6), with the dataset consisting of international names. Wiki-Gendersort has an F1 score of 89.1, so the tool does give accurate results.

Disadvantages: The main disadvantages of this tool are its lack of verification and its difficulty in implementation. The number of citations for the paper detailing the method (Reference 6) could not be found, so no external verification for the method is available. The tool also has no user guide or dedicated team to contact when attempting to implement the tool into any existing system. The maintainability is another weakness of the model, as it does not receive regular updates, though due to it being open source, the code can be maintained by any team with the correct technical knowledge. A potential concern with the model is its sustainability. As the method categorises gender as a binary, by default it cannot categorise non-binary people accurately. As this group of people gets larger, this would make the tool less accurate.

4.2.2.5 Age and Gender Classification using Convolution Neural Networks

Upfront Cost	Training data	Accuracy	Sustainability	Maintainability	Method transparency	Ease of implementation	Verification status	Total
4.4	0	19.5	11.2	4.1	22.5	1.8	4.1	**67.7**

Advantages: The tool is open source, which provides the usual benefits for the upfront costs and method transparency. As the model can be downloaded from GitHub for free, there are no fees that must be paid to use the tool. As the code for the model is feely available, it is also possible for members of the public to verify and scrutinise the methodology, which would help with trusting the model. The tool has been highly cited (576 citations) which shows the paper has been scrutinised by many academics, though it is not possible to know how many of them verified the method as opposed to debunking it. The tool does have a user guide to help with implementation, though there is no dedicated support team like there would be for commercial tools.

Disadvantages: The main disadvantages of this tool are its lack of clarity on its training data, and problems with sustainability. The amount of training data used by the model was not explicitly stated in the study detailing the methodology (Reference 7). The tool also only recognises gender as a binary, which may become unsuitable over time as more people identify as non-binary. There is also an issue with maintenance, as this tool does not receive regular updates; this issue can be overcome as the tool is open source, so the tool could be maintained internally.

4.2.2.6 Gender API

Upfront Cost	Training data	Accuracy	Sustainability	Maintainability	Method transparency	Ease of implementation	Verification status	Total
2.2	8.2	22.1	11.2	8.2	0	3.6	0	**55.6**

Advantages: Gender API is a commercial tool, with its main advantages being its large set of training data, high accuracy, and its high level of maintenance. Gender API uses a database of over 6,000,000 names, with a UK specific database consisting of over 100,000 names (Reference 8), and the database is updated every month. In one study, Gender API had an F1 score of 98.5 (Reference 12), which is a very high score (the highest of all methods tested in this report). As this is a commercial tool, the maintenance is handled by the Gender API team, implementing consistent updates to the tool. Gender API can be very easily implemented into existing systems, with the API working in multiple languages and able to accept a range of file types, along with a support team that can be contacted with any queries.

Disadvantages: The main disadvantages of Gender API are its lack of transparency, upfront costs, and sustainability. The model is a commercial tool and does not have the source code or exact methodology open for scrutiny. Gender API also do not provide any external verification for the tool, though academic studies have tested the tool (Reference 12). The tool is provided on a subscription model, so there are recurring fees to using the tool, though these costs may be mitigated by reduced maintenance and implementation costs. For sustainability, the model is consistently updated with new names (mainly from publicly available data: government data with manual additions), which should allow the model to account for demographic shifts associated with names. The main issue is that the gender detection cannot detect non-binary people, which can become an issue, especially as trends point towards more people identifying as non-binary in the future.

4.2.3 Results

The scores for each model defined in section 4.2.2 have been combined with the weightings presented in Section 4.2.1 to provide the overall ranking for each method presented in Table 4.

	Namsor	Social media	Ethnicolr	Wiki-Gendersort	Neural networks	Gender API
Upfront Cost	0	4.4	4.4	4.4	4.4	2.2
Training data	8.2	8.2	4.1	4.1	0	8.2
Accuracy	22	20.6	16.4	20	19.5	22.1
Sustainability	11.2	11.2	22.5	11.2	11.2	11.2
Maintainability	8.2	4.1	4.1	4.1	4.1	8.2
Method transparency	0	22.5	22.5	11.2	22.5	0
Ease of implementation	3.6	1.8	1.8	0	1.8	3.6
Verification status	8.2	4.1	4.1	0	4.1	0
Total	**61.4**	**76.9**	**79.9**	**55.1**	**67.7**	**55.6**

Table 4 - Final AHP rankings

The results presented in Table 4 show similar rankings between the methods, particularly between Namsor, Social Media and Ethnicolr. This is potentially interesting as they are very different methods using different proxies. Namsor scores very poorly on upfront cost and transparency as it is commercial tool, but its ease of implementation and maintainability outweigh any open-source methods available. It is important to remember that the weightings provided in Section 4.2.1 will have a big impact on the overall ranking of methods. They are shown to aid the discussion in Section 0 but should be adapted for each use case depending on stakeholder priorities.

One of the limitations of AHP used here is the fact that different key metrics can often help achieve similar objectives and compensate for them. For example, the main benefit from both verification status and transparency is an increase in trust in the model.

A model could have a high score in one of these metrics and a low score in another (e.g., Namsor is a closed source tool with many external audits performed on it to verify its accuracy). Whilst the model can be deemed trustworthy due its verification, the low transparency may bring down its score more than it reasonably should, as one of the negatives of a low transparency (low trust) has been mitigated by another metric. The process does not account for scenarios where a low score in one metric is compensated for by another metric, adding a level of uncertainty to a model’s final score. Where a specific use case is well understood the weighting on each metric will become more important to the final result. The metrics cannot just be merged, as there are other factors that are captured by the different metrics, with transparency also helping to increase an understanding of the model that can help with performing maintenance or implementation of the model.

4.3 Discussion

4.3.1 Accuracy

It was not in the scope of this task to perform independent quantitative accuracy assessments of the methods being reviewed. For the accuracy results we have relied on already reported accuracy figures. These can give an indication of performance, but it is important to note the main shortfalls with this approach so that unjustified confidence is not placed in the outcome:

- For some of the methods assessed the accuracy is only self-reported; this is mainly seen in the academic papers reviewed as they include their own testing. Therefore, without independent verification, there is the risk of bias when the author has chosen the dataset and the accuracy metrics to report.

- It is very difficult to distil the accuracy down to a singular figure when classifying groups with large population imbalances. This is because of **the accuracy paradox**, where accuracy is found to be a poor measure of performance when dealing with small groups in a population (Reference 16). Therefore, different metrics need to be used. For gender detection methods, the accuracy paradox is less of a problem if you are considering only male or female. For ethnicity predictors or gender predictors which can predict non-binary genders, the accuracy paradox is a much bigger problem. For this report the **F1 score** was used instead of accuracy. The F1 score is the harmonic mean of the precision and recall. Precision measures valid positive predictions, and the recall measures valid negative predictions (Reference 17).

- The accuracy metric also only covers specific aspects of the models e.g., Namsor's gender prediction had F1 score used as an accuracy measure, but this does not represent its accuracy in predicting diaspora from names. The accuracy of a multi-faceted tool cannot be simplified down to a single figure, with each aspect considered separately.

- For the accuracy metric, Ethnicolr is the only model which is using its F1 score for predicting ethnicity as opposed to gender. Gender and ethnicity predictions are very different tasks, with gender predictions often being a binary choice, as opposed to a large range of potential ethnicities. This should be kept in mind when comparing the accuracy of Ethnicolr to the other models.

- **The tests won't have been performed on the same dataset** (except in the case of Reference 12, where 4 proxy methods are assessed in one review). A good comparison of accuracy can be made here; it is recommended more assessments like this are sought or otherwise carried out independently before implementing a method.

