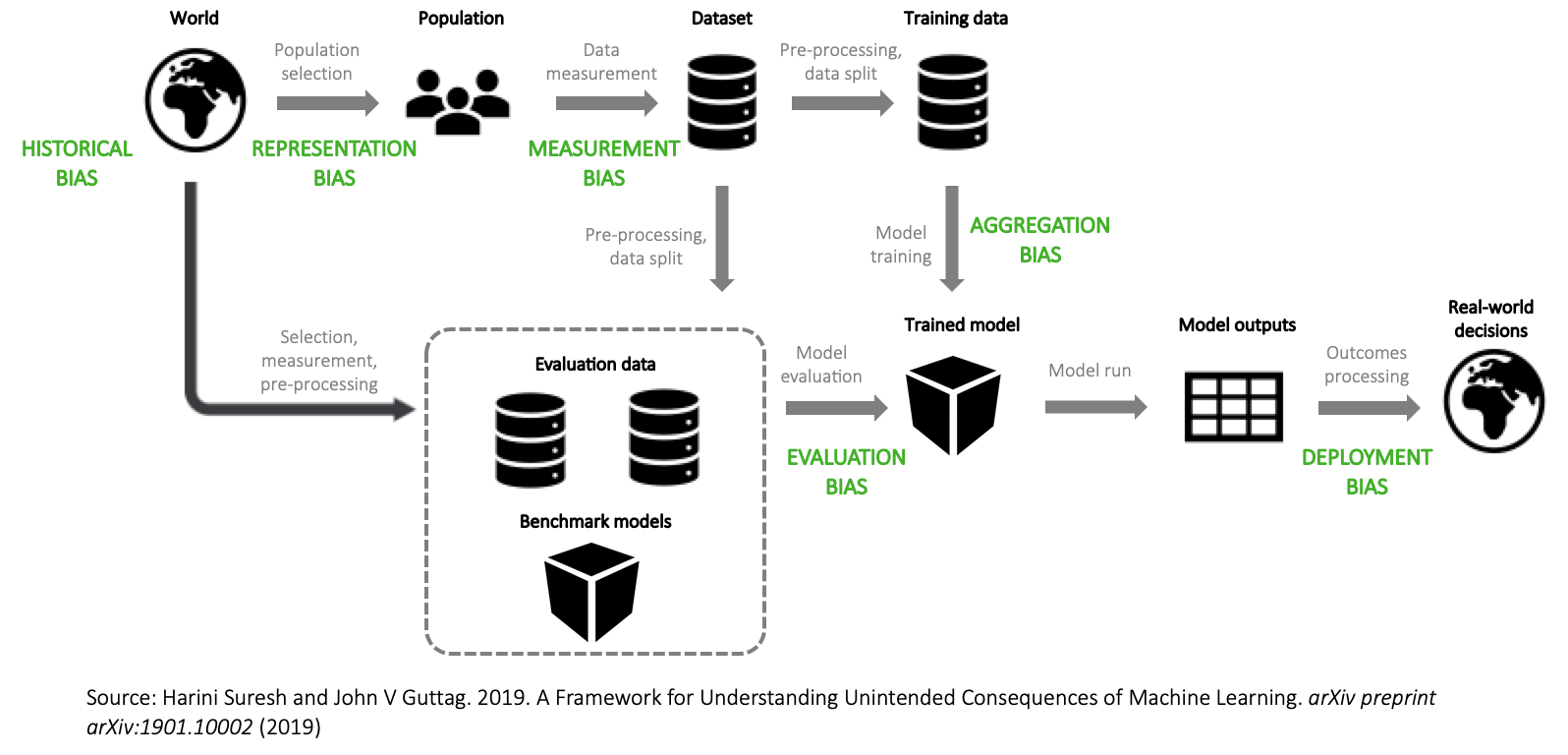
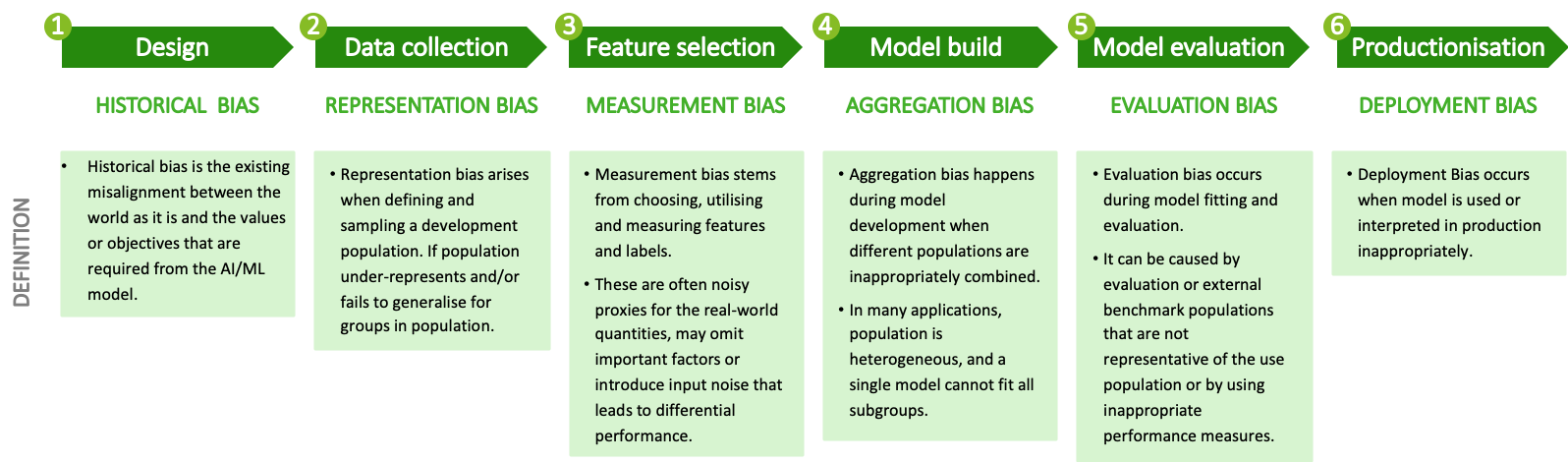
Bias in model development lifecycle

