These checklists were originally provided as appendices in the final report for the Collaborative Data Project: Fair Algorithms. This project was part of the minor Data Wise (2020/2021) at the University of Groningen. The project was initiated by DataFryslân. During development of the checklists the team has collaborated with Statistics Netherlands (CBS).

Authors:

Alexandros Christopoulos	*S3742431*	*a.christopoulos.1@student.rug.nl*
Conrad Bauer	*S3679926*	*c.b.bauer@student.rug.nl*
Matthias Kooij	*S3391078*	*m.kooij.3@student.rug.nl*
Tjardo van der West	*S3599752*	*t.h.van.der.west@student.rug.nl*
Xueling Bai	*S3346951*	*x.bai.4@student.rug.nl*

Technical checklist

Start

This checklist focuses on increasing fairness of algorithms used in a context which involves decisions about people and is not meant to cover all areas relevant to a project involving algorithms (e.g. privacy or data management). Within this context, sensitive attributes can lead to unwanted discrimination, or unfairness. Algorithmic fairness is difficult to define, as different contexts warrant different definitions of the concept. In social contexts, algorithmic fairness is mostly used as "does not discriminate". While the goal of fair algorithms is to eliminate discrimination, perfect fairness is often impossible, but discrimination can still be mitigated as much as possible. Some aspects, such as tradeoffs between accuracy and fairness, have to be considered on a case by case basis. For each project, it is important to discuss terminology in the user group. See the glossary for definitions of terms used in this checklist.

Algorithms have the potential to cause various types of fairness-related harms. These include, but are not limited to, the following:

- Allocation of opportunities, resources, or information
- Failure to provide the same quality of service all people
- Reinforcement of existing societal stereotypes
- Denigration of people by being actively derogatory or offensive
- Over- or underrepresenting groups of people, or even treat them as if they do not exist

These harms were adapted from Madaio et al. (2020).

For whom is this checklist?
This checklist has to be applied to algorithms and machine learning practices which make decisions about people. While some questions are aimed specifically at machine learning (including deep learning and neural networks), the checklist is also suitable for simple rule-based algorithms. When you are not using sensitive attributes this checklist might not be relevant to your project.

For the best results, answer each question in the checklist. Some answers lead to follow up questions that should all be answered as well, while other answers do not require further information. Because the questions are relevant to different roles in a data project, we recommend that you complete this starter kit together with the team working on the algorithm. Input from the various roles is likely to benefit the discussion. Consider taking notes or minutes as they can be helpful when reviewing

your answers or communicating your results. After completing the checklist you should consider what aspects of your project need to be improved.

The original version of this checklist is part of the innovation project AI met Impact (AI with impact), and was compiled by Statistics Netherlands, municipalities, VNG and various partners. The project group "Fair Algorithms" translated and further developed this checklist with a focus on fairness during the Collaborative Data Project in the Minor Data Wise at the University of Groningen September 2020 - January 2021.

Technical

Exploration

The legitimacy and conditions of the project are central to the exploratory process. In this phase, it is determined, among other things, whether there is data of sufficient quality, what the risks are and whether the project complies with legislation and regulations. The director, the project manager, the data analyst as well as the privacy and security officer are involved in this process.

Context

1. Have you identified possible risks and benefits of creating/using this algorithm?
   Write down the identified risks and benefits to the different stakeholders. Check whether the risks have been addressed after the completion of this checklist.
   No
   Yes
   If no:
   Deliberate what the potential risks of this project are.
   If yes:
   Deliberate whether the potential benefits outweigh the risks associated with this project. Consider changing the project.

2. Are there reasons why this algorithm should not be used?
   Deliberate whether the risks are acceptable. Document your findings.
   No
   Yes
   If yes:

Deliberate whether the potential benefits outweigh the risks associated with this project. Consider changing the project before and after using the algorithm.

3. Have you mapped out the ethical aspects of the project, for example with the help of De Ethische Data Assistant (DEDA)?

   This is a toolkit which aids data analysts, project managers and policy makers to detect ethical issues in data projects, data management and data policies.

   No

   Yes

   If no:

   Consider the ethical aspects, such as the impact of the algorithm on citizens. You can find the Ethical Data Assistant here.

4. Is it clear what context this algorithm will be developed for and deployed in?

   It is important to decide and document for what context an algorithm is developed. If you are reusing the algorithm in a different context than originally intended, explain why this is appropriate.

   No

   Yes

   If no:

   Decide on the context of the algorithm and document your decisions.

5. Is there sufficient deployment available of the required roles for the project? For example the role of data engineer, privacy officer, data analyst, and project manager.

   No

   Yes

   If no:

   Create an overview of the required roles and skills, for all processes up to and including the use of the algorithm.

6. Are the responsibilities for internal and external communication clear for all roles in the project?

   No

   Yes

   If no:

   See if the municipality has a communication strategy, and use this as a basis for making a communication plan for the project.

   If yes:

   Check whether the communication plan complies with the municipality's communication strategy (if available).

7. Has a communication strategy been drawn up for communication about the use of the algorithm and the purpose and capabilities of the algorithm to (indirect) stakeholders?

   It is advisable to have simple communication in order for laymen to understand it as well. When making decisions about citizens, the organisation should be more transparent in their communication compared to, for example, an internal experiment in the organisation.

   No

   Yes

   If no:

   In the context of transparency, it is important to actively communicate about the use of algorithms. Consider how and when you will inform (indirect) stakeholders.

8. Have all relevant parties approved the project?

   For example, the municipality or partners of the project.

   No

   Yes

   If no:

   Ensure that there is sufficient support internally for the project and that the risks and bottlenecks concerning approval are identified.