- **The dataset is very important to the accuracy performance** – a review of the same methods as in Reference 12 but done on Chinese names in Pinyin format showed significantly poorer performance (Reference 18). This problem has been highlighted in much of the literature reviewed; non-Western names are likely to see poorer performance. This is a concern as when used in the Western world they would represent the minorities that bias detection is targeted at. It is recommended that if performing an independent assessment, testing data as close as possible to the actual use case should be used.

- The metric reported can impact the perceived performance – when **classification accuracy** is used as the metric it can therefore be misleading. Reference 19 highlights a weighting in an ethnicity estimator tool that favours ethnic minorities over white British. This should lead to a higher accuracy performance for ethnic minorities, e.g., less likely to misclassify a person as 'white' at the expense of misclassifying 'white' names.

- Do the accuracy metrics cover all the necessary information? In most cases when assessing the accuracy of a gender estimator tool, performance is measured for men and women. It was shown in Reference 2 that performance can vary within this category, a tool might have good accuracy for women but when looking deeper into the data performance is poor for women from ethnic minorities. If carrying out an independent review, filtered accuracy results would increase understanding of performance or applying weightings to minority results (this is sometimes referred to as '**balanced accuracy**'). (Reference 20)

While reported accuracy scores can give a good indication of performance when initially reviewing methods and potentially discounting the use of some, they should not be used in place of first-hand testing to provide confidence in a model. There is also a risk of models being used because they have been cited many times and so appear trusted by the community, again this should not be used in place of testing as the use case could be different or the method unsustainable and now out of date. Before any proxy method is used, to provide full transparency, an independent review should be carried out on its performance where the model is queried for each use case on a representative dataset. Performance should be assessed for each predicted category, considering the effect of false positives and false negatives.

4.3.2 Minimum detectable bias

For a tool to be used in bias detection there must be considerable **confidence that the results are more accurate than random chance**. As discussed in the previous section, measures of performance are complicated and without an independent review the minimum detectable bias for a tool is unknown. A conservative way to provide an initial estimate is to use the classification accuracy metric.

As an example:

A model is correct 75% of the time, meaning any bias in the system that is less than 25% could potentially be missed by the system and so we can class this as the 'minimum detectable bias'. In a dataset where 15% of people are female, there is the potential for them all to be misclassified by the model and for bias to go undetected.

4.3.2.1 Gender predictors

Out of the methods assessed in this report Ethnicolr is the only tool that does not makes a gender prediction, so a conservative minimum detectable bias metric is given for the other five methods in Table 5. It is important to note that these values will vary with the dataset used, Reference 18 reports much lower accuracy scores using Chinese names and so would predict much higher minimum detectable bias.

Method	Minimum detectable bias	Notes
Namsor	2%	Average of male and female classification accuracy values (Reference 12)
Demographic Inference and Representative Population Estimates from Multilingual Social Media Data	10%	No accuracy value available, based on F1 metric (Reference 4)
Wiki-Gendersort	6.6%	Average of male and female classification accuracy values (Reference 12)
Age and Gender Classification using Convolutional Neural Networks	22.5%	Average of accuracy values for two datasets tested on (Reference 7)
Gender API	1.8%	Average of male and female classification accuracy values (Reference 12)

Table 5 - Conservative minimum detectable bias predictions, where accuracy score was not available F1 is used.

4.3.2.2 Race/ethnicity predictors

Out of the methods assessed in this report, three claim to make race/ethnicity or country of origin predictions. However, there is no performance data for Gender API in the open literature, so the two available methods have been presented in Table 6.

Method	Minimum detectable bias	Notes
Namsor	29%	Accuracy of 'country of origin' metric (Reference 21)
Ethnicolr	27%	Average F1 score from global race predictions (Reference 5)

Table 6 - Conservative minimum detectable bias predictions, where accuracy score was not available F1 is used.

4.3.3 Ease of Implementation and Maintainability Versus Cost

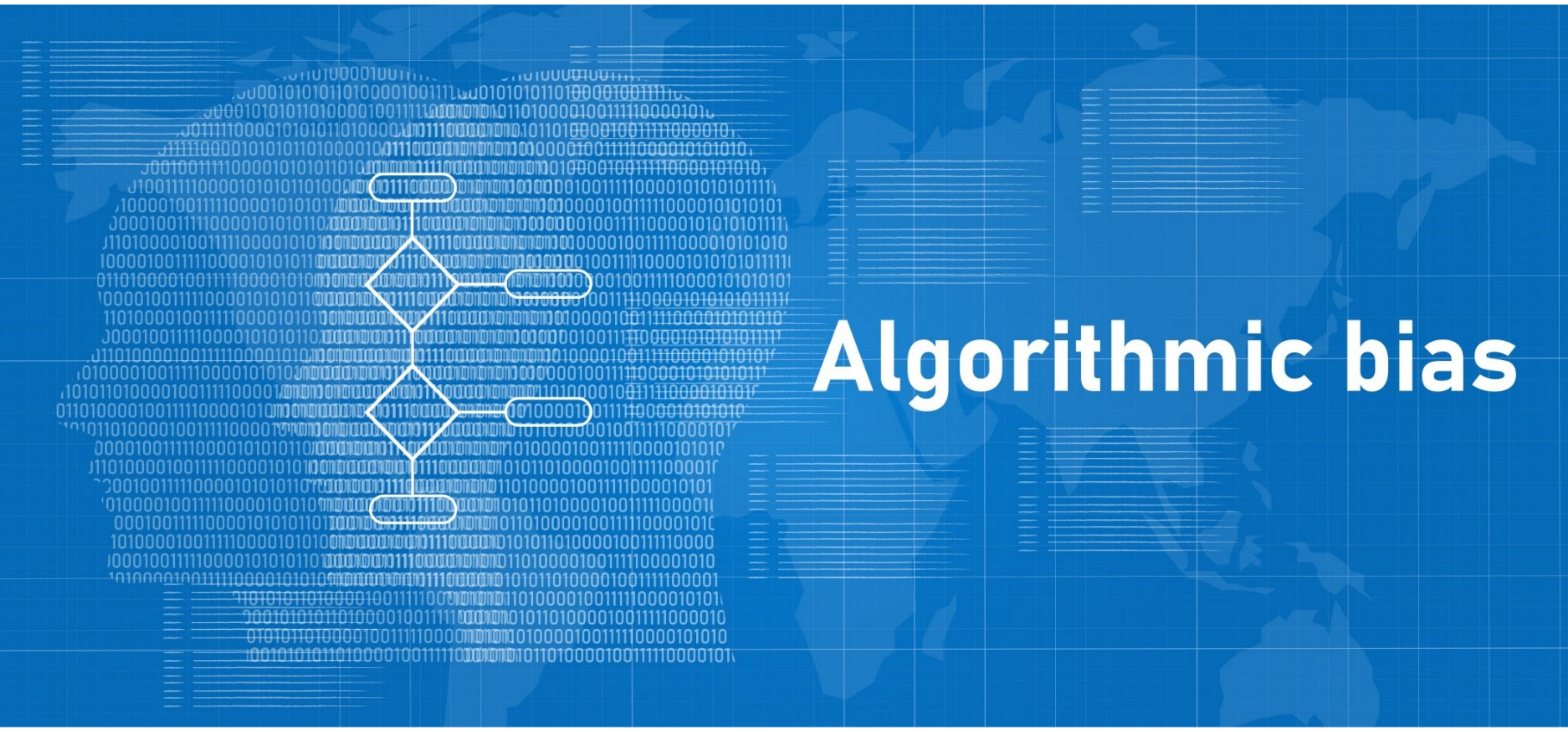
When reviewing proxy methods there is a clear trade-off between commercial tools and open-source models, that has been summarised in Table 7. With a commercial tool the transparent costing and support team can help to de-risk its use, so it is an important consideration to make.