**Project Title**: Bias in model development lifecycle questionnaire

**Lead researcher:** *Michelle Seng Ah Lee (*[*sal87@cam.ac.uk*](mailto:sal87@cam.ac.uk)*)*

This survey will go through an assessment to identify the potential biases in a model development lifecycle.  
  
Think of a model that you would like to assess for potential unfair bias and discrimination.  
  
Below are the biases and their definitions for your reference.





**Ethical risk assessment: design KPIs/KRIs**

1. Describe a model you would like to assess for potential unfair and discriminatory bias. Ex) a supervised machine learning model to predict whether a mortgage loan will default
2. What positive impact can this model have on the target population? Ex) predicting default risk can help prevent unaffordable loans being approved
3. What is the benefit of higher accuracy / precision for the target population? Ex) better credit risk evaluation model leads to greater financial inclusion, better hiring algorithm leads to overall higher employee performance / reduction in attrition
4. Can these objectives be measured and quantified? If yes, list how they can be formalised. Ex) unaffordable loan approval can be measured based on false negative rate (i.e. loans predicted to be repaid but defaulted), and greater financial inclusion can be measured as the total amount of loans given out
5. What are any potential allocative harms (withholding of opportunities / resources)? Ex) model may be more likely to give loans to certain groups, e.g. race and gender, which would replicate and widen the societal inequalities
6. Is there any representational harm (diminished identity)? Ex) for an image search algorithm for “CEO”, returning more men than women reinforces the bias in identity
7. Are there any fundamental rights at stake? Ex) right to self-determination, liberty, due process of law, freedom of movement, privacy, freedom of thought, freedom of religion, freedom of expression, right of peaceful assembly, right to freedom of association
8. Can these potential harms be measured and quantified? If yes, list how they can be formalised. Ex) we can measure the loan denial rates for previously disadvantaged groups (e.g. minority race)

