Bias Reflectlist Cheat Sheet

About this Document

This document is a cheat sheet for the bias ~~checklist~~ reflectlist activity.

It includes the following:

1 The Bias Taxonomy: a table of the biases, with descriptions, examples, and deliberative prompts.

2 Bias Mitigation Techniques: a list of bias mitigation techniques that can be employed to help address the above biases.

3 Project Lifecycle Heatmap: a visualisation of the biases across the project lifecycle, showing where they are most significant and the scope of their impact.

For further details, please see the full activity handout on the [Turing Commons website.]

Bias Taxonomy

SOCIAL BIASES	
Name	**Historical Bias**
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.
Example	Examples of historical bias include social dynamics that contribute to prejudicial arrest rates in policing, or social determinants of criminal behaviour and outcomes, such as poverty that can create higher risks of recidivism.
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?

Name	**Representation Bias**
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).
Example	Representation bias could arise in a symptom checking application that has been trained on data collected exclusively through smartphone use or other online interaction. This dataset would likely underrepresent groups, such as elderly people who may lack access to a smartphone or connectivity.
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?

Name	**Label Bias**
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.
Example	For example, designers of a predictive model in public health may choose "patient wellbeing" as their label, and then define it in terms of disease prevalence and hospitalisation. However, subpopulations who suffer from health disparities and socioeconomic deprivation may understand wellbeing more in terms of basic functioning, 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.
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?

Name	**Annotation Bias**
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.
Example	An example of annotation bias is when police officers misidentify the race or ethnicity of a criminal suspect in an arrest report due to uncertainty or personal bias. Data sets produced in this context may misrepresent the prevalence of arrests amongst demographic subgroups, leading to erroneous conclusions about crime trends.
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)?

Name	**Chronological Bias**
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.
Example	An example of chronological bias could be where a dataset used to build a predictive risk model in children's social care has data that spans over several years, in which large-scale care reforms, policy changes, adjustments in relevant statutes (such as changes to legal thresholds or definitions) have occurred. As such, there may also have been changes in data recording methods that could create major inconsistencies in the data points extracted from person to person.
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?

Name	**Selection Bias**
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.
Example	An example of selection bias is where pregnant women are routinely not selected for drug trials, due to increased risks. However, while safeguarding them during pregnancy, their lack of inclusion also leads to lower efficacy for their cohort (e.g. real-world lack of efficacy for certain pain killers).
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?

Name	**Implementation Bias**
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.
Example	Consider a biometric identification system that was initially designed by a public authority to assist in the detection of potential terrorist activity but is now repurposed to target and monitor non-violent activists or political opponents.
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?

Name	**Status Quo Bias**
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.
Example	This bias can occur in cases where people are more critical of technological systems, even though they may outperform biased human decision-making. Suppose a school is thinking about implementing a decision support system used to help school staff identify under-performing children. This system has some biases, but these are less significant than the existing human biases that allow some students to fall "beneath the radar". If the school does not change to the less biased decision support system, it may be due to status quo bias.
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?

Name	**De-Agentification Bias**
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.
Example	An example is the choice to design, develop, or deploy a system for monitoring historically marginalised communities, such as refugees and religious minorities. Such communities are often not represented in key decisions concerning the adoption and use of such systems though they may be significantly affected by them.
Deliberative Prompts	› Have you considered consulting, engaging, and working with protected and marginalized groups as part of your project? How have their perspectives and experiences been considered?

STATISTICAL BIASES

Name	**Missing Data Bia**
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.
Example	For instance, missing data bias may arise in predictive risk models used in social care 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.
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?

Name	**Measurement Bias**
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.
Example	A recidivism risk model that uses prior arrests or criminal records of relatives as proxies to predict future 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.
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?

Name	**Wrong Sample Size Bias**
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.
Example	An example comes from not using a big enough sample size to understand or explain a phenomenon, and then extrapolating the findings to the overall or general population. If a machine learning model is trained with insufficient data, its classifications or predictions are unlikely to be very good, as they are unlikely to adequately represent the overall data distribution of interest.
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?

Name	**Aggregation Bias**
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.
Example	Examples of aggregation bias include clinical decision-support systems in medicine, where clinically significant variations between patient cohorts (e.g. different 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.
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?