Commercial tool	**Open source**
Set up to be easily used with a GUI or an API implementation so very little time or investment is required to get going.	If there is a GitHub repository or equivalent the setup will be less straightforward, but there is often a walkthrough guide. More time and potentially cost investment is needed at the outset.
Usually maintained, this means they should work on most systems and update as technology and datasets are updated.	Sometimes maintained, but often not or at least not as rigorously as commercial tools, so some ongoing cost of maintenance/continued training should be considered.
Cost is normally based on number of requests so is easily calculated but scales with the size of the dataset.	Cost is somewhat harder to determine as it will be based on the cost to set up the tool and potentially hiring/training someone to use and maintain it.
For well-known tools there are many documented use cases and published accuracy results from independent sources.	There is less of a requirement to publish accuracy results or, in the case of academic papers, no independent review of the model/results.
There is a risk they could be discontinued, so if they are part of a system this could cause problems later down the line.	There are privacy and security implications when using open-source tools and/or datasets.

Table 7 - Trade-offs between commercial tools and open-source models.

4.3.4 Transparency

The transparency score given in Table 4 is a measure of how much of a model's methodology and training data is available to the public. Some open-source methods are completely transparent in that they provide full source code and use open-source training data. This does not guarantee full transparency though - machine learning algorithms by nature are not very transparent (this is often called the 'black box problem'). Often, for complicated problems, even the people who have created and trained them do not fully understand how all the variables relate to each other to make predictions. Where a model has an accompanying academic paper there is more transparency, in that for the choice of technique and any parameter-tuning justification is provided. However, this is only transparent for people with the required knowledge base to understand the paper. One could argue that a proxy method with very good accuracy performance does not need to be transparent to work, but it is hard to garner trust in a system people do not understand. There are numerous media examples of AI making errors and introducing bias; this has consequently increased the need for transparency. More interpretable models are likely to receive **higher stakeholder confidence** as they cannot hide bias as easily. (Reference 19)



5 Feasibility of Methods

Note - The aim of this report is to provide insight and guidance on the feasibility of using proxy methods. It is not to provide legal advice and does not provide legally approved conclusions.

5.1 Accuracy Assessment

5.1.1 Measures of Accuracy

The standard measure of classifier models is usually their classification accuracy (that is, how many predictions they were able to make correctly during the testing phase). However, this figure alone is insufficient in guiding our judgement for most use cases.

For instance, it is necessary to weigh relative costs of both false positives and false negatives, as these costs may differ depending on the usage context. If two models have identical classification accuracies (with one returning more false negatives and the other more false positives), but we consider false positives to be a worse outcome in context, then we should go with the one that has a lower false positive rate. For this purpose, we could potentially ease the problem of computing and evaluating our models using tools such as confusion matrices. From this information we can calculate additional metrics such as **F1 score**; these give us a more contextualised understanding of model performance (Reference 22). It may sometimes even be the case that a model with lower classification accuracy ends up performing better in context – this is an example of the accuracy paradox (Reference 23).

It should be noted that F1 scores (or equivalent) and/or confusion matrices are given by the developers for Ethnicolr, Namsor, Wiki-Gendersort, Gender API, and the Social Media models. In the case of Age and Gender CNN, a confusion matrix is only given for age determination. The figures stated are indicative but should be independently

5.1.2 Sustainability of Accuracy

Accuracy figures are very likely to be fluid over time, reflecting a constantly shifting demographic landscape in the United Kingdom. Take the example of an algorithm which sorts by surname. It is likely that increased multiculturalism will alter distributions over time – a surname that was almost certainly indicative of a particular demographic just 50 years ago may be less of an accurate indicator today, and there is potential for this trend to continue. Some models we have looked at (such as Gender API) attempt to account for these scenarios. To this end, it is important to consider how long a particular method may continue to be viable for use; this has further economic implications such as investment payback time and costs stemming from increased future misclassifications.

***Example Scenario 1:** Consider an algorithm ‘A’ that scans facial features to ascribe labels for gender. When presented with a face that is somewhat androgynous by our societal standards, we might expect it to struggle to correctly classify this person. As fashions change, so will the ways in which people present themselves; this may diminish the effectiveness of the algorithm over time. For example, in the future it may be more common for men to wear heavier makeup. Model A has learned to associate makeup with women and will therefore become less accurate.*

This type of scenario is especially important when we consider that a demographic organisation may want to monitor and prevent bias for its transgender individuals (with the category of gender reassignment being a protected characteristic in the Equality Act 2010 (Reference 24)). Due to disparate access to medical interventions (among other factors), facial scanning algorithms could be expected to fail to recognise this demographic a reasonable amount of the time. Furthermore, if the trend of increasing numbers of younger generations in the UK openly identifying as transgender continues (Reference 25), the model may therefore become less and less accurate with passing generations.

***Example Scenario 2:** Algorithm 'A' is now presented with a subject whose gender identity is non-binary. Most methods and datasets also do not account for non-binary identities, so A has no way of correctly classifying the subject.*

In addition to this, an increasing number of people in the UK identify as a 'mixed' ethnic category (Reference 26), ethnicity being another protected characteristic (Reference 24). This could also be expected to diminish the accuracy of face-scanning approaches to ascribing ethnic backgrounds, due to the lines between trained-in algorithmic notions of what the 'average' member of each demographic looks like becoming blurred.

***Example Scenario 3:** Algorithm 'B' ascribes labels for ethnicity based on a picture of a face. It is presented with someone who comes from a family with many different ethnic backgrounds across multiple generations. It therefore struggles to correctly classify this person, as it does not have a strong idea of which ethnicity the person appears to be closest to.*

These examples raise tangentially related questions of importance:

- How can we ensure our algorithm itself is not unfairly biased?
- How ethical is it to create a tool which ascribes identifiers such as race based on what are essentially highly stereotypical depictions of often-protected demographic groups?
- What are the associated challenges with training models to recognise demographics such as mixed ethnicities or non-binary genders which by their very nature blur traditional expectations of categorisation?

A potential workaround for diminishing accuracy is the implementation of algorithms which may either undergo regular retraining or continuous learning. However, the main drawback here is that it is often far more difficult to assure agents like this for use in sensitive contexts, due to both the relative lack of explainability of 'black box' models and the unpredictability of future performance. An algorithm that is constantly learning and retraining itself also gives rise to additional concerns about the increased likelihood of a system being manipulable by threat actors, for example by dataset poisoning (which itself could significantly damage the accuracy of outputs). It may be possible to balance risks by reaching some kind of 'middle ground'. One manifestation of this might be a system which is separate from the live deployment, retraining itself at regular intervals and requiring re-approval before being pushed to live.

Either way, initial training will impact accuracy. This raises a potential concern in how training data will be sourced if it is not readily available to organisations in a sufficiently representative form. If training data is to be volunteered by data subjects, there may be systematic biases in who tends to volunteer this data. This could result in a biased model. Bias in the training set is a particularly prevalent concern in photographs of faces (Reference 27). Additionally, any pre-existing training data runs the risk of being potentially outdated due to its historical nature.