   If yes:

   Describe the considerations of the parties that have not agreed so that risks and bottlenecks concerning approval are clearly identified.

9. Are there any parts of the project that do not follow a specific set of guidelines?

   No

   Yes

   If yes:

   Check whether there are guidelines available for that part of the project.

   If no:

   Consider documenting your actions and consider writing your own guidelines if necessary.

10. Is the code of the algorithm well documented in order for access and reviewing in the future?

    No

    Yes

    If no:

    Consider documenting the code well as an algorithm that is fair today might cause unwanted discrimination in the (far) future and may need

to be revised.

Laws and regulations

The legal aspects are described in detail in the policy tab.

1. Does the data contain special personal data? More information about special personal data in the General Data Protection Regulation (GDPR) can be found here.
   No
   Yes
   If yes:
   Consider why it is justifiable to use these data and consider the added value of these data on the explanatory power of the model. The justification must be a valid exception according to the GDPR.

2. Does the data contain variables that can lead to unwanted discrimination?
   No
   Yes
   If yes:
   Explore in what ways the algorithm can discriminate. Ensure a correct interpretation of the model. Investigate mitigating measures for this problem, such as model correction, counterfactual fairness, or the use of fairness metrics. More information about ways to discriminate can be found here (in Dutch).

3. Is there automatic decision-making, whether or not employing profiling?
   No
   Yes
   If no:
   Make sure that human intervention has a real impact on the result, otherwise, it will be regarded as automatic decision-making according to the GDPR.
   If yes:
   In principle, automatic decision-making is not permitted by the GDPR. Check whether this is an exceptional situation.

4. Has data minimization been applied?
   Data shall be adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed.
   No
   Yes
   If no:
   Reason why all the data are relevant, for example by creating a correlation matrix with the target variable or asking a domain expert.

Data inventory

1. Is metadata available for each dataset?
   This is important as otherwise it can be unclear what the data represents (e.g. sampling information) which can lead to unfairness.

   If no metadata standard is defined, use a commonly used standard such as the Dublin Core Metadata Initiative.

   No

   Yes

   If no:

   Contact the data supplier or make sure that a meta-data file is stored during the data collection.

   If yes:

   Make sure the metadata adheres to the standard and is up to date.

2. Is there a difference between the delimited population and the rest of the population that may cause unwanted discrimination? *Such differences may be regarding sensitive attributes as described in the GDPR. (see Glossary)*

   No

   Yes

   If no:

   Consider delimiting the population differently and argue why this population was chosen.

   If yes:

   Argue why there can be no unwanted discrimination.

3. What kind of data do you use?

   directory

   survey

   combination

   If directory:

   Are there any reasons why this register may have been filled in incorrectly? Could there be measurement errors, for example?

   No

   Yes

   If yes:

   Assess whether the quality of the data is sufficient. Correct for possible errors in the development phase.

   If survey:

   Is there interviewer bias (interviewer influence), response bias (socially desirable responses) or nonresponse bias (certain groups less response)?

   No

   Yes

If yes:

Assess whether the quality of the data is sufficient. Correct for possible errors in the development phase.

If combination:

Are there any reasons why this register may have been filled in incorrectly? Could there be measurement errors, for example?

No

Yes

If yes:

Assess whether the quality of the data is sufficient. Correct for possible errors in the development phase.

Is there interviewer bias (interviewer influence), response bias (socially desirable responses) or nonresponse bias (certain groups less response)?

No

Yes

If yes:

Assess whether the quality of the data is sufficient. Correct for possible errors in the development phase.

Is there enough overlap in the data so sources can be combined?

No

Yes

If no:

Is it possible to conduct the analysis with fewer sources?

Are there important differences between the quality of the sources?

No

Yes

If yes:

Document the difference in quality between the sources. Consider if it is necessary to approach the sources differently in the analytical phase.

4. Did you make use of an appropriate method of sampling that does not introduce bias in the data?

Avoid convenience samples if possible and consider cultural norms and biases that may have affected the data collection. For example, if you were to have a meeting about improving a workplace environment, cultural norms could mean that only those with strong complaints participate. This constitutes a bias against the workers who do not have strong enough complaints to participate, and subsequent changes may discriminate against those that did not participate in the meeting.

No

Yes

If no:

Reevaluate your data collection OR argue why the data can still be used and state explicitly what biases might be introduced in the data collection or why there is no bias to be expected.

5. Are there important factors for which no data is available?
   No
   Yes

   If yes:

   If there are variables that are expected to correlate with these factors, investigate methods of correction using these factors.

6. Are all categories of variables that could be considered discriminatory equally represented in the data?

   If you have different representations of variables your sample may be biased.

   No
   Yes
   If no:

   Collect more data, use transfer learning from, for example, another municipality, change the research framework or stop the research. If necessary: Complete the data with a sample from the population. Be aware of the shortcomings of previous research.
   If yes:

   Is there too little data or does the data contain too little variation? See the answer above. Also, check whether data augmentation is possible and appropriate or use methods such as bootstrapping.

7. Has the data been collected over multiple periods?
   No
   Yes

   If no:

   Were there explicit, successful efforts to ensure that the quality of the data is the same for the periods? Are there events or trends in the periods that may affect the data?
   If yes:

   If there are variables that correlate with these factors, treat the factors as a latent class.

8. Is the data recent?
   No
   Yes

   If no:

What changes (e.g. new legislation, change of neighborhood composition) are important? If there are major changes: correct for changes, remove the variables, or stop the study.

9. Have you taken into account the expiration date of the data?
   For how long can you store the data? Do you have a valid reason to keep it?
   No
   Yes
   If no:
   Check the legal retention period and ensure that the data is deleted in time.

