Bias Self-Assessment

A document to help teams identify sources of bias during the project lifecycle

Instructions

The following document introduces a set of social, statistical, and cognitive biases that may occur throughout the activities and stages of a project's lifecycle. These biases require ongoing reflection and deliberation to minimise the possible negative impact upon downstream activities or the risk of discriminatory outcomes.

For each bias, there is information about the following:

- Lifecycle scope: the range of stages and activities in the project lifecycle that are likely to be impacted by the respective bias
- Significant stage(s): those stages and activities where intervention is most likely to have an impact in mitigating the respective bias

The description of the bias also provides illustrative examples to help you and your team reflect upon how the respective bias may affect your own work.

This self-assessment is not intended as a checklist that needs to be completed all at once. Rather, it is something you should familiarise yourself with and return to for reference during key stages of bias mitigation (e.g., initial project planning).

List of Biases

Legend Bias Categories:

Social Bias

Social bias has to do with the way that pre-existing or historical patterns of discrimination and social injustice—and the prejudices and discriminatory attitudes that correspond to such patterns—can be drawn into the AI lifecycle. In particular, it relates to how these patterns and attitudes can be perpetuated, reinforced, or exacerbated through the development and deployment of data-driven technologies.

Statistical Bias

Statistical bias refers to a systematic deviation from an expected statistical result that arises due to the influence of some additional factor. This understanding is common in observational studies where bias can arise in the process of sampling or measurement. Statistical biases can involve errors (deviations from a true state) or differences between measured or calculated values and true values.

Cognitive Bias

Cognitive bias refers to a systematic deviation from a norm of rationality that can occur in processes of thinking or judgement and that can lead to mental errors, misinterpretations of information, or flawed patterns of response to decision problems.

Historical Bias

Historical biases exist prior to the inception of any AI project, and they can exist even where data are responsibly sampled, collected, and processed. They arise in AI innovation contexts when there is a gap or misalignment between the state of the world and the goals or objectives of the project and system being developed. Such a gap allows for historical patterns of inequity or discrimination to be reproduced, or even augmented, in the development and use of the system even when the system is functioning to a high standard of accuracy and reliability. For instance, even with perfect sampling and feature selection, a project will exhibit bias where it perpetuates (or exacerbates) socioeconomic inequalities through the outcomes it promotes, or the deployment of the system being developed. For this reason, correcting historical biases fall under the remit of Application Fairness.

Lifecycle Scope:	Pre-exists lifecycle
Significant Stage(s):	Project Planning, Problem Formulation, System Use and Monitoring

Confirmation Bias

Confirmation biases arise from tendencies to search for, gather, or use information that confirms pre-existing ideas and beliefs, and to dismiss or downplay the significance of information that disconfirms one's favoured hypothesis. This can be the result of motivated reasoning or sub-conscious attitudes, which in turn may lead to prejudicial judgements that are not based on reasoned evidence. For example, confirmation biases could surface in the judgment of the user of an AI decision-support application, who believes in following common sense intuitions acquired through professional experience rather than the outputs of an algorithmic model and, for this reason, dismisses its recommendations regardless of their rational persuasiveness or veracity.

Lifecycle Scope:	Whole of lifecycle
Significant Stage(s):	Problem Formulation, Data Analysis, System Use and Monitoring

Self-Assessment Bias

A tendency to evaluate one's abilities in more favourable terms than others, or to be more critical of others than oneself. In the context of a project team, this could include the overly-positive assessment the group's abilities (e.g. through reinforcing groupthink). For instance, during project planning, a project team may believe that their resources and capabilities are sufficient for the objective of the project, but in fact be forced to either cut corners or deliver a sub-par product.

Lifecycle Scope:	Whole of lifecycle
Significant Stage(s):	Project Planning

Availability Bias

The tendency to make judgements or decisions based on the information that is most readily available (e.g., more easily recalled). When this information is recalled on multiple occasions, the bias can be reinforced through repetition—known as a 'cascade'. This bias can cause issues for project teams throughout the project lifecycle where decisions are influenced by available or oft-repeated information (e.g. hypothesis testing during data analysis).