Regular retraining would also protect against bias creep as the general population distribution shifts in time (in directions which may mean an initial training set that was once sufficiently representative is no longer so). Representative distribution is especially important when we consider the potential indicators of group that may differ from one demographic to another. For example, the stereotypical (or societally expected) presentations for different genders may vary by age group or ethnic background. This too is likely to be highly susceptible to time. Take the example of having long hair – in many Western cultures this is typically associated with women, but in many Native American cultures is much more a unisex practice. We have no way of knowing if this will continue to be the case in future, though. The rate required for retraining would vary by model and usage context, but a reasonable periodicity may be defined by tracking and testing accuracy rates and noting when the model drops below a certain threshold.

In the case of the spread of models we have investigated, Namsor and Gender API are known to use continuously updating datasets; the Social Media model and Age and Gender CNN do not. While it is uncertain whether Ethnicolr and Gendersort continue to be maintained, they are based on the Wikipedia dataset which stems from a continually updating source, perhaps lending itself to more straightforward retraining.

5.1.3 Summary of Accuracy Risks

Risk	Likelihood	Potential impact	Mitigations	Post-mitigation likelihood
We choose the wrong model based on accuracy figures alone.	Very likely	The model performs poorly, adding costs of misclassification and remediation.	Use additional measures such as confusion matrices, F1 score, and specific use case analysis.	Very unlikely
We see diminishing returns on classification algorithms due to demographic shift.	Moderately likely	The model becomes less usable over time, diminishing returns on investment and increasing any misclassification costs.	To an extent this can be avoided by regular retraining or models that learn from changing data, though this comes with its own risks. New emerging demographics may be much more difficult to classify and mitigate.	Somewhat likely

Table 8 - Risks to accuracy and their potential mitigations. Any likelihood figures are estimates only.

5.2 Granularity Assessment

For our purposes, 'granularity' and 'privacy' can be closely linked terms depending on how we are defining the latter. We can take privacy to have any of the following meanings in our context: protecting people's data being put into a system; protecting the inner workings of the systems from being exposed and therefore at risk of manipulation by threat actors; or protecting individuals from having sensitive information identifiably attached to them which they did not voluntarily surrender (i.e., the consequence of insufficient coarseness of granularity). In this section, we will largely be considering the last of these three. Therefore, this section will primarily discuss aspects granularity through their relationship to **data protection**.

We can broadly split techniques for aggregation depending on whether they are designed around blurring the data or simply masking it. It should be noted that some methods naturally combine better with certain tools and approaches; this should be a consideration when selecting the right algorithm for the context. We have learned that the level of granularity required for each model depends upon the contextual use-case, but in general a small dataset will produce a lack of anonymisation. The ICO recommends applying a 'motivated intruder' test to ensure the removal of identifying data is accurate (Reference 28); we believe this would be a useful technique in determining the correct level of granularity has been achieved for a given usage scenario.

5.2.1 Data Blurring

Basic aggregation of all the data together does not necessarily guarantee privacy. For example, if we apply very specific sets of filters that have very few results each, it may be possible to determine who belongs to the results subset (i.e., what personally identifiable labels we have ascribed to them) (Reference 29). Simple aggregation functions exist for imported datasets in R, SQL, and Python - however, a pre-existing set of non-anonymised data is generally required to feed in. Basic aggregation would therefore have to be 'baked in' the original algorithmic program to be potentially allowable. Aggregate data can have random noise added to prevent reverse-engineering, but this obviously affects the values in the output dataset, which may not be acceptable in a particular use case.

Data generalisation (also sometimes called **clusters**, **binning** or **blurring**) can be achieved in various ways, such as providing less-specific values or bins (e.g., identifying by birth month instead of full date). Generalisation may also be **automated** or **declarative**. See the example paper by Samarati and Sweeney for more details (Reference 30). Another example could be pixelating visual data (in the case of photographs of faces), though this might severely limit the usefulness of applying the algorithm.

A general limitation of these kinds of techniques is that they are not effective for smaller datasets. These techniques also increase the likelihood of us drawing incorrect conclusions from the data (Reference 31).

5.2.2 Data Masking

Differential privacy takes a maths-based definition of privacy for the purposes of data protection. As a technique, it describes patterns of the whole group while withholding individual identifiers. Some ways of pursuing differential privacy include the Laplace mechanism (addition of noise) and randomised response (perturbation) (Reference 32). However, there is a risk of introducing bias using the latter approach if the spread of truthful and randomised responses is not correctly balanced. It is also possible to add in noise during the model training process, or to the outputs. Again, data with added noise carries the risk of being less helpful for its intended use case.

Tokenization is an approach that replaces sensitive information with a different datum of equal type and length. A database may be required to link tokens to their true meanings; this presents a vector for accidental or malicious leak of personal data and requires extra computational requirements for implementation that expand with the amount of category labels being counted.

Pseudonymization is defined in GDPR as 'processing of personal data in such a way that the data can no longer be attributed to a specific data subject without the use of additional information, as long as additional information is kept separately and subject to technical and organizational measures to ensure non-attribution to an identified or identifiable individual' (Reference 33). **Data coding** is when private information is removed and replaced with numbers or codified categories, similar to tokenization above. Whether or not this is helpful depends on the individual use case, particularly the data to be codified and its sensitivity.

Algorithms used to mask data can sometimes be reverse engineered. The efficacy of these techniques would be limited for small datasets. Techniques such as randomising response can also affect the calculations we may want to perform on the data. **K-anonymity** is the measure that defines how re-identifiable data records may be. True anonymity occurs if quasi-identifiers for each person in the dataset match at least k-1 other people also in the set.

5.2.3 Summary of Granularity Risks

Risk	Likelihood	Potential impact	Mitigations	Post-mitigation likelihood
Insufficient coarseness of granularity results in an unnecessarily high privacy risk.	Moderately likely	Personal information is uniquely identifiable, and privacy is violated. This also presents legal and ethical consequences.	Only process datasets of sufficient size and select an appropriate privacy-preserving method for the use case. Apply the motivated intruder test as a check.	Very unlikely
Noise or blurring significantly impacts model usefulness or the conclusions we draw.	Moderately likely	The model is less useful, presenting decreased return on investment and increased total misclassification costs.	Conduct experiments to find an appropriate level of noise / blur.	Unlikely
Computational requirements exceed those allotted for the use case.	Moderately likely (but highly context dependent)	Additional computational resource must be secured, presenting a cost. This cost would also be ongoing due to upkeep.	Steer away from privacy-preserving techniques which have intrinsically high computational resource requirements (such as tokenization). Use models which are easy to integrate with techniques.	Unlikely
Masked data is reverse engineered by a threat actor.	Unlikely	A privacy violation occurs, with legal and ethical implications.	Use a blurring technique. Apply the motivated intruder test as a check.	Very unlikely

Table 9 – Risks to granularity and their potential mitigations. Any likelihood figures are estimates only.

5.3 Privacy Assessment

We must also consider how to protect the inner workings of systems from being exposed and therefore manipulated. We could also discuss this in terms of some of the cyber security aspects of the model; the IEEE remarks that ‘the goals of data privacy and information security are overlapping in some instances’ (Reference 34). This section will therefore look more closely at **model security**.