10. What kind of variables does the data contain?
    Multiple answers are possible.
    Numerical
    Categorical
    If numerical:
    Has the accuracy been chosen correctly?
    If categorical:
    Are there categories with few observations? If so, possibly merge categories to avoid overfitting, make changes in the data collection, or stop the project.
    If both categorical and numerical:
    Has the accuracy been chosen correctly?
    Are there categories with few observations? If so, possibly merge categories to avoid overfitting, make changes in the data collection, or stop the project.

11. Is the data and the algorithm publicly accessible?
    It is ideal if both the data and the algorithm is accessible to those that are influenced by the decisions made using the data and/or algorithm.
    No
    Yes
    If no:
    Is it possible to make the data public? If not, ensure sufficient transparency about the algorithm.
    If yes:
    Be explicit and transparent about what data you used.

Development

Often, the development of an algorithm is iterative and starts with an exploration of the available data. When exploring data, it is important to consider one's own assumptions and biases and how they influence the interpretation of the data. Models

are trained and evaluated based on performance metrics, and these metrics must be critically evaluated regarding assumptions and biases as the data interpretation itself. Having an accurate model is important because having a fair model requires that the data can be accurately described. The data engineer and data analyst play a major role in this process.

Exploratory data analysis

Exploratory data analysis is a tool to determine where and when we might have biases and how to mitigate their effect on the outcomes as much as possible. By exploring the data, potential fairness issues can be discovered early in the model creation.

1. Are there similar cases or is there similar data which show a different outcome from your data?

   Check whether you are using data that shows a different outcome from a significant amount of related data. If you use this data you may come to wrong conclusions about your project as you could be using incomplete data. This can occur intentionally or unintentionally and is known as "cherry picking" based on the hypothetical that a farmer may only pick the nicest cherries from a tree and a buyer who only sees the selected fruit may wrongly conclude that the tree is in great condition. The full picture (cherry tree) may be wrongly represented by the data (picked cherries).

   No

   Yes

   If no:

   Try to find out if there is a general trend within similar data and whether this allows you further conclusions.

   If yes:

   Try to find out why your data/outcomes differ(s) from other data. Make sure that you are comparing to similar data and that you do not have systemic bias in your data.

2. Have you exhaustively searched (brute-force) your data for significant correlations?

   This may be problematic as it increases the risk for false positives. If you search for correlations with a significance level of .05, you will get a significant result after every 20th try (on average).

   No

   Yes

   If yes:

   Consider the tests that gave you nonsignificant results. Did you expect different results? You may have to collect additional data to test the findings and make sure that they were not based on chance.

3. Is there any data which could be relevant to your project that you do not have data or are not taking into account?

   In order to get the best results you should ask yourself if there is relevant data which you do not have. If you make your decision without this data your results may be influenced by survivorship bias. Survivorship bias is the logical error of focusing only on the part of the data which made it past a selection process while disregarding the data which did not.

   No

   Yes

   If yes:

   Try to find and/or incorporate the data in your model. Otherwise argue why you are not using the data.

4. Are you making conclusions about causal relationships? (A causes B)

   Third variables can explain the relationship. Also, it is difficult to conclude on the direction of causality. This is related to other fallacies: If you have a lot of data you will find more relationships, but it is not always possible to conclude in a causal direction.

   No

   Yes

   If yes:

   Make sure that the background information needed about the causal model and the context in which it is obtained are reasonable. Try to eliminate plausible alternatives and establish the time order of the relation. In other words, make sure that there is no third variable which explains the relation and establish that A precedes B.

   If no:

   Subgroup analysis can be used to analyze which subgroups were treated most (un)fairly and to distinguish the type of bias exhibited. Alternatively, a stochastic approach can be used to provide transparency about how (classification) decisions are made. Directed acyclic graphs (DAGs) are a common means to represent conditional independence assumptions between variables.

5. Did you separate the data to look for (a) strong correlation(s) between variables?

   Often, relationships in part of the data change when all data is aggregated. This is known as Simpson's paradox. A famous example are cases of alleged gender bias in university admissions based on the lower overall admission rates of women. When taking into account that women and men tended to apply to different departments, it was concluded that women tended to apply to more competitive departments with overall lower admissions (also for men).

No
Yes
If yes:
Make sure to select a sample that is diverse in features and *sensitive attributes equally weighted.* Check if the observed relations change when you aggregate the data.

6. Are observations in certain variables missing or are there duplicates?
   No
   Yes
   If yes:
   If there are not too many and it does not appear in a certain group: imputate the missing observations. If there are many, do not use the variable.

7. Are there any abnormal observations (outliers)?
   Outliers can affect the outcomes of models, particularly least squares regression models.
   No
   Yes
   If no:
   Does the data contain enough variety? Has outlier detection been applied to the data, in which several variables were also combined?
   If yes:
   Consider the reason(s) for the outliers. If these observations are caused by an error, such as a measurement error or a typing error, correct or delete the observations. Keep also in mind that the outlier might exist due to discrimination. Investigate whether the model can handle abnormal observations.

Model selection

1. Are there labels available?
   No
   Yes
   If no:
   Choose unsupervised methods, or label the data.
   If yes:
   How did the labels come about? What is the quality of the labels, what risks are there?

2. Is the robustness of the model sufficient?
   Does the model have the sufficient ability to resist noise? The sufficient level of robustness should be determined for the specific project.

No
Yes
If no:
Try to improve the model's ability to resist noise. See this and this website for tips.

3. Is numerical feature transformation used?
   No
   Yes
   If no:
   Numerical features can be transformed using methods such as normalization, min-max normalization, logistic transformation, and others.
   If yes:
   Make sure that the relations between the transformed data and other attributes are not broken.