Name	**Evaluation Bias**
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.
Example	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. If benchmarks overly represent a segment of the populations (such as adult light-skinned males), they may not be good enough at capturing performance for other skin tones and thus reinforce the biased criteria for optimal performance.
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?

Name	**Confounding**
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.
Example	Clear examples can be found in the use and analysis of electronic health records (EHRs). EHRs 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 which may be reflecting the labour shortages of medical staff rather than the progression of a disease during hospitalisation. Contextual awareness and domain knowledge are therefore crucial elements for identifying and redressing confounders.
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?

Name	**Training-Serving Skew**
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.
Example	Imagine a recommendation system for an e-commerce website which was trained using all available user data at the time, but when it is finally deployed to the website, its recommendations do not perform as well as during training. Upon investigation, the team discover that the model is biased towards recommending products which were popular at the time of training, but which no longer are. This is an example of training-serving skew as the performance of the model is biased due to changes in the underlying distribution of the data which occurred between the training and the deployment of the system.
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)?

COGNITIVE BIASES

Name	**Confirmation Bias**
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.
Example	Consider a policymaker or minister who has strong attitudes on the economic impacts of immigration. If their pre-existing stance leads to them ignoring or downplaying models and empirical evidence that serve as evidence against their views, and only considering evidence that support their existing attitudes, they are suffering from confirmation bias.
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?

Name	**Self-Assessment Bias**
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).
Example	Consider a project team that is carrying out an assessment about whether they have sufficient skills and resources to develop fair and explainable ML algorithms. Self-assessment bias could create a situation where the team are unlikely to acknowledge or notice gaps in their skills, which may significantly affect their ability to deliver a responsible product.
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?

Name	**Availability Bias**
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).
Example	If a team uses a dataset that they already have access to, even though it is not actually the best data for their problem, this is a form of availability bias. However, this is different from the form of availability bias that may affect people when recalling certain facts throughout a project's lifecycle. Here, availability refers to the individual's ability to recall information, rather than to an ability to access data.
Deliberative Prompts	› Have you considered alternative sources, references, datasets, and methods that can help minimise gravitating towards readily available or memorable information?

Name	**Naïve Realism**
Description	Naïve 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.
Example	An example of naïve realism would be treating 'employability' as something that is objectively measurable and, therefore, able to be predicted by a machine learning algorithm based on objective factors (e.g., exam grades, educational attainment), without recognising the social construction inherent in these factors.
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?