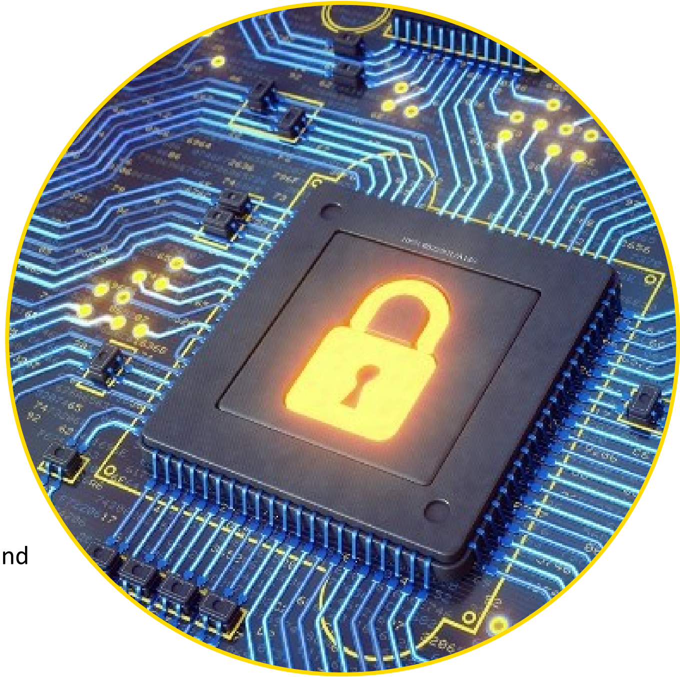
A commonly known weakness with classifier models is the ‘evasion attack’ – the ability of a near-undetectable level of unexpected noise to affect model conclusions. Depending upon the intended model use case and context, it may have a widely varying likelihood of being adversarially targeted in such a way. We must ask ourselves how attractive the system is as a target and what may stand to be gained by misclassifying an input. For the use case of bias monitoring this is potentially less likely than, say, labelling faces for personal identification purposes. It would also require the insertion of invalid data for receipt of services, which would heavily inconvenience the user in many cases. There are still some mitigating measures. Techniques such as denoised smoothing (Reference 35) look promising in the literature but require extra computational resource, whereas other measures such as training to withstand input noise are viable but can affect overall model accuracy and require inclusion from the algorithm design phase.

Obtaining a trustworthy training set is critical. During any training phase (be that the initial model training or ongoing ‘update’ training), it is possible for a threat actor to ‘poison’ the model by inserting crafted inputs. These inputs will have been designed to influence the model to perform in a particular way, to the threat actor’s advantage. Data poisoning is one of the key reasons a continuously-learning model is much harder to assure for use, due to the inherent risk.

It is far easier to craft such a malicious payload if there is freely available information concerning the inner workings of a model and/or the initial training set is publicly known. This presents a key consideration for any model which is open-source or derived from open-source training sets. A publicly usable model is also a more likely victim of ‘model stealing’ attacks (Reference 36), which can be used to glean the data that was used to train them. This is an obvious potential privacy risk, relating back to our earlier discussion of data protection in Section 5.2.

Methods for lowering the risk of adversarial attacks include:

- Limiting how much of the model’s inner workings are public knowledge.
  This is not always possible depending on the model and required use case, and in some cases must be balanced with any requirements for transparency. On a related note, only reliable training sets should be used.
- Rate restrictions (i.e., slowing down the rate of input queries). This is unlikely to be useful for the bias monitoring use case as live models will not typically be accessible to everyone.
- Having separate test and live branches which can be used to
- The inclusion of adversarial detection methods. This is the most detection methods which sort the ‘good’ data from the ‘bad’ – and vector.



The incentives for an adversary to attack a model may vary. In the case of bias detection, likely motives may include a desire to disrupt the organisation's ability to monitor for bias, or a desire to be misclassified due to a perceived gain. It is also a common trend throughout recent history for attackers to attempt to compromise tools simply for the perceived challenge.

The impacts of an attack can be minimised by having both a robust incident response plan and the resources to regularly assess that a model is still performing as expected. Note that the privacy requirements for the deployed system will vary widely by stakeholder group – some groups will need (and should have) greater and/or different transparency needs.

The IEEE Standard for Data Privacy Process (Reference 34) provides recommendations for deploying systems that use personal data. Some of these include producing a detailed map of data transits, identifying an optimal set of procedures, and ongoing monitoring.

To some extent, all the models we have examined in detail use openly available data such as government census information or Creative Commons photo databases. In this report, two of the models we have looked at use datasets derived from Wikipedia. This is a source which may be susceptible to intentional sabotage, though the scale at this which would be required in order to significantly affect model outputs may be infeasible for most potential threat actors. Namsor and Gender API are the only tools we have looked at which are not open-source – the rest are available via GitHub repository. It may be possible for open-source tools to be retrained on a different set if a suitable data source is available, however the feasibility and/or complexity of this could vary.

5.3.1 Summary of Privacy Risks

Risk	Likelihood	Potential impact	Mitigations	Post-mitigation likelihood
An evasion attack is used to purposely misclassify an input.	Very unlikely	Misclassification metrics increase.	Use denoised smoothing or train the model to withstand noise. An adversary detection method could be deployed. However, it is arguably reasonable to accept this risk due to the very low likelihood of occurrence.	Very unlikely
Training datasets are poisoned, intentionally or accidentally.	Somewhat likely (but highly context dependent)	The model becomes less usable and requires a cost to be fixed.	Vet datasets for use and ensure both they and/or the model's inner workings are kept private where this is feasible. Use separate live and test deployments.	Unlikely
A model stealing attack is used to determine training data.	N/A	Identifiable information from the training set is leaked.	Not applicable here, as the deployment model is not expected to be publicly usable for our use case(s).	N/A

Table 10 – Risks to privacy and their potential mitigations. Any likelihood figures are estimates only.

5.4 General

Given that granularity is extremely use case dependant, without a specific application, it is not possible to give a reasonable requirement for the exact requirement and doing so for a very simple use case will not provide significant benefit. It is **recommended** that this be reconsidered with a real data set applied to a use case and requirement.

6 Conclusion

In this report, six proxy methods for predicting social demographics with the aim of monitoring bias have been reviewed. The methods are assessed as likely to be viable for use in the UK as they have been trained on UK or international datasets. They cover a range of different proxies: **age, gender, and race/ethnicity** as well as types of method: **open-source, commercial tools, and academic papers**. The report also provides an assessment on the general feasibility of using proxy methods for bias monitoring purposes, considering in detail the accuracy, granularity and privacy risks involved.

It was shown in Section 3 that evaluating the appropriateness or feasibility of a proxy method is very case specific, the requirements will change depending on the proxy available, the required output, the area of use and the scope of the organisation wishing to implement it. The three most important metrics for feasibility considered in this report were the **accuracy, sustainability, and transparency** of a method of prediction.

The **accuracy** of the results is important, as without considerable confidence in the outputs being correct the method has no purpose. In the open-source literature, a wide range of accuracy metrics were presented for classification models; classification accuracy, F1 score and confusion matrices. It is difficult to compare methods where the metrics are not comparable and where the testing datasets are not identical. This has the potential to mask where models are overfitting to the training data. This is most common when a model has been trained to a small or very specific dataset and so has high accuracy when tested on similar data but cannot be extrapolated outside of this window. There have been very few published comparisons of multiple models that could be used to definitively compare accuracy. It is recommenced that before use of any proxy method an independent review is carried out using representative test data to ensure performance is high enough to be useful as a bias detection tool.

One of the main trade-offs to consider is using a **commercial tool** or an **open source one**. Commercial tools provide a lot of benefit in term of ease of use and financial scoping, they are also likely to be more **sustainable** due to continuous training and support teams that are current on social demographic changes. However, their lack of **transparency** can make them harder to trust. Transparency is becoming increasingly important due to the number of reported instances of ingrained algorithmic bias (Reference 2), the very thing bias monitoring aims to prevent. If stakeholders cannot trust in the AI tools not to introduce bias, how can they trust the detection tools not to encounter the same problems. Having a fully transparent process does not guarantee it will be bias free, but it makes it easier to detect any bad traits a ML algorithm is picking up.