4. Is importance reweighting used in the classification?
   No
   Yes
   If no:
   Consider to indicate a frequency count for an instance type to place lower/higher importance on "sensitive" training samples.
   If yes:
   Make sure of the stability and robustness of the classification.

5. Is the task of the model classification or regression?
   Other classification includes multi-class, multi-label or hierarchical models.
   Binary classification
   Other classification
   Regression
   If binary classification:
   Are false positives as important as false negatives in the context of the problem?
   If other classification:
   In the context of the problem, are there one or more classes that are more important to estimate right or wrong than other classes.

6. Have the assumptions of the model been met? Such as the distribution of errors, or independence of features?
   No
   Yes

If no:
Change your model OR describe why you still choose this model, and note which assumptions have not been met.

7. Are there latent variables?
   No
   Yes
   If yes:
   Has a model been chosen that takes this into account? Describe the effect of the latent variables on the data and the outcome of the model as accurately as possible.

8. Have the assumptions about the data been sufficiently substantiated and described?
   No
   Yes
      If no:
      Argue whether these assumptions are reasonable and necessary. There
   may be an alternative model that requires fewer assumptions.
   If yes:
   Try to validate the assumptions as much as possible, with data or via a domain expert.

9. Is the model selection justified?
   Have the arguments for choosing a particular model been reported sufficiently and in an understandable way, so that a colleague without a technical background understands it?
   No
   Yes
      If no:
   Explain the choices in understandable language. A diagram or figure can be supportive. (such as the ROC curve)

10. Is the model selection well documented?
    Does the documentation match the background and knowledge of the person who has to make the choice about the model? Has the choice of model been summarized briefly, concisely and clearly? This is important for explainability.
    No
    Yes
    If no:

Adjust the documentation so that it is understandable. Make sure that the person who has to decide on the model has sufficient knowledge, for example by giving a presentation about the model.

Model Performance

Assessing model performance is an important part of any data project. To ensure fairness it is important to go beyond standard performance metrics and also take fairness metrics into account.

1. Do you use frequently used performance metrics?
   Many performance metrics focus only on efficiency, but it is also important to evaluate how to assess fairness.
   No
   Yes
   If no:
   Can you justify why you chose this specific metric?

2. Do you use a measure of fairness?
   Using a fairness metric will help you to assess how fair your model is.
   No
   Yes
   If no:
   Make use of a measure of fairness such as the FairTrade method.
   If yes:
   Make sure you state why you chose this specific measure of fairness.

3. Is unsupervised machine learning being used?
   No
   Yes
   If yes:
   Pay enough attention to the validation of the model. Investigate whether the validation is
   sufficient for the impact of the model. Use an appropriate evaluation metric.

4. Is the uncertainty of the prediction / classification known?
   Calculating and sharing the uncertainty of the analyses is important for transparency. *Communicating uncertainty can increase trust in evidence-based decisions in the long term and is a form of transparency.*
   No
   Yes
   If no:

For example, use Monte Carlo methods to get an estimate of the
Uncertainty.

5. Is there over- or underfitting?
   Make sure that your model performs well both on the training data and other data.
   No
   Yes
   If yes:
   Does the model suit the task? Possibly apply regularization, change models or use resampling methods such as cross-validation or bootstrapping, or apply an early stopping criterion.

6. Have you compared the performance of the algorithm with other, simpler algorithms?
   In principle, a simpler model is preferable for the same performance on measures of accuracy and fairness. Simple models have higher computational efficiency, are explained more easily (increasing transparency), prevent overfitting (beware of underfitting) and can be used to compare more complex models later on.
   No
   Yes
   If no:
   Compare the results of different models and justify which model you want to use.
   If yes:
   Note whether the performance differs significantly from the other
   Models.

7. Has hyperparameter tuning been addressed?
   Addressing fairness during hyperparameter tuning can lead to greater fairness while preserving other performance measures.
   No
   Yes
   If no:
   Analyze and note the influence of the hyperparameters on the outcomes of the model.
   If yes:
   Describe how you arrive at the choice for these hyperparameters.

8. Has a sensitivity analysis been carried out?

A sensitivity analysis describes how various aspects of the feature vector affect a given outcome.

No

Yes

If no:

Perform a sensitivity analysis to map how different values of independent variables influence the dependent variable, given the assumptions of the model.

If yes:

Ensure that sensitivity analyses can better help understand uncertainty

about fairness.

9. Is Blinding used in some common classifier models?

Blinding is the approach of making a classifier "immune" to one or more sensitive variables.

No

Yes

If no:

Make sure you seek to train a race blind classifier (among others) in that each of race groups have the same threshold value, i.e. the provided loan

rate is equal for all races

If yes:

Make sure you are making a trade-off between blindness and model performance.

Implementation

After a proof of concept is available and approved, the algorithm can be implemented. In this process, the model is applied, the research questions from the exploration are reflected on and the model is transferred to the executive team. Good communication between the data analyst and the executive department is important.

Apply model

1. Is interpretation of the model desirable or necessary?

No

Yes

If no:

Explain why this is less important for this application.
If yes:
Explain what the model has learnt. Does this agree with the hypotheses, why not or why? Make sure you can explain the model to laymen.

2. Does the model explain the data well enough so that the results can be used for the intended purpose?
   No
   Yes
   If no:
   Describe the model fit in relation to the use of the algorithm in practice

3. Have statements about the data been statistically tested? An example of a statement is: 'Residents who follow VWO are more likely to drop out of school early'
   No
   Yes
   If yes:
   Have the assumptions of the test been met? Has a parametric or has a non-parametric test been used?