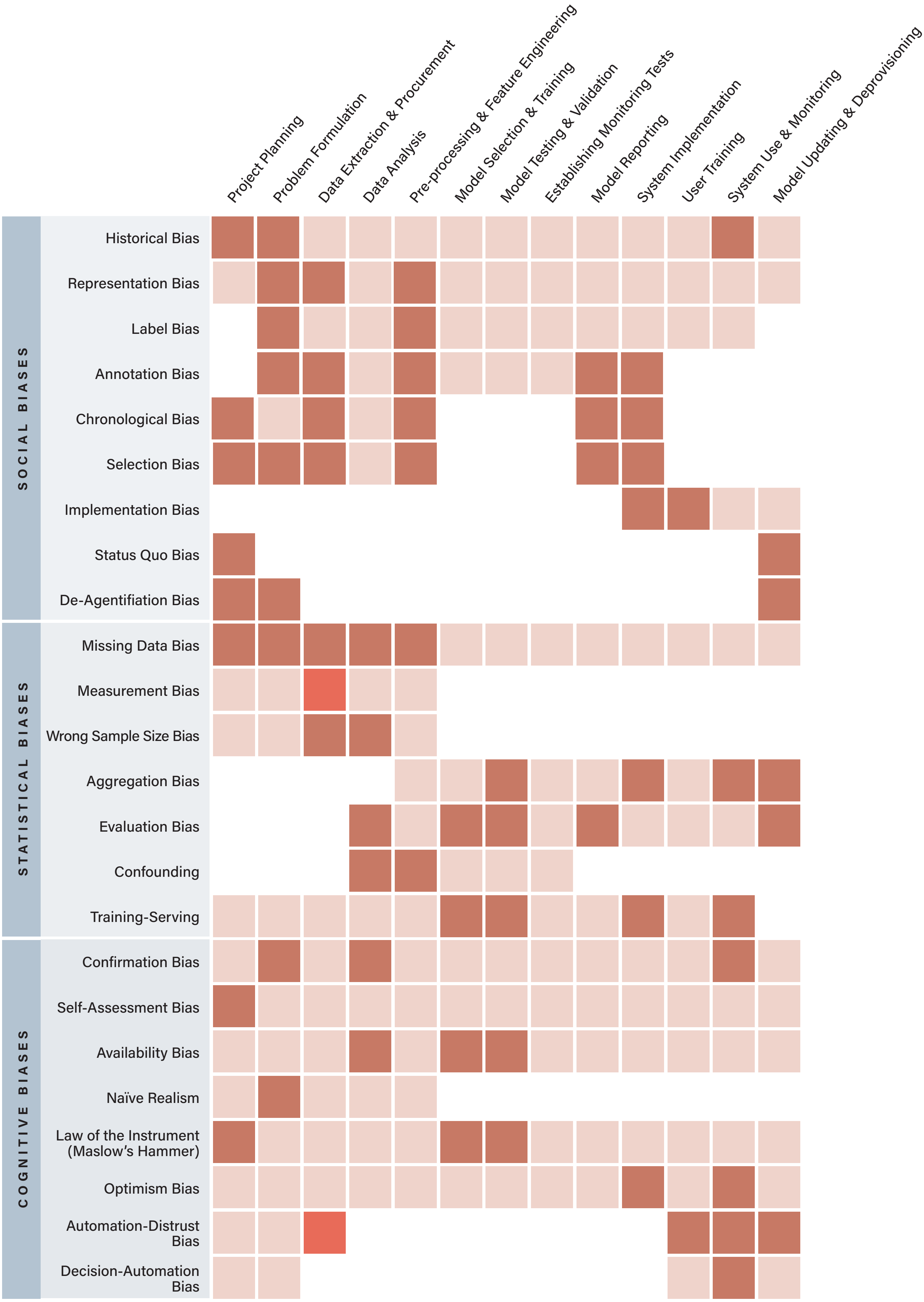
Name	**Law of the Instrument (Maslow's Hammer)**
Description	This bias is best captured by the popular phrase 'If all you have is a hammer, everything looks like a nail.' The phrase cautions against the cognitive bias of over-reliance on a particular tool or method, perhaps one that is familiar to members of the project team. For example, a project team that are experts in a specific ML technique, may over-use 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.
Example	If an organisation develops a system to parse natural language, and successfully deploys it for one task, but then uses it in a new project without considering whether it is the right tool, they are falling prey to this bias.
Deliberative Prompts	› Is 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?

Name	**Optimism Bias**
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.
Example	During project scoping, a project management team incorrectly assume that it will only take three months to design, develop, and deploy a new algorithmic system, because a previous (and similar) project took this long. However, despite the success of the previous project, their assessment this time turns out to be an underestimate because they did not consult with their developers to fully understand important differences between the two projects.
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?

Name	**Decision-Automation Bias**
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.
Example	An immigration officer is using facial recognition software, which purportedly claims to detect instances of lying during asylum claim interviews. Over time, the officer stops relying on their own faculties, and leans too heavily on the predictions of this system, despite visual cues that contradict the facial recognition system's predictions.
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)?

Name	**Automation-Distrust Bias**
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.
Example	Members of a profession (e.g. judges, doctors) who rule out decision support systems based on (potentially unfounded) fears of these technologies, may be influence by this bias. However, in such instances, their aversion to using technology in a constructive way, may prevent them from identifying and mitigating some of their own cognitive biases, or improving evidence-based decisions in their respective fields.
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?

Heatmap of Biases



Bias Mitigation Techniques

Name	**Peer Review**
Description	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).

Name	**Additional Data Collection**
Description	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).

Name	**Participatory Design Workshops**
Description	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.

Name	**Stakeholder Engagement**
Description	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.

Name	**Human-in-the-loop**
Description	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.

Name	**Identify Under-represented Groups**
Description	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.