In terms of the issues, we have discussed surrounding privacy, accuracy and granularity, we perceive the biggest ongoing potential risk to come from an inability to wholly account for **future changes in UK demographics** while maintaining the same levels of model effectiveness seen at the outset. Consequently, this also links to the related issues of assuring models which learn continuously, and to some extent being able to guarantee that datasets have the lowest possible risk of being poisoned either intentionally or accidentally (though likely the latter for our case).

The risks relating to **privacy** violations conversely have the most damaging perceived consequences, though when mitigating measures are applied these risks can be **greatly reduced**. Determining a granularity requirement without a realistic data set and using an actual algorithm is extremely difficult and may be misleading and as such was not able to be explicitly defined here.

This report has shown that **it is feasible to use proxy methods to predict demographic characteristics** about a person from a range of proxies. However, several ethical matters are associated with these techniques. The main concern is the performance of the models. It could be argued that a tool predicting correctly most of the time is better than using no tool at all. But this is not the case in bias detection where often the aim is to assess the fairness in algorithms towards minorities. Care should be taken when using any of the methods discussed in this report to ensure that accuracy of the bias detection method is not also introducing bias.

7 Recommendations

- Before any proxy method is used or recommended for use an independent review should be carried out on using a **representative dataset** to ensure performance is high enough to be useful as a bias detection tool. Filtered accuracy results or weighted accuracies should be used to increase understanding of performance across the whole population. It is recommended that F1 score be used over classification accuracy, as it considers the effect of false positives and false negatives.

- In general, some more comparisons of proxy methods should be carried out similar to the reports in references 12 and 18. Using the **same testing datasets and performance criteria** would give potential users of the data a more transparent resource to choose the correct method to use.

- It is recommended that the requirement for **granularity** be reconsidered with a real data set applied to a use case and requirement.

- In some use cases a **combination** of proxies may provide performance benefits. For instance, where names are used to predict gender, some cultural context can provide benefit. The name 'Jean' is more often considered feminine in English speaking countries and masculine in French speaking countries. Passing a dataset through a 'country of origin' predictor and then a gender prediction is likely to provide a better prediction than one model alone.

- Different proxies and different population groups will evolve at different rates, so there can be no clear rules for guidance on frequency of retraining models or how often they should be revalidated for use. It is recommended that further work is carried out, looking at **historic** data as well as **current** social and cultural trends to provide some predictions on potential future modelling shortfalls.

8 References

1. Centre for Data Ethics and Innovation (2020) Review into bias in algorithmic decision-making.
2. Shin, T. Real-life Examples of Discriminating Artificial Intelligence. Available at https://towardsdatascience.com/real-life-examples-of-discriminating-artificial-intelligence-cae395a90070 (Accessed: March 7, 2023)
3. Namsor (2023) Namsor, a name checking technology. Available at: https://namsor.app/about-us (Accessed: January 25, 2023).
4. Wang, Z., Hale, S., Ifeoluwa Adelani, D., Grabowicz, P., Hartmann, T., Flöck, F., & Jurgens, D. (2019). Demographic Inference and Representative Population Estimates from Multilingual Social Media Data. WWW'19.
5. Sood, G., Laohaprapanon, S. (2018). Predicting Race and Ethnicity from the Sequence of Characters in a Name. arXiv:1805.02109
6. N Berube, G Ghiasi, M Sainte-Marie & V Lariviere (2020). Wiki-Gendersort: Automatic gender detection using first names in Wikipedia. Available at: https://osf.io/preprints/socarxiv/ezw7p/ (Accessed: January 25, 2023).
7. Levi, G., Hassncer, T. (2020) Age and Gender Classification using Convolutional Neural Networks. The Open University of Israel
8. Gender API (2023) About Us. Available at https://Gender API.com/en/about-us (Accessed: February 21, 2023).
9. Jagoda, J, A., Schuldt, S, J., Hoisington, A, J. (2020) What to Do? Let's Think It Through! Using the Analytic Hierarchy Process to Make Decisions. Available at: https://kids.frontiersin.org/articles/10.3389/frym.2020.00078 (Accessed: February 22, 2023).
10. Korstanje, J (2021) The F1 score. Available at: https://towardsdatascience.com/the-f1-score-bec2bbc38aa6 (Accessed: March 1, 2023).
11. 1000minds. (2023) Pairwise comparison method. Available at: https://www.1000minds.com/decision-making/pairwise-comparison (Accessed: February 21 2023).
12. Sebo, P. (2021) Performance of gender detection tools: a comparative study of name-to-gender inference services. Journal of the Medical Library Association.
13. Bursztyn, L et al. (March 2022) "The Immigrant Next Door: Long-term Contact, Generosity and Prejudice" available at https://namsor.app/files_to_download_p/immigrant-next-door_march2022.pdf
14. Rieke, A; Svirsky, D; Southerland, V; Hsu, M (2022) "Imperfect Inferences: A Practical Assessment" available at https://namsor.app/files_to_download_p/uber_benchmark.pdf
15. Science-Metrix Inc (January 2018) "Analytical Support for Bibliometrics Indicators: Development of Bibliometric Indicators to Measure Women's Contribution to Scientific Publications" available at https://namsor.app/files_to_download_p/science-metrix_bibliometric_indicators_womens_contribution_to_science_report.pdf
16. "Accuracy Paradox", Wikipedia (accessed March 2, 2023). Available at https://en.wikipedia.org/wiki/Accuracy_paradox
17. Lao, R. Machine Learning | Accuracy Paradox. Available at https://www.linkedin.com/pulse/machine-learning-accuracy-paradox-randy-lao (Accessed March 2, 2023).
18. Sebo, P (2022) How accurate are gender detection tools in predicting the gender for Chinese names? A study with 20,000 given names in Pinyin format. Journal of the Medical Library Association.
19. Jain, V., Enamorado, T., Rudin, C. (2022) The Importance of Being Ernest, Ekundayo, or Eswari: An Interpretable Machine Learning Approach to Name-Based Ethnicity Classification. Harvard Data Science Review.
20. Rudin, C., Radin, J. (2019) Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From an Explainable AI Competition. Available at: https://doi.org/10.1162/99608f92.5a8a3a3d (Accessed: March 2, 2023).
21. Sebo, P. (2022) NamSor's performance in predicting the country of origin and ethnicity of 90,000 researchers based on their first and last names.
22. Sasaki, Y. (2007) The truth of the F-measure. Toyota Technological Institute. Available at: https://www.toyota-ti.ac.jp/Lab/Denshi/COIN/people/yutaka.sasaki/index-e.html (Accessed: February 2, 2023).
23. Afonja, T. (2017) Accuracy paradox, Medium. Towards Data Science. Available at: https://towardsdatascience.com/accuracy-paradox-897a69e2dd9b (Accessed: February 2, 2023).
24. Equality Act 2010 c. 1. (2010) Available at: https://www.legislation.gov.uk/ukpga/2010/15/contents (Accessed: January 27th, 2023).