4. Has the algorithm been audited by internal or external auditors?
   No
   Yes
   If no:
   Consider organizing such audits. Often, the source of unfairness does not lie within but outside the code. The auditor should look closely at the exact outcomes of the algorithm to determine if discrimination exists.
   If yes:
   Has auditing been committed both internally and externally? It is possible to audit the outcomes of an algorithm without having access to the algorithm by external parties.

Reflection

1. Has a pilot taken place?
   By testing your method, potential issues can be addressed beforehand.
   No
   Yes
   If no:
   Describe the consequences if the algorithm fails or is not positively assessed by the administration or citizens. Consider running a pilot.

2. Has a Social Impact Statement been performed and published together with the algorithm?

   This aids data scientists and policy makers on topics as responsibility, explainability, accuracy, auditability and fairness during the design stage, pre-launch and post-launch of an algorithm. Check the website.

   No
   Yes

   If no:
   Consider performing it and publish it together with the algorithm.

Use

This process consists of the usage of the algorithm for the first time as well as and regular checks of the data and the algorithm. External communication plays a major role in this process. Depending on the feedback on the use, everyone involved can play a role in this process.

1. Are you using the algorithm for the first time?
   If a pilot study was conducted take it into account.
   No
   Yes

   If yes:
   1.1 Has the algorithm been applied to the target group that was predefined?

   No
   Yes

   If no:
   Justify why the algorithm was applied to a different target group and investigate the effect this has on the quality of the model.

   1.2. Is there sufficient feedback from (indirect) stakeholders on the project?
   No
   Yes

   If no:
   Investigate the cause of this. Are the feedback channels accessible enough for those involved and is the feedback processed in a timely manner?
   1.3. Is there a built-in possibility for data subjects to object to the use of or decisions made by the algorithm?
   No

Yes

If no:

Make sure there is a built-in option to process an objection.

2. Do you expect the context and environment of the algorithm to change over time?

No

Yes

If no:

Write down your findings and repeat this question regularly.

If yes:

Investigate how the changes affect the impact of the algorithm. Ask: Can there be new / unforeseen causes of unwanted discrimination? Can the algorithm be adjusted or does the algorithm need to be retrained?

3. Do those involved have the opportunity to request information and provide feedback?

No

Yes

If no:

Openness and transparency are important as they promote accountability and open a channel of communication between developers and society.

Ensure that data subjects can pass on feedback and information and process this information accurately.

If yes:

Make sure that the questions and comments of those involved are

processed timely and accurately.

4. Is the system improved on a regular basis through feedback from all stakeholders?

No

Yes

If no:

Regularly improve the algorithm based on the feedback.

5. Does the system contain feedback loops?

A feedback loop is when the output of the model affects the input of the model at a later time, thus creating a self-amplifying effect. A common example of this are recommendation algorithms: A user may get an algorithmically computed recommendation (e.g. for buying an object), follows the recommendation (buys the object), and this action is then fed back to the

algorithm and influences future buying recommendations. Future recommendations may focus too much on this action while underrepresenting other interests. In situations where there is unfairness, the impact of the initial unfairness can become worse.

No
Yes

If yes:

Regularly examine how representative the data is. Use a sample from the population in addition to output data from the model if possible.

6. Is the data quality stable over time?
Consult with your data scientist.
No
Yes

If no:

Investigate the effect of the change in data quality on the output of the algorithm. Determine whether the data quality is sufficient to continue using the algorithm.

7. Does the project (still) comply with current guidelines?
No
Yes

If no:

Make adjustments where needed.

8. Is there regular communication about the use of algorithms?
No
Yes

If no:

Increase transparency by regularly communicating about the use of algorithms, including why and how they are used.

9. Are there long term plans to evaluate the working of the algorithm?
This includes measures that ensure the realization of these plans.

No
Yes

If no:

Consider that part of society that is influenced by your algorithm. Most likely this part of society is not static but subject to changes. If the working of your algorithm is not evaluated in the reasonably far future and remains static, it could produce unwanted discrimination that is never detected.

Policy checklist

Appendix B contains a checklist that mainly refers to policy makers. This checklist's questions are aiming at informing policy makers about the related laws and regulations revolving around algorithmic development.

Policy

This part of the checklist contains some questions specifically for policymakers / decision-makers. It is advisable for the technical colleagues to also be involved as much as possible in answering these questions. Similarly, we advise policymakers to be involved in answering the technical questions and to try to understand them as best they can.

Context

1. Have you identified possible risks and benefits of creating/using this algorithm?

   Write down the identified risks and benefits to the different stakeholders. Check whether the risks have been addressed after the completion of this checklist.

   No

   Yes

   If no:

   Deliberate what the potential risks of this project are.

   If yes:

   Deliberate whether the potential benefits outweigh the risks associated with this project. Consider changing the project.

2. Are there reasons why this algorithm should not be used?

   Deliberate whether the risks are acceptable. Document your findings.

   No

Yes

If yes:

Deliberate whether the potential benefits outweigh the risks associated with this project. Consider changing the project before and after using the algorithm.

3. Have you mapped out the ethical aspects of the project, for example with the help of De Ethische Data Assistant (DEDA)?

   This is a toolkit which aids data analysts, project managers and policymakers to detect ethical issues in data projects, data management and data policies.

   No

   Yes

   If no:

   Consider the ethical aspects, such as the impact of the algorithm on citizens. You can find the Ethical Data Assistant here.

4. Is it clear what context this algorithm will be developed for and deployed in?

   It is important to decide and document for what context an algorithm is developed. If you are reusing the algorithm in a different context than originally intended, explain why this is appropriate.

   No

   Yes

   If no:

   Decide on the context of the algorithm and document your decisions.

5. Is there sufficient deployment available of the required roles for the project? *For example the role of data engineer, privacy officer, data analyst, and project manager.*

   No

   Yes

   If no:

   Create an overview of the required roles and skills, for all processes up to and including the use of the algorithm.