Design: historical/external bias

1. Is there documented historical discrimination in the domain area against a protected class, as defined in the UK Equality Act? Ex) academic studies demonstrate lower mortgage approval rates for racial minorities in the US, especially black and Hispanic applicants
   * age
   * disability
   * gender reassignment
   * marriage or civil partnership (in employment only)
   * pregnancy and maternity
   * race
   * religion or belief
   * sex
   * sexual orientation
2. Describe which groups are at a disadvantage
3. For that group, select which of the following layers of inequality is a justifiable source of differences in outcome. The justification may include 1) well-founded causal relationship to an outcome of interest (e.g. income to risk of default) and/or 2) a feature that is within the individual’s control and transparently disclosed (e.g. history of paying bills on time)
   * Disability
   * Race
   * National origin
   * Socioeconomic status
   * Talent/ education
   * Personality traits
   * Preferences
   * Culture
   * Discrimination in related markets
4. Which of the features in your data may be associated with the unjustifiable sources of differences in outcome, e.g. postcode with race?
5. Is there any misalignment between the ground truth (world as-is) and the organisation’s values? For example, there may be more male senior executives, but the organisation’s objective is to have stronger female representation in leadership.

Data collection: Representation bias

1. Is the marketing / targeting / data collection strategy returning a representative sample of the population? Ex) is the mortgage company advertised in majority-white neighborhoods?
   * Yes
   * No
2. Are any of the recorded features affected by human judgment? Ex) the data set may include the interviewer’s scores on the candidates’ performance.
   * Yes
   * No
3. Are any of the recorded features produced by a third party data set or model? Ex) the credit scores may be provided by a specialist agency
   * Yes
   * No
4. Is the ground truth of actual outcomes known? Ex) whether denied loans would have defaulted is unknown
   * Yes
   * No
5. Is there sufficient sample in each subgroup of interest for this analysis? Ex) only 5% of applicants are Native Americans
   * Yes
   * No

Feature engineering: measurement bias

1. Are there differences in the measurement process between groups for either input features or the target outcome? Ex) high-minority neighborhoods are more frequently patrolled, leading to higher arrest rates
   * Yes
   * No
2. Are there differences in the data quality between groups? Ex) schools in poor districts have lower quality recorded data on student performance
   * Yes
   * No
3. Are there any features added by the model developer that could it be affected by his/her judgment? Ex) the data scientist added flags of what he/she considers an important feature from a job applicant’s CV, e.g. “participated in university extracurricular activities” or “held a leadership position”
   * Yes
   * No
4. Are there proxies of outcome that may be also proxies of a protected group membership, especially those with a history of discrimination in this domain area? Ex) In US mortgage lending, employees were more likely to recommend loan types that are expensive to finance to black and Hispanic applicants, which would affect their recorded loan type.
   * Yes
   * No
5. Does the measurement closely match what the model intends to track? Ex) arrest ≠ crime rate, final grades ≠ student success
   * Yes
   * No

Model build and training: aggregation bias

1. Are the populations heterogeneous in a way that a single model cannot account for all subgroups? (See: Simpson’s paradox) Ex) Medical diagnosis algorithm should be different for men and women given their different body compositions
   * Yes
   * No
2. Are there other heterogeneous mechanisms in play that are being inaccurately aggregated that may be associated with protected features? Ex) differences in behavior across products, different time periods, different data sets, etc.
   * Yes
   * No
3. Have you tested on the fairness metric of choice on all protected subgroup combinations? Ex) the model has similar error rates for men and women and for black and white applicants, but it has high error rates for black women
   * Yes
   * No

Model evaluation: evaluation bias

1. Are all trade-offs on objectives identified for all available models? All objectives should be quantified into metrics where possible to enable model comparison Ex) mapped the trade-off between financial inclusion and minority race denial rates for mortgage lending for 10 versions of predictive models
   * Yes
   * No
2. Do the metrics cover all measurable objectives related to positive and negative impacts on the target population? Ex) The assessment of mortgage default prediction algorithm covers unaffordable loans (false positive), financial inclusion, minority race denial rate. It also includes qualitative assessment of explainability.
   * Yes
   * No
3. Do your metrics align with the relative importance of False Positives vs. False Negatives? Ex) those predicted to repay but defaulted represent unaffordable loans / cost to the company, and those predicted to default but would have repaid represent missed opportunity / allocative harm
   * Yes
   * No
4. Is there a metric the model may be over-fitting to? Ex) the main credit risk evaluation accuracy metric
   * Yes
   * No
5. Are there any disparities in sub-