25. ONS (2023) Gender identity: Age and sex, England and Wales: Census 2021, Gender identity: age and sex, England and Wales - Office for National Statistics. Office for National Statistics. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/culturalidentity/genderidentity/articles/genderidentityageandsexenglandandwalescensus2021/2023-01-25 (Accessed: January 27th, 2023)
26. Morgan, A. (2022) Change over time in admin-based ethnicity statistics, England: 2016 to 2020, Change over time in admin-based ethnicity statistics, England: 2016 to 2020 - Office for National Statistics. Office for National Statistics. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/culturalidentity/ethnicity/articles/changeovertimeinadminbasedethnicitystatisticsengland2016to2020/2022-05-23 (Accessed: January 27, 2023).
27. Leslie, D. (2020) Understanding bias in facial recognition technology, The Alan Turing Institute. Available at: https://www.turing.ac.uk/sites/default/files/2020-10/understanding_bias_in_facial_recognition_technology.pdf (Accessed: February 16, 2023).
28. ICO (2012) Anonymisation: Managing data protection risk code of practice, ICO. Available at: https://ico.org.uk/media/1061/anonymisation-code.pdf (Accessed: February 1, 2023).
29. McIntosh, V. (2022) Understanding aggregate data, de-identified data &amp; anonymous data, Comparitech. Available at: https://www.comparitech.com/blog/information-security/aggregate-vs-anonymous-data/ (Accessed: February 1, 2023).
30. Samarati, P. and Sweeney, L. (1998) “Generalizing data to provide anonymity when disclosing information (abstract),” Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems - PODS '98 [Preprint]. Available at: https://doi.org/10.1145/275487.275508.
31. Table of de-identification techniques - San Jose State University (2019) Table of De-Identification Techniques. Available at: https://www.sjsu.edu/research/docs/irb-deidentification-techniques-table.pdf (Accessed: February 1, 2023).
32. ICO (2012) Anonymisation: Managing data protection risk code of practice, ICO. (Accessed: February 1, 2023).
33. Data Protection Act 2018. Available at: https://www.legislation.gov.uk/ukpga/2018/12/contents/enacted (Accessed: February 1, 2023).
34. IEE (2022) "IEEE Standard for Data Privacy Process," in IEEE Std 7002-2022. Available at: https://doi.org/10.1109/IEEESTD.2022.9760247. (Accessed: February 16, 2023).
35. Salman, H. et al. (2020) “Denoised Smoothing: A Provable Defense for Pretrained Classifier.” Available at: https://doi.org/ https://doi.org/10.48550/arXiv.2003.01908.
36. Shen, Y. et al. (2022) “Model stealing attacks against inductive graph Neural Networks,” 2022 IEEE Symposium on Security and Privacy (SP) [Preprint]. Available at: https://doi.org/10.1109/sp46214.2022.9833607.

A.1 Alternative Proxy Methods

Model	Type	Proxy	Notes
Agify.io, Genderize.io, Nationalize.io	Commercial tool	Name	Not enough performance information
Bias detection by using name disparity tables across protected groups	Academic paper	Name	Currently based on an American database
AGEify	Commercial tool	Image or ID	Similar to other tools considered
Predicting Twitter User Socioeconomic Attributes with Network and Language Information	Academic paper	Social media	Not available for testing or commercial use
Comparing Bayesian Improved Surname Geocoding to Machine Learning Methods	Academic paper	Name	Currently based on an American database
Predicting age groups of Twitter users based on language and metadata features	Academic paper	Social media	Similar to other tools considered
User-Level Race and Ethnicity Predictors from Twitter Text	Academic paper	Social media	Not available for testing or commercial use
Bayesian Improved Surname Geocoding (BISG)	Open source	Name	Currently based on an American database
Inferring User Gender from User Generated Visual Content on a Deep Semantic Space	Academic paper	Image	Similar to other tools considered
rethnicity	Open source	Name	Not available for testing or commercial use
Face ++	Commercial tool	Image	Not enough performance information
Microsoft Face API	Commercial tool	Image	Similar to other tools considered
Genderperformr	Academic paper	Name	Not enough performance information
What your username says about you	Academic paper	Email address or social media	Similar to other tools considered

Table 11 - Alternative proxy methods considered.

A.2 Weightings Justification

A.2.1 Upfront Cost

Upfront cost and training data: A higher upfront cost would be worth considering for a method which uses a large high quality, relevant training dataset. There would however reach a point of diminishing returns, where cheaper alternatives may still provide the accuracy needed, even with an inferior training dataset. The metrics were judged to be equal when compared to each other.

Upfront cost and accuracy: All methods considered require a high enough accuracy to be actually monitoring bias. A high-cost method would be worth it for significant gains in accuracy, especially gains for minority groups which are the most likely to be miscategorised for many methods. Accuracy is judged to be the more significant metric, with the upfront cost being weighted at 0.67 against accuracy.

Upfront cost and sustainability: A method with a very high sustainability would potentially last for a very long time. If a method can adapt to shifting demographics without needing to be significantly overhauled or replaced, this could easily justify high upfront costs when considering a method. The increased longevity of a sustainable model is judged to be more important than higher upfront costs, so costs are weighted at 0.67 against sustainability.

Upfront cost and maintainability: An easily maintained method would have much lower running costs than a method which required specialists to perform routine maintenance to keep it running. This means an easy to maintain method can easily justify a high upfront cost, as a hard to maintain method would likely be more expensive in the long run. The lower cost of an easily maintained model is likely to save more money in the long run compared to a low upfront cost, so upfront cost was weighted at 0.67 against maintainability.

Upfront cost and transparency: Many commercial tools will not provide full transparency on their methodology (e.g., release the source code online). The lack of full transparency for a method would make self-verification more difficult, along with trusting the methodology. A fully transparent tool is much easier to verify and trust, though is more susceptible to hacking. A lack of method transparency can be tolerated if the method has been verified by trusted sources. If this was the case, a high-cost method that has been verified and produces accurate results could be useful even if the full methodology is unknown, though this would still create an issue with public trust. Transparency is considered more important than upfront costs due to the additional public trust being more valuable than extra upfront costs. Cost is weighted at 0.67 against transparency.

Upfront cost and ease of implementation: An easy to implement method would be worth a high upfront cost, as a method that is difficult to implement it would require extra staff and more time to implement. This would likely be more expensive than paying an upfront fee to use an easy to implement solution. The upfront subscription costs of a model could very well be within the short-term costs associated with a hard to implement model, so the metrics are judged to be equal to each other.

Upfront cost and verification status: Results from a heavily verified method are easier to trust than a method that is only verified internally. An upfront cost would be worth the added security of knowing a method has been externally verified, as this reduces the possibility of the method not being as useful as advertised. Using a verified method could potentially be a long-term saving, as it is entirely possible a method's claims of accuracy don't hold up to scrutiny. Due to their being circumstances were both high cost and non-externally verified models can be considered suitable, the metrics were judged to be equal.

A.2.2 Training Data

Training data and accuracy: The quality and relevance of the training data is very important when judging the accuracy of a method. For example, if the model was trained and evaluated on US data and gives a high accuracy, this will not necessarily translate to a high accuracy in the UK, as the model may be overfitted to only work well in the US. It is important that any training data is either international or from the UK, as this would improve the accuracy and performance within the UK. Accuracy is deemed more important than the training data, as the training data is chosen specifically to improve the accuracy of a model. Training data is weighted at 0.67 against accuracy.

Training data and sustainability: A more up to date set of training data would help with method sustainability as the method should be able to stay relevant for small demographic shifts. However, if demographics change significantly, training data would become outdated, and would need updating/replacing. As both sustainable model and reliable training data are required for a long-term, accurate tool, they were judged to be equal.

Training data and maintainability: A method that is, or can be routinely maintained could be retrained and updated with new data. Being able to update the training data of a method makes it less important for the initial training data to be a perfect representation of an area being modelled. However, this means the model would need to be retrained and tested again before it could be used. As model longevity relies on both being able to maintain and provide high quality training data to a model, the metrics were judged to be equal.

Training data and transparency: The transparency of the method will help with verifying the method, along with understanding how training data is processed. High quality training data is necessary for identifying biases, but it is also important to know if the method is fairly processing the data. Being able to see the full method of a model, along with having high quality training data was judged to be equal.