6. Are the responsibilities for internal and external communication clear for all roles in the project?

   No

   Yes

   If no:

   See if the municipality has a communication strategy, and use this as a basis for making a communication plan for the project.

   If yes:

Check whether the communication plan complies with the municipality's communication strategy (if available).

7. Has a communication strategy been drawn up for communication about the use of the algorithm and the purpose and capabilities of the algorithm to (indirect) stakeholders?

   It is advisable to have simple communication in order for laymen to understand it as well. When making decisions about citizens, the organisation should be more transparent in their communication compared to an internal experiment in the organisation.

   No

   Yes

   If no:

   In the context of transparency, it is important to actively communicate about the use of algorithms. Consider how and when you will inform (indirect) stakeholders.

8. Have all relevant parties approved the creation & use of the algorithm?

   For example, the municipality or partners of the project.

   No

   Yes

   If no:

   Ensure that there is sufficient support internally for the project and that the risks and bottlenecks concerning approval are identified.

   If yes:

   Describe the considerations of the parties that have not agreed so that risks and bottlenecks concerning approval are clearly identified.

9. Are there any parts of the project that do not follow a specific set of guidelines?

   No

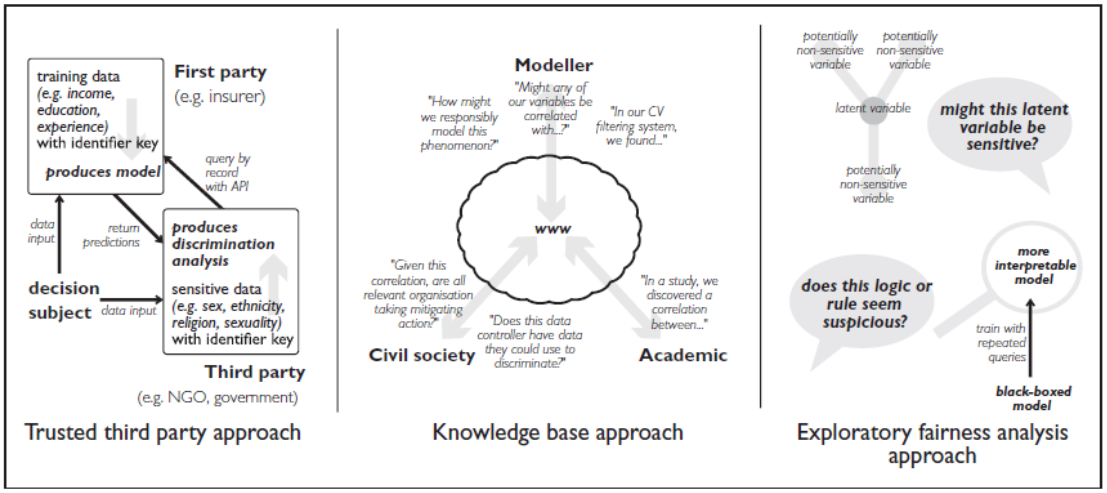
   Yes

   If yes:

   Check whether there are guidelines available for that part of the project.

   If not, consider documenting your actions and writing your own guidelines if necessary.

10. Are you currently applying any of the three approaches described in Veale and Binns (2017) in your project?



The first approach is about having another organization as a third party which possesses the sensitive data. This third party tries to detect discrimination.

The second approach is about sharing knowledge about fair algorithms on a platform by civil society, data scientists and academics.

The third approach is about asking questions after analyzing possible latent variables and the logic of or a rule in an algorithm.

No

Yes

If no:

The authors offered a method to mitigate discrimination without collecting sensitive data. Consider reading the article and check if there is anything useful for your project.

Legislation and regulations

General Administrative Law Act (Algemene wet bestuursrecht, AWB)

The Advisory Division of the Council of State recommends a sharper interpretation of the general principles of good governance with a view to digitization. Next to the laws listed below they also pertain to transparency. In the preparation of each decision (including decisions based on algorithms) especially the following articles should be taken into account:

- Equality principle, art. 1 Gw. The government must treat similar cases in the same manner.
- Fair-play-principle, art. 2:4 Awb. The Fair-play-principle provides a fair treatment of a decision.
- Carefulness principle, art. 3:2 Awb. The government should prepare and take a decision carefully.

- Détournement de pouvoir, art. 3:3 Awb. The administrative authority shall not use the power to take a decision for any other purpose without permission.
- Proportionality principle, art. 3:4 par. 2 Awb. The government must ensure that the burdens or adverse consequences of a government decision for a citizen do not outweigh the general interest of the decision.
- Motivation principle, art. 3:46 Awb. The government should motivate its decisions well.

Public Access to Government Information Act (Wet openbaarheid bestuur, WOB)

This law can be relevant for algorithmic fairness when, for instance, citizens suspect unfairness and file a WOB request (they demand transparency). It is therefore important to be able to explain the decision process of your algorithm as best as possible.

1. Is your project part of a governmental organization?
   No
   Yes
   If no:
   The General Administrative Law Act and the Public Access to Government Information Act do not apply to non-government organisations.
   If yes:
   The General Administrative Law Act and the Public Access to Government Information Act apply to your project. This means you should be ready to provide information to the public about all parts of your project. Consider if your project complies with the act, e.g. détournement de pouvoir.

2. Has privacy been taken into account, for example by carrying out a Data Protection Impact Assessment (DPIA) to identify privacy risks?
   No
   Yes
   If no:
   The AVG states that a DPIA must in any case be performed if:
   The algorithm systematically and comprehensively evaluates personal aspects based on automated processing, including profiling, and makes decisions that affect people;
   Processes special personal data or processes criminal data on a large scale;

Widely and systematically follows people in a publicly accessible area (for example with camera surveillance).