Lifecycle Scope:	Whole of lifecycle
Significant Stage(s):	Data Analysis, Model Training, Testing, and Validation

Naïve Realism

A disposition to perceive the world in objective terms that can inhibit recognition of socially constructed categories. For instance, treating 'employability' as something that is objectively measurable and, therefore, able to be predicted by a machine learning algorithm on the basis of objective factors (e.g., exam grades, educational attainment).

Lifecycle Scope:	Project Planning – Pre-processing and Feature Engineering
Significant Stage(s):	Problem Formulation

Representation Bias

When a population is either inappropriately represented (e.g., not allowing sufficient self-representation in demographic variables) or a sub-group is under-represented in the dataset, the model may subsequently fail to generalise and under-perform for a sub-group (or sub-groups). For example, representation biases could arise in a symptom checking application that has been trained on a data collected exclusively through smartphone use or online interaction as this dataset would likely underrepresent groups within the general population like elderly people who may lack access to smartphones or connectivity.

Lifecycle Scope:	Project Planning – Pre-processing and Feature Engineering
Significant Stage(s):	Problem Formulation, Pre-processing and Feature Engineering

Label Bias

A label (or feature) used within an algorithmic model may not mean the same thing for all data subjects. There may be a discrepancy between what sense the designers are seeking to capture in a label or feature, or what they are trying to measure in it, and the way that affected individuals understand its meaning. Where there is this kind of variation in meaning for different groups within a population, adverse consequences and discriminatory impact could follow. For example, designers of a predictive model in public health may choose "patient wellbeing" as their label, defining it in terms of disease prevalence and hospitalisation. However, subpopulations who suffer from health disparities and socioeconomic deprivation may understanding wellbeing more in terms of basic functionings, the food security needed for health promotion, and the absence of the social environmental stressors that contribute to the development of chronic medical conditions. Were this predictive model to be used to develop public health policy, members of this latter group could suffer from a further entrenchment of poor health outcomes.

Lifecycle Scope:	Project Planning – Pre-processing and Feature Engineering; System Use and Monitoring
Significant Stage(s):	Problem Formulation, Pre-processing and Feature Engineering

Missing Data Bias

Missing data can cause a wide variety of issues within an AI project, and these data may be missing for a variety of reasons related to broader social factors. Missingness can lead to inaccurate inferences and affect the validity of the model where it is the result of non-random but statistically informative events. For instance, missing data bias may arise in predictive risk models used in social care to detect potentially harmful behaviour in adolescents where interview responses and longitudinal data collected over extended periods of time are used as part of the dataset. This can be seen in cases where interview questions about socially stigmatised behaviours or traits like drug use or sexual orientation trigger fears of punishment, humiliation, or reproach and thus prompt non-responses, and in cases where data collection over time leads to the inconsistent involvement and drop-out of study participants.

Lifecycle Scope:	Whole of lifecycle
Significant Stage(s):	Data Analysis, Model Training, Testing, and Validation

Measurement Bias

This bias addresses the choice of how to measure the labels or features being used. It arises when the measurement scale being applied fails to capture data pertaining to the subjects in a fair and equitable manner. For example, a recidivism risk model that uses prior arrests or arrested of relatives as proxies to measure criminality may surface measurement bias insofar as patterns of arrest can reflect discriminatory tendencies to over-police certain protected social groups or biased assessments on the part of arresting officers.

Lifecycle Scope:	Project Planning – Pre-processing and Feature Engineering
Significant Stage(s):	Data Extraction or Procurement

Chronological Bias

Chronological bias arises when individuals in the dataset are added at different times, and where this chronological difference results in individuals being subjected to different methods or criteria of data extraction based on the time their data were recorded. For instance, if the dataset used to build a predictive risk model in childrens' social care spans over several years, large-scale care reforms, policy changes, adjustments in relevant statutes (such as changes to legal thresholds or definitions), and changes in data recording methods may create major inconsistencies in the data points extracted from person to person.