Training data and ease of implementation: A high quality set of training data should help in creating an accurate model for bias detection. It is judged it would be worth increased effort being required to implement a model if the training data provided was of a high standard and relevant to the group being monitored, as this would produce more accurate results. Having a high-quality set of data was judged to be more important than the costs associated with a hard to implement model. Training data was weighted at 1.5 against ease of implementation.

Training data and verification status: It is complex task to measure the quality of training data or detect any biases it may have which could skew the results produced by a model. The verification status of a method would help in giving confidence in a method's accuracy, along with its training data. Being able to verify the high quality of a training data set was judged to be important, as a bias detection tool needs to use training data that is trusted to represent a group of people. Training data is weighted at 0.67 against verification status.

A.2.3 Accuracy

Accuracy and sustainability: A highly accurate method that is unsustainable would only be useful for short term use before demographic shifts cause the method to either need overhauling or replacing. On the other hand, a low accuracy yet sustainable method would not be useful at any point in time. It is important for a method to have a high level of accuracy and sustainability if the method is to be used for an extended period. The current and future accuracy of a model are judged to be as important as each other, so the accuracy and sustainability are considered to be equal.

Accuracy and maintainability: A highly accurate method would be worth a high level of maintenance if the method is expected to last for an extended period. A method that with a low maintenance level that produced low accuracy results would not be fit for purpose and would require replacing. High costs associated with a difficult to maintain model were judged to be worth the additional costs for a highly accurate model. Accuracy has a weighting of 1.5 against maintainability.

Accuracy and transparency: Method transparency helps with verification and confidence in a method, however if the method has a high accuracy (which has been verified), it could be worth the risk of trading knowledge of the exact methodology for highly accurate results. This however does heavily depend on being able to trust reported accuracy figures. Provided the method has been externally verified, the accuracy of a model was judged to be more important than full transparency of the method. Accuracy has a weighting of 1.5 against transparency.

Accuracy and ease of implementation: The ease of implementing a method will lower staffing costs associated with setup and will make it easier to exploit a method's potential, but if the method has a low accuracy, the saved costs would not be worth it. A high accuracy method is required to reliably monitor bias, so would be worth the extra effort required if the method was difficult to implement. It would be worth the extra difficulty and costs associated with a hard to implement model if the model had a high degree of accuracy. Accuracy has a weighting of 1.5 against ease of implementation.

Accuracy and verification status: Being able to trust the results of a bias monitoring algorithm is very important due to sensitivities around the monitoring of protected characteristics. A highly accurate method should produce trustworthy results, but it is important for those unfamiliar with the method to also be able to trust it. External verification would help with this, but the accuracy itself would be more important as this is a prerequisite for an external verification to give the okay on a method. As verification status increases trust in the accuracy measure of a method, they are judged to be equal against each other.

A.2.4 Sustainability

Sustainability and maintainability: If a method is not sustainable, its maintainability would not be very important as the method itself would become outdated and need either significant changes or replacement. The longevity of a method will depend on both its maintainability and sustainability, with maintainability being more relevant in the short term, and sustainability more important for the long term. A hard to maintain but sustainable model is a workable model, however an unsustainable yet easy to maintain model is not, due to the unsustainability lowering model accuracy. For this reason, sustainability is deemed more important and has a weighting of 1.5 against maintainability.

Sustainability and transparency: If a method was guaranteed to be highly sustainable, it would not be as important to have full access to the model, as one of the reasons to see the model in full is to see if the methodology will still apply in the future. It would be difficult to guarantee that a method was sustainable without access to the methodology, though high quality external verification would mitigate this risk. As public trust is also very important when considering the long-term benefits of a model, sustainability and transparency have been weighted equally against each other.

Sustainability and ease of implementation: If a method was highly sustainable, it would be much easier to accept a method that is hard to implement. This is because the method is likely to still be in use after a long time, so the effort required when initially implementing the method would not have to be repeated any time soon. For this reason, sustainability was deemed more important and given a weighting of 1.5 against implementation.

Sustainability and verification status: When considering long term use of a method, it is more important for a method to be sustainable than to have rigorous external verification, as changes in demographics can cause older verifications to become outdated, and the method would need to be verified again. In the short term, verification would be beneficial as it allows one to trust the information given about the method. When considering long term use of a model, sustainability is more important than verification status, so sustainability is weighted at 1.5 against verification status.

A.2.5 Maintainability

Maintainability and transparency: Self maintenance of a method would be much more difficult without access to the methodology. Commercial products are usually maintained, so method transparency is not required. Some open-source models are maintained by an online community and can also be maintained internally as the full methodology is known. Transparency makes it much easier for the public to trust the results produced by a method. Public trust and long-term use of the model are both necessary for a good tool, so maintainability and transparency have been weighted equally against each other.

Maintainability and ease of implementation: A hard to implement tool may require a lot of training to effectively maintain the tool, so would make the tool more expensive to keep running. Conversely, a tool that could easily be implemented but not well maintained could become outdated or susceptible to cyber-attacks, so would have a shorter lifespan. For these reasons the metrics have been weighted equally against each other.

Maintainability and verification status: As a tool is maintained and changes, this could affect it verification status by changing parts of the method (e.g., training data, accuracy measurements etc). A tool that has been verified allows one to trust a tool more readily, however maintainability allows one to extend the life of a method, which more useful. Maintainability has been weighted at 1.5 against verification status.

A.2.6 Transparency

Transparency and ease of implementation: For open-source tools, full knowledge of the methodology would help with implementation, but would not guarantee it would be easy. For commercial tools, they are often built with easy implementation in mind, but will not provide method transparency. Transparency of a method will help with trusting and understanding the model, which could be worth the additional costs for a difficult to implement model. The metrics have been weighted equally against each other.

Transparency and verification status: The main benefit of a transparent method is that it allows one to verify and understand the methodology used. The fully verified method can allow one to trust a method but would have a less thorough understanding of the method. It would also take less resources to trust an external verification. As both metrics help with understanding and trusting the model/methodology, they are weighted equally against each other.

A.2.7 Ease of Implementation

Ease of implementation and verification status: It is more important to have assurances in the methods efficacy then ease of use with the tool. Extra time taken to learn how to implement a method that is known to work would be a better use of resources on using a tool which might have currently unknown biases/inaccuracies. Some form of verification (either external verification or a transparent method so it can be internally verified) would be worth the extra resources required for a hard to implement method. For this reason, ease of implementation is weighted at 0.67 against verification status.

	Cost	Training data size	Accuracy	Sustainability	Maintainability	Method transparency	Ease of implementation	Verification status
Cost	1	1	0.67	0.67	0.67	0.67	1	1
Training data size	1	1	0.67	1	1	1	1.5	0.67
Accuracy	1.5	1.5	1	1	1.5	1.5	1.5	1
Sustainability	1.5	1	1	1	1.5	1	1.5	1.5
Maintainability	1.5	1	0.67	0.67	1	1	1	1.5
Method transparency	1.5	1	0.67	1	1	1	1	1
Ease of implementation	1	0.67	0.67	0.67	1	1	1	0.67
Verification status	1	1.5	1	0.67	0.67	1	1.5	1

Table 12 - Trade-offs for each metric.

FRAZER-NASH
CONSULTANCY
A KBR COMPANY

5th Floor Malt Building
Wilderspool Business Park
Greenalls Avenue
Warrington
WA4 6HL

Tel: +44 (0)1925 404000

fnc.co.uk

SYSTEMS ● ENGINEERING ● TECHNOLOGY