More information about DPIA can be found here (in Dutch). DPIA is also referred to as Privacy Impact Assessment (PIA) or data impact assessment (Dutch gegevenseffectbeoordeling; GEB).

GDPR

3. Is the algorithm processing personal data in any way?
   No
   Yes

   If no:
   The GDPR does not apply.
   If yes:
   The GDPR applies to your project. Since its introduction, most focus has gone to the privacy protection parts of the GDPR. Fairness is however also a key principle of the GDPR. There is growing attention for fairness in the legal literature and public debates, which is why we expect this principle to become a more important part of the GDPR. This means that, in order to comply with the GDPR, your project should not only protect subjects' privacy but also avoid unfairness as best it can.

4. Have appropriate technical and organizational measures been taken to be able to demonstrate that the processing is carried out in accordance with the GDPR? (Art. 24 and 25 GDPR)

   No
   Yes

   If no:
   Make sure appropriate measures are taken.

5. Does the data contain personal data?

   More information about special personal data in the General Data Protection Regulation (GDPR) can be found here.

   No
   Yes

   If yes:
   Consider why it is justifiable to use these data and consider the added value of these data on the explanatory power of the model. The justification must be a valid exception according to the GDPR.

6. Are you able to detect and deal with unwanted discrimination?

Discuss this with your data scientist.

Information about ways to discriminate can be found here (in Dutch).

No

Yes

If yes:

Discuss with your data scientist how this can be prevented.

7. Is there automatic decision-making, whether or not employing profiling? (Art.22 Par.1 GDPR) *Discuss this with your data scientist.*

   No

   Yes

   If no:

   Make sure that human intervention has a real impact on the result, otherwise, it will be regarded as automatic decision-making according to the GDPR.

   If yes:

   In principle, automatic decision-making is not permitted by the GDPR. Check whether this is an exceptional situation.

Data inventory

1. Is the data recent?

   No

   Yes

   If no:

   What changes (e.g. new legislation, change of neighbourhood composition) are important? If there are major changes: correct for changes, remove the variables, or stop the study.

2. Have you taken into account the expiration date of the data?

   For how long can you store the data? Do you have a valid reason to keep it?

   No

   Yes

   If no:

   Check the legal retention period and ensure that the data is deleted in time.

3. Is the data and the algorithm publicly accessible?

   It is ideal if both the data and the algorithm is accessible to those that are influenced by the decisions made using the data and/or algorithm.

   No

   Yes

   If no:

Is it possible to make the data public? If not, ensure sufficient transparency about the algorithm.
If yes:
Be explicit and transparent about what data you used.

Model selection & performance

1. Is the model selection justified and well documented?
   Has the documentation for choosing a particular model been reported briefly, concisely and in an understandable way? Does the documentation match the background and knowledge of the person who has to make the choice about the model?
   No
   Yes

   If no:
   Discuss this with your data scientist. Adjust the documentation so that it explains the choices in understandable language. Make sure that the person who has to decide on the model has sufficient knowledge.

2. Do you understand the used performance metrics, including measures of fairness?
   Many performance metrics focus only on efficiency, but it is also important to evaluate how to assess fairness. Using a fairness metric will help you to assess how fair your model is.
   No
   Yes

   If no:
   Discuss the performance metrics with your data scientist. Make sure that it is clear why these specific measures were chosen.

Implementation

After a proof of concept is available and approved, the algorithm can be implemented. In this process, the model is applied, the research questions from the exploration are reflected on and the model is transferred to the executive team. Good communication between the data analyst and the executive department is important.

1. Has the algorithm been audited by internal or external auditors?
   No
   Yes

   If no:
   Consider organizing such audits. Often, the source of unfairness does not lie within but outside the code. The auditor should look closely at the exact outcomes of the algorithm to determine if

discrimination exists.
If yes:
Has auditing been committed both internally and externally? It is possible to audit the outcomes of an algorithm without having access to the algorithm by external parties.

2. Has a pilot taken place?
By testing your method, potential issues can be addressed beforehand.
No
Yes

If no:
Describe the consequences if the algorithm fails or is not positively assessed by the administration or citizens. Consider running a pilot.

3. Has a Social Impact Statement been performed and published together with the algorithm?

This aids data scientists and policymakers on topics as responsibility, explainability, accuracy, auditability and fairness during the design stage, pre-launch and post-launch of an algorithm. Check the website.

No
Yes

If no:
Consider performing it and publish it together with the algorithm.

Use

This process consists of the usage of the algorithm for the first time, and regular checks of the data and the algorithm. External communication plays a major role in this process. Depending on the feedback on the use, everyone involved can play a role in this process.

1. Are you using the algorithm for the first time?
If a pilot study was conducted take it into account.
No
Yes

If yes:
1.1 Has the algorithm been applied to the target group that was predefined?
No
Yes

If no:

Justify why the algorithm was applied to a different target group and investigate the effect this has on the quality of the model.

1.2. Is there (sufficient) feedback from (in-)direct stakeholders on the project?
No
Yes

If no:
Investigate the cause of this. Are the feedback channels accessible enough for those involved and is the feedback processed in a timely manner?

1.3. Is there a built-in possibility for data subjects to object to the use of or decisions made by the algorithm?
No
Yes

If no:
Make sure there is a built-in option to process an objection.

2. Do you expect the context and environment of the algorithm to change over time?
No
Yes

If no:
Write down your findings and repeat this question regularly.

If yes:
Investigate how the changes affect the impact of the algorithm. Ask: Can there be new / unforeseen causes of unwanted discrimination? Can the algorithm be adjusted or does the algorithm need to be retrained?