Lifecycle Scope:	Project Planning – Data Analysis
Significant Stage(s):	Project Planning, Data Extraction or Procurement

Selection Bias

Selection bias is a term used for a range of biases that affect the selection or inclusion of data points within a dataset. In general, this bias arises when an association is present between the variables being studied and additional factors that make it more likely that some data will be present in a dataset when compared to other possible data points in the space. For instance, where individuals differ in their geographic or socioeconomic access to an activity or service that is the site of data collection, this variation may result in their exclusion from the corresponding dataset. Likewise, where certain socioeconomically deprived or marginalised social groups are disproportionately dependent on a social service to fulfil basic needs, it may be oversampled if data is collected from the provision of that service.

Lifecycle Scope:	Project Planning – Data Analysis
Significant Stage(s):	Project Planning, Data Extraction or Procurement

Wrong Sample Size Bias

Using the wrong sample size for the study can lead to chance findings that fail to adequately represent the variability of the underlying data distribution, in the case of small samples, or findings that are statistically significant but not relevant or actionable, in the case of larger samples. Wrong sample size bias may occur in cases where model designers have included too many features in a machine learning algorithm. This is often referred to as the "curse of dimensionality", a mathematical phenomenon wherein increases in the number of features or "data dimensions" included in an algorithm means that exponentially more data points need to be sampled to enable good predictive or classificatory performance.

Lifecycle Scope:	Project Planning – Model Training, Testing, and Validation
Significant Stage(s):	Data Extraction or Procurement, Pre-processing or Feature Engineering

Aggregation Bias

Aggregation bias arises when a "one-size-fits-all" approach is taken to the outputs of a trained algorithmic model (i.e., that model results apply evenly to all members of the impacted population) even where variations in subgroup characteristics mean that mapping functions from inputs to outputs are not consistent across these subgroups. In other words, in a model where aggregation bias is present, even when combinations of features affect members of different subgroups differently, the output of the system disregards the relevant variations in condition distributions for the subgroups. This results in the loss of information, lowered performance, and, in cases where data from one subgroup is more prevalent than those of others, the development of a model that is more effective for that sub-group. Good examples of aggregation bias come up in clinical decision-support systems in medicine, where clinically significant variations between sexes and ethnicities—in terms of disease aetiology, expression, complications, and treatment—mean that systems which aggregate results by treating all data points similarly will not perform optimally for any subgroup.

Lifecycle Scope:	Pre-processing or Feature Engineering – System Use and Monitoring
Significant Stage(s):	Pre-processing or Feature Engineering

Law of the Instrument (Maslow's Hammer)

This bias is best captured by the popular phrase 'If all you have is a hammer, everything looks like a nail'. The phrase cautions against the cognitive bias of over-reliance on a particular tool or method, perhaps one that is familiar to members of the project team. For example, a project team that are experts in a specific ML technique, may over-use the technique and mis-apply it in a context where a different technique would be better suited. Or, in some cases, where it would be better not to use ML/AI technology at all.

Lifecycle Scope:	Whole of lifecycle
Significant Stage(s):	Project Planning, Model Selection, Model Training, Testing, and Validation

Evaluation Bias

Evaluation Bias occurs during model iteration and evaluation from the application of performance metrics that are insufficient given the intended use of the model and the composition of the dataset on which it is trained. For example, an evaluation bias may occur where performance metrics that measure only overall accuracy are applied to a trained computer vision system that performs differentially for subgroups that have different skin tones. Likewise, evaluation biases arise where the external benchmark datasets that are used to evaluate the performance of trained models are insufficiently representative of the populations to which they will be applied. In the case of computer vision, this may occur where established benchmarks overly represent a segment of the populations (such as adult light-skinned males) and thus reinforce the biased criteria for optimal performance.

Lifecycle Scope:	Data Analysis – Model Updating and Deprovisioning
Project Lifecycle Stage(s):	Data Analysis, Model Training, Testing, and Validation

Confounding

Confounding is a well-known causal concept in statistics, and commonly arises in observational studies. It refers to a distortion that arises when a (confounding) variable independently influences both the dependant and independent variables (e.g., exposure and outcome), leading to a spurious association and a skewed output. Clear examples of confounding can be found in the use of electronic health records (EHRs) that arise in clinical environments and healthcare processes. EHRs are observational data and often reflect not only the health status of patients, but also patients' interactions with the healthcare system. This can introduce confounders such as the frequency of inpatient medical testing reflecting the busyness or labour shortages of medical staff rather than the progression of a disease during hospitalisation, differences between onset of a disease and the date of diagnosis, and health conditions that are missing from the EHRs of a patient due to a non-random lack of testing. Contextual awareness and domain knowledge are crucial elements for identifying and redressing confounders.

Lifecycle Scope:	Data Analysis – Model Reporting
Significant Stage(s):	Data Analysis

Training-Serving Skew

Occurs when the model is deployed on individuals whose data are not similar to or representative of the individuals whose data were used to train, test, and validate the model. This can occur, for instance, where a trained model is applied to a population in a different geographical area from that where the original data were collected or to the same population but at a time much later than that at which the training data were first collected. In both cases, the trained model may fail to generalise because the new, out-of-sample inputs are being drawn from populations with different underlying distributions.

Lifecycle Scope:	Data Extraction or Procurement – System Use and Monitoring
Significant Stage(s):	Model Training, Testing, and Validation, System Use and Monitoring

Optimism Bias

Also known as the planning fallacy, optimism bias can lead project teams to under-estimate the amount of time required to adequately implement a new system or plan. In the context of the project lifecycle, this bias may arise during project planning, but can create downstream issues when implementing a model during the 'model productionalisation' stage, due to a failure to recognise possible system engineering barriers.

Lifecycle Scope:	Whole of lifecycle
Significant Stage(s):	Model Productionalisation, System Use and Monitoring

Implementation Bias

Implementation bias refers, generally, to any bias that arises when a system is implemented or used in ways that were not intended by the designers or developers but, nevertheless, made more likely due to affordances of the system or its deployment. For example, a biometric identification system that is used by a public authority to assist in the detection of potential terrorist activity could be repurposed to target and monitor activists or political opponents.

Lifecycle Scope:	Model Productionalisation – System Use or Monitoring
Significant Stage(s):	User Training, System Use and Monitoring

Decision-Automation Bias

Decision-automation bias arises when users of automated decision-support systems become hampered in their critical judgment, rational agency, and situational awareness as a result of their faith in the efficacy of the system. This may lead to over-reliance or errors of omission, where implementers lose the capacity to identify and respond to the faults, errors, or deficiencies, which might arise over the course of the use of an automated system, because they become complacent and overly deferent to its directions and cues. Decision-automation bias may also lead to over-compliance or errors of commission where implementers defer to the perceived infallibility of the system and thereby become unable to detect problems emerging from its use for reason of a failure to hold the results against available information.

Lifecycle Scope:	User Training – System Use and Monitoring
Significant Stage(s):	User Training

Automation-Distrust Bias

Automation-distrust bias arises when users of an automated decision-support system disregard its salient contributions to evidence-based reasoning either as a result of their distrust or scepticism about AI technologies in general or as a result of their over-prioritisation of the importance of prudence, common sense, and human expertise. An aversion to the non-human and amoral character of automated systems may also influence decision subjects' hesitation to consult these technologies in high impact contexts such as healthcare, transportation, and law.

Lifecycle Scope:	User Training – System Use and Monitoring
Significant Stage(s):	System Use and Monitoring

Status Quo Bias

An affectively motivated preference for "the way things are currently", which can prevent more innovative or effective processes or services being implemented. This bias is most acutely felt during the transition between projects, such as the choice to deprovision a system and begin a new project, in spite of deteriorating performance from the existing solution. Although this bias is often treated as a cognitive bias, we highlight it here as a social bias to draw attention to the broader social or institutional factors that in part determine the status quo.

Lifecycle Scope:	Model Updating or Deprovision – Project Planning
Significant Stage(s):	Model Updating or Deprovision, Project Planning