

01

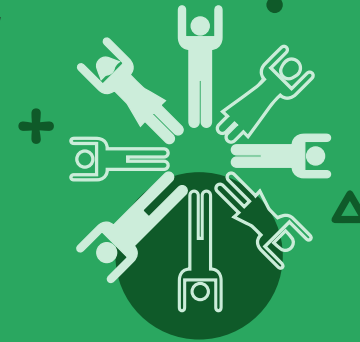


Social Biases

Historical Bias

Historical biases affect many areas of society and exist prior to the start of a project. If not carefully considered, a project can exacerbate current socioeconomic inequalities.

02



Social Biases

Representation Bias

Representation bias can arise when a population is not appropriately represented within a dataset, leading to the model underperforming for the respective sub-group.

Description

This bias can arise 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. In these cases, the AI model may subsequently fail to generalise, and under-perform for a sub-group (or sub-groups).

Deliberative Prompts

- › How have you measured and evaluated the representativeness of the dataset to ensure that the sample is adequate?
- › Have you consulted stakeholder groups to verify that your dataset is representative?

Description

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 objectives of the 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.

Deliberative Prompts

- › Which groups and communities will be affected by the use of your model or system?
- › Are there groups or communities that will be excluded from your model or experience barriers to using your system? If so, why?
- › Is there a risk of worsening or perpetuating socioeconomic inequalities in the development and deployment of your model?

03



Social Biases

Label Bias

Labels or features used by algorithmic systems may have different meanings for different groups, leading to adverse consequences and discrimination.

04



Social Biases

Annotation Bias

Annotation bias arises when annotators introduce subjective perceptions, error, or systemic sociocultural biases into the data annotation process, often due to fatigue or a lack of focus.

Description

Annotation bias occurs when annotators incorporate subjective perceptions or error into the work of annotating data. Data annotation often occurs under less-than-ideal scenarios, including contexts in which human error may occur due to fatigue or lack of focus, or from annotators not receiving sufficient training.

Annotation bias can also result from positionality limitations that derive from demographic features, such as age, education, or first language, as well as other systemic cultural or societal biases that influence annotators.

Deliberative Prompts

- › Who carried out the annotation of your dataset? What methods did they follow?
- › Were there processes in place to ensure that multiple annotators followed the same standards (e.g. inter-rater reliability)?

Description

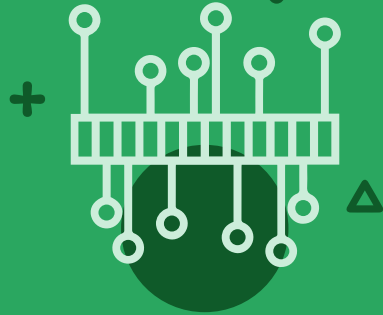
A label (or feature) used within an algorithmic model may not mean the same thing for all data subjects. There may be a discrepancy between what sense the designers are seeking to capture or what they are trying to measure in a label or feature, and the way that affected individuals understand its meaning.

Where there is this kind of variation in meaning for different groups within a population, adverse consequences and discriminatory impact could follow.

Deliberative Prompts

- › How have you identified problematic labels (or features), which may be imperfect proxies, within your dataset?
- › Does your target variable have multiple meanings or interpretations?
- › Are labels used across the project lifecycle and have they been clearly defined?

05



Social Biases

Chronological Bias

Chronological bias occurs when data are recorded at different times, such that different methods or criteria are used to determine their values.

06



Social Biases

Selection Bias

Selection bias occurs when systematic barriers affect the rate of inclusion for certain sub-groups of data points or subjects within a dataset.

Description

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. If for instance 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 exclusions from the corresponding dataset based on those differences.

Deliberative Prompts

- › Have you examined the different stakeholders that are included or not included within the data and datasets being considered?
- › Are there stakeholder groups you can consult with to help minimise the likelihood of you and your team missing key stakeholder considerations?

Description

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.

Deliberative Prompts

- › Have you worked with domain experts to map the data journey and identify systematic variations between groups of data subjects or objects?
- › Is there a wide variation in terms of when your data were recorded?

07



Social Biases

Implementation Bias

Implementation bias can occur when a system is used in ways that were not originally intended by the designers or developers of the system.

08



Social Biases

Status Quo Bias

Status quo bias arises from an affective attachment to the current state of things, even when it prevents more effective processes or services being implemented.

Description

An affectively motivated preference for “the way things currently are”, which can prevent more effective processes or services being implemented. This bias is most acutely felt during the transition between projects. For example, it may be difficult for a team to decide to deprovision a system and instead begin a new project, even 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.

Deliberative Prompts

- › Have you assessed how your team members feel about the use or lack of use of technology in your project? Is this different to how things have usually been done within your team?
- › Are you able to consult with someone outside of your team to see if your project as well as the proposed problem and solution are appropriate?

Description

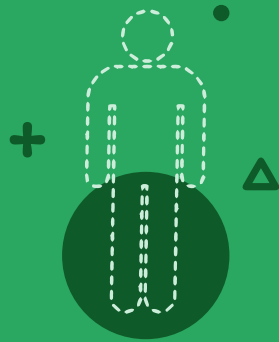
Implementation bias refers, generally, to any bias that arises when a system is implemented or used in ways that were not intended by the designers or developers but, nevertheless, made more likely due to affordances of the system or its deployment.

Design choices made during the implementation of a system can create so-called, ‘choice architectures’ that make specific actions or decisions more or less probable, whether intentionally or not.

Deliberative Prompts

- › Has your system been repurposed from another project or team? If so, is the system fit-for-purpose?
- › Does the use of the system now differ from how it was previously used?

09

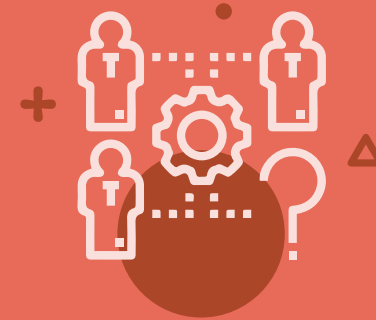


Social Biases

De-Agentification Bias

De-agentification bias arises when minoritised or marginalised groups are excluded from participating meaningfully in the development of algorithmic systems, or otherwise enjoying their benefits.

01



Statistical Biases

Missing Data Bias

Missing data bias arises when relevant data is missing from a dataset, causing inaccurate inferences which affect model validity, particularly when the missingness is non-random.

Description

Relevant data may be missing in a project for a variety of reasons related to social factors and can cause a wide variety of issues within an AI project.

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.

That is, when data is missing in a non-random manner, it is likely that the data is missing for reasons which are relevant to the model's performance.

Deliberative Prompts

- › How have you dealt with and recorded your handling of missing data (e.g. choice of imputation or augmentation method)?
- › Have you consulted with domain experts to help you identify possible explanations for the missing data and whether they may be informative?

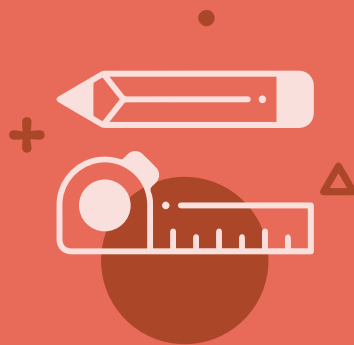
Description

De-agentification bias occurs when social structures and innovation practices systemically exclude minoritised, marginalised, vulnerable, historically discriminated against, or disadvantaged social groups from participating or providing input in AI innovation ecosystems. Protected groups may be prevented from having input into the development, use, and evaluation of models. They may lack the resources, education, or political influence to detect biases, protest, and force correction.

Deliberative Prompts

- › Have you considered consulting, engaging, and working with protected and marginalised groups as part of your project? How have their perspectives and experiences been considered?

02

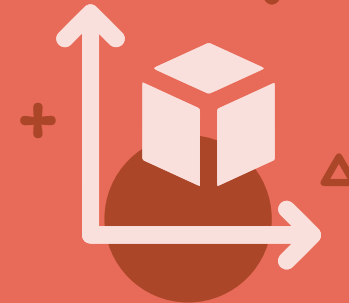


Statistical Biases

Measurement Bias

Measurement bias refers to the unfair or inequitable consequences of using an inappropriate or limited scale for measuring the labels or features used in a model.

03



Statistical Biases

Wrong Sample Size Bias

Improper sample size can lead to chance findings or statistically significant but irrelevant outcomes, particularly when too few or too many features are included in a machine learning algorithm.

Description

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.

It may also 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.

Deliberative Prompts

- › Which methods or statistical indicators (e.g. p-values, confidence intervals) have been used and reported to help ensure that the findings did not arise by chance?
- › Have you considered the likely use case for the results? How will this be reported (e.g. in 'limitations' section) to help readers assess the relevance of the results?

Description

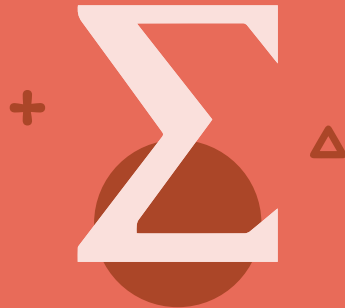
Measurement bias occurs when the measurement method used to collect data and define the features or labels used by a model is flawed or fails to capture relevant information about the objects or subjects being studied.

Measurement bias can arise from a variety of factors, such as biased sampling techniques or flawed data collection processes. However, a common source of the bias is a limitation with the measurement scale being used, which may fail to capture some key characteristic of the object or subject being represented.

Deliberative Prompts

- › Are there multiple scales that could be used to measure your features? Is there reasonable disagreement about which of these scales is preferred? If so, how has this disagreement been addressed?

04



Statistical Biases

Aggregation Bias

Aggregation bias occurs when a uniform approach is applied to a trained algorithmic model's outputs, ignoring the variations in subgroup characteristics.

05



Statistical Biases

Evaluation Bias

Evaluation bias arises when the performance metrics used to evaluate a model are inadequate for the model's intended use or the dataset on which it is trained.

Description

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.

This bias can arise when 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.

Deliberative Prompts

- › How will you divide your dataset into separate training and testing datasets?
- › Will you validate the model against an external benchmark population? If not, have you taken steps to report these limitations?

Description

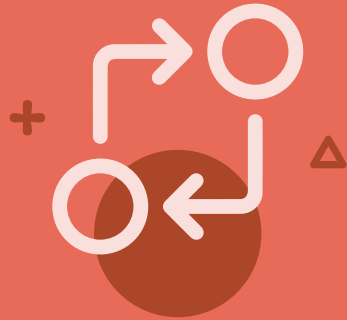
Aggregation bias arises when a “one-size-fits-all” approach is taken to the outputs of a trained algorithmic model even where variations in subgroup characteristics mean that mapping functions from inputs to outputs are not consistent across subgroups.

In other words, if 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 conditional distributions for the subgroups. This may result in the loss of relevant information, lowered performance, and the development of a model that is more reliable some sub-groups.

Deliberative Prompts

- › Which evaluation methods (e.g. model comparison) have you employed to help you identify aggregation bias and its impact on the various subgroups in your dataset?

06

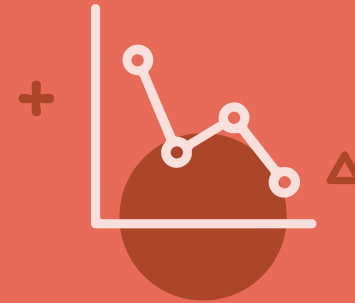


Statistical Biases

Confounding

Confounding occurs when a (confounding) variable affects both the dependent and independent variables, leading to skewed or distorted output and spurious associations.

07



Statistical Biases

Training-Serving Skew

Training-serving skew refers to the deployment of a model in a context or environment that differs substantially from the environment represented by its training data.

Description

This bias 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. It may arise if, for instance, 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 much later time than that when the training data were collected. The trained model may then fail to generalise because the new, out-of-sample inputs are being drawn from populations with different underlying distributions.

Deliberative Prompts

- › What steps have you taken to measure and evaluate the performance of your model within the intended domain (e.g. use of synthetic data, external validation on similar datasets)?
- › Have you engaged domain experts to ensure these steps are adequate (e.g. sufficiently representative of the impacted users)?

Description

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.

Deliberative Prompts

- › Are there methods you can use (e.g. propensity score matching, causal diagrams) that could help reduce bias that results from confounding (e.g. in the estimation of the average treatment effect)?
- › Is the sample size sufficient (i.e. large enough) to minimise the impact of confounders?

01



Cognitive Biases

Confirmation Bias

Confirmation bias is the tendency to favour information that confirms one's pre-existing beliefs and ignore or downplay evidence conflicting one's beliefs.

02



Cognitive Biases

Self-Assessment Bias

The overestimation of one's abilities and underestimation of others, leading to an overly positive assessment of one's own capacities or those of one's group.

02

Description

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 of the group's abilities (e.g., through reinforcing groupthink).

Deliberative Prompts

- › As part of the planning for your project, have you considered things that may go wrong or have a negative impact?
- › Are you able to be more flexible with your timeline to accommodate for identifying and addressing gaps of knowledge and skills within your team?
- › Have you and your project team considered obtaining constructive criticism and suggestions from others?

01

Description

Confirmation biases arise from a typical human tendency 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.

Deliberative Prompts

- › What mechanisms do you have in place within your team that can help ensure a diversity of viewpoints that may mitigate the effects of confirmation bias?

03



Cognitive Biases

Availability Bias

The tendency to make decisions based on easily available or recalled information, which can lead to biased judgments or decisions.

04



Cognitive Biases

Naïve Realism

A disposition to perceive the world in unrealistic, objective terms that can inhibit the recognition of socially constructed categories.

Description

Naive realism is a disposition to perceive the world in objective terms, which can inhibit the recognition of socially constructed categories.

As a result of this disposition, people are less inclined to identify how their own personal experiences contribute to their understanding or interpretation of a phenomenon or object being studied, or to reject alternative perspectives as mistaken or irrational.

For instance, individuals may fail to identify how their cultural or political beliefs influence how they perceive categories such as emotions or social behaviours, and falsely describe these phenomena in objective terms rather than recognising their subjective or intersubjective elements.

Deliberative Prompts

- › Have you identified non-quantifiable or difficult-to-measure qualitative factors that may contribute to and affect your model or decision-making process? How are these documented and accounted for?

Description

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).

Deliberative Prompts

- › Have you considered alternative sources, references, datasets, and methods that can help minimise gravitating towards readily available or memorable information?

05



Cognitive Biases

Law of the Instrument (Maslow's Hammer)

The over-reliance on a particular tool or method without consideration of whether it is the right tool for the job, leading to "fitting the problem" to the capabilities of the tool.

06



Cognitive Biases

Optimism Bias

Optimism bias occurs when a team underestimates the amount of time required to complete a project or plan.

Description

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 'system implementation' stage, due to a failure to recognise possible system engineering barriers.

Deliberative Prompts

- › Have you and your team been realistic with what can be achieved within the time allocated to the project?
- › Are you able to be more flexible with your time and resources, particularly where stakeholder engagement is involved?

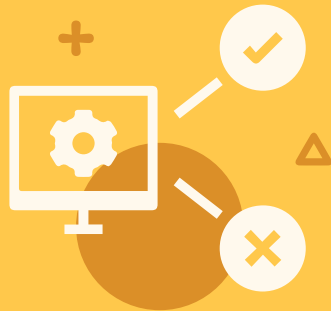
Description

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 overuse the technique and misapply 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.

Deliberative Prompts

- › Is the technology you're developing the best way forward for your project? Who has determined this?
- › If you're repurposing an existing technology, is it fit-for-purpose for the task and project at hand?
- › Does your team have the appropriate knowledge and skillset to adopt the current system, model or tool?

07



Cognitive Biases

Decision-Automation Bias

The tendency to rely too heavily on automated decision-support systems, leading to errors of omission or commission.

08



Cognitive Biases

Automation-Distrust Bias

Automation-distrust bias arises when users of an automated decision-support system disregard its contributions due to illegitimate distrust or scepticism of the system.

Description

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.

Deliberative Prompts

- › Have you engaged the intended users of your system early on in project planning to identify barriers and co-design solutions that would increase the level of trust they have in your system?
- › Is there information you could provide to help reduce any concerns users would have about how your model or system operates?

Description

This bias arises when users of automated decision-support systems become hampered in their critical judgement as a result of their faith in the efficacy of the system. This may lead to over-reliance (errors of omission), where implementers lose the capacity to identify and respond to the faults which might arise when using an automated system because they become complacent and overly deferent to its directions and cues. Decision-automation bias may also lead to over-compliance (errors of commission) where implementers defer to the perceived infallibility of the system and thereby become unable to detect problems emerging from its use.

Deliberative Prompts

- › Have you considered user requirements such as transparency or interpretability when designing your model?
- › Does the intended context of use demand a greater need for interpretability, and how may this affect the model's accuracy (e.g. reducing model complexity)?
- › Could long-term use of your model or system have a detrimental effect on the professional judgement of users (e.g. leading to deskilling)?

01



Bias Mitigation Techniques

Peer Review

Targeted review of work by a committee, red team, or other group to identify and evaluate any gaps or issues.

Can be internal or external (e.g. independent auditor).

02



Bias Mitigation Techniques

Additional Data Collection

Return to the data extraction (or procurement) stage to carry out additional data collection or reconsider methods of data extraction (e.g. revised experimental methods, more inclusive and accessible forms of engagement).

03



Bias Mitigation Techniques

Participatory Design Workshops

A form of stakeholder engagement that seeks to involve stakeholders within the design process to identify needs and preferences, co-create solutions, and ensure usability and acceptance.

04



Bias Mitigation Techniques

Stakeholder Engagement

Carry out meaningful forms of engagement to consult or partner with wider stakeholders. This could include hosting community fora, conducting online surveys or interviews, or even running a citizen jury or assembly.

05



Bias Mitigation Techniques

Human-in-the-loop

Agree on guidelines to ensure the use of data-driven technologies support human decision making by providing recommendations or automating routine tasks, while still allowing humans to make final decisions and have clear oversight.

06



Bias Mitigation Techniques

Identify Under-represented Groups

Analyse gaps in demographic data in consultation with community groups and domain experts. Develop appropriate methods to address gaps and limitations based on context-aware reflection.

07



Bias Mitigation Techniques

Skills and Training

Organise and facilitate skills and training events, such as webinars, workshops, self-directed learning, to upskill project team members or users (e.g. understanding and communicating uncertainty of predictive models, interactions with system interface).

08



Bias Mitigation Techniques

Data Augmentation

Augment your dataset using techniques appropriate to the objective (e.g. addressing sparsity), such as data linkage or mixing, synthetic data generation, imputation, adding noise, transformation.

09



Bias Mitigation Techniques

Diversify Evaluation Metrics

Use additional evaluation metrics for your model to determine whether its performance applies equally for all individuals or sub-groups. Where relevant carry out intersectional analysis of multiple demographic or identity characteristics to identify biases that may not be apparent when considering a single characteristic.

10



Bias Mitigation Techniques

Multiple Model Comparison

Train and test multiple models, both within the same class of models and also across classes to assess a broader range of possible performance values.

11



Bias Mitigation Techniques

External Validation

Go beyond the *internal* validation of your model (i.e. training-testing split of data) and perform *external* validation with an entirely new dataset. You could engage with another team or organisation to help validate your study or model development in a new environment (e.g. different population of data subjects, novel geographical environment).

13



Bias Mitigation Techniques

Open Documentation

Where possible, document the actions and decisions made throughout your project to support reproducibility and replicability efforts, assist users of your system, and promote best practices of transparency.

14



Bias Mitigation Techniques

Regular Auditing

Work with another team, committee, or organisation to perform regular audits of your project, focusing on key areas such as transparency and explainability, data quality, model performance, user satisfaction, and equitable impact.

15



Bias Mitigation Techniques

Employ Model Interpretability Methods

During data analysis, model testing and validation, and system use and monitoring, use appropriate model interpretability methods (e.g. local, model-agnostic, data visualisation) to ensure that your model is meeting the original objectives for your project.

16



Bias Mitigation Techniques

Quality Control Procedures

Conduct regular assessments of your model or system against established quality control procedures (e.g. analytical quality assurance) to ensure that issues are identified early on (e.g. clerical errors in data input that may arise from time-pressured human inputters or annotators).

12



Bias Mitigation Techniques

Double Diamond Methodology

The Double Diamond methodology is a process for design that is well-suited to creative approaches to problem-solving and exploring multiple perspectives and possibilities. The method consists of four phases:

Phases

- 1 **Discover:** gain insight and identify the problem, understanding needs and challenges, and gather information in a highly exploratory manner.
- 2 **Define:** clarify the information from the previous stage to gain a narrower, well-defined area to focus on.
- 3 **Develop:** generate and test possible solutions, exploring the feasibility and desirability of the solutions, while also identifying areas that need additional work.
- 4 **Deliver:** deliver a final product or service that meets the original specification (e.g. minimum viable product), and which can be used to gather additional feedback.