3. Do those involved have the opportunity to request information and provide feedback?
No
Yes

If no:
Openness and transparency are important as they promote accountability and open a channel of communication between developers and society.
Ensure that data subjects can pass on feedback and information and process this information accurately.

If yes:

Make sure that the questions and comments of those involved are processed timely and accurately.

4. Is the system improved on a regular basis through feedback from all stakeholders?

No

Yes

If no:

Regularly improve the algorithm based on the feedback.

5. Does the system contain feedback loops?

A feedback loop is when the output of the model affects the input of the model at a later time, thus creating a self-amplifying effect. A common example of this are recommendation algorithms: A user may get an algorithmically computed recommendation (e.g. for buying an object), follows the recommendation (buys the object), and this action is then fed back to the algorithm and influences future buying recommendations. Future recommendations may focus too much on this action while underrepresenting other interests.

No

Yes

If yes:

Regularly examine how representative the data is. Use a sample from the population in addition to output data from the model if possible.

6. Is the data quality stable over time?

Consult with your data scientist.

No

Yes

If no:

Investigate the effect of the change in data quality on the output of the algorithm. Determine whether the data quality is sufficient to continue using the algorithm.

7. Does the project (still) comply with current guidelines?

No

Yes

If no:

Make adjustments where needed.

8. Is there regular communication about the use of algorithms?

No

Yes

If no:

Increase transparency by regularly communicating about the use of algorithms, including why and how they are used.

9. Are there long term plans to evaluate the working of the algorithm?
This includes measures that ensure the realization of these plans.

No

Yes

If no:

Consider that part of society that is influenced by your algorithm. Most likely this part of society is not static but subject to changes. If the working of your algorithm is not evaluated in the reasonably far future and remains static, it could produce unwanted discrimination that is never detected.

Glossary

Algorithm	Automated steps / instructions that convert input data into an intended output or to perform a particular task
Algorithmic Fairness	A fair algorithm is an algorithm which prevents unwanted discrimination from happening or where unwanted discrimination does not exist.
Artificial Intelligence (AI)	The science and development of machines with competences that can be considered intelligent according to the standard human intelligence.
Auditing	An internal or external inspection of the algorithm, the input and output.
Blinding	Blinding is the approach of making a classifier “immune” to one or more sensitive variables. A classifier is, for example, race-blind if there is no observable outcome differentiation based on the variable race. Blinding (or partial blinding) has also been used as a fairness audit mechanism. Specifically, such approaches explore how partially blinding features (sensitive or otherwise) affect model performance. This is similar to the idea of causal models and can help identify problematic sensitive or proxy variables with black-box-like analysis of an ML model.
Categorical Variables	A categorical variable is one that has two or more categories. For example, gender is a categorical variable having two categories (male and female).
Causal Relationships	A causal relation between two events exists if the occurrence of the first causes the other.
Classification vs Regression	Classification is about predicting a label and regression is about predicting a quantity.
Convenience Sampling	Convenience sampling is a type of nonprobability sampling in which people are sampled simply because they are "convenient" sources of data for researchers.
Data minimization	Keeping the data adequate, relevant and limited to what is necessary.

Data Protection Impact Assessment (DPIA)	A DPIA is a process designed to help you systematically analyse, identify and minimise the data protection risks of a project or plan described in the GDPR.
De Ethische Data Assistant (DEDA)	A toolkit which aids data analysts, project managers and policymakers to detect ethical issues in data projects, data management and data policies.
Exploratory Data Analysis	Exploratory data analysis is a tool to determine where and when we might have biases and how to mitigate their effect on the outcomes as much as possible.
Fairness	Just treatment or behaviour without discrimination.
Hyperparameter tuning	Hyperparameters are parameters whose values are set prior to the commencement of the learning process. By contrast, the value of other parameters is derived via training.
Latent variables	A latent variable is a variable that cannot be observed. The presence of latent variables, however, can be detected by their effects on variables that are observable.
Metadata	A set of data that describes and gives information about other data.
Numerical feature transformation	A function that transforms features from one representation to another. ... feature values may cause problems during the learning process, e.g. data represented in different scales.
Numerical Variables	A numerical variable is a data variable that takes on any value within a finite or infinite interval (e.g. length, test scores, etc.)
Outliers	An outlier is an observation that lies an abnormal distance from other values in a random sample from a population.
Over and underfitting	Overfitting: Good performance on the training data, poor generalization to other data. Underfitting: Poor performance on the training data and poor generalization to other data.
Profiling	Any form of automated processing of personal data whereby certain personal aspects of a natural person are evaluated on the basis of personal data, in particular with the aim of analyzing or predicting his professional performance, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements.

Reweighting in Classification	Tackle the label noise classification problem by importance reweighting. Liu et al.(2016) formulated this problem and proved a consistency guarantee that the learned classifier using importance reweighting would converge to the optimal classifier in the noise-free case for any surrogate loss function.
Sensitive Attributes	Variables sensitive to discrimination. Race, ethnicity, religion or belief, nationality, gender, sexuality, disability, marital status, genetic features, language, age, and to a degree nationality.
Sensitivity analysis	A sensitivity analysis determines how different values of an independent variable affect a particular dependent variable under a given set of assumptions.
Systemic bias	Systemic bias comes from the way that data is created
Unsupervised machine learning	In a supervised learning model, the algorithm learns on a labelled dataset, providing an answer key that the algorithm can use to evaluate its accuracy on training data. An unsupervised model, in contrast, provides unlabeled data that the algorithm tries to make sense of by extracting features and patterns on its own.
Unwanted discrimination	Making an unjustified difference in the treatment of people on the basis of a sensitive attribute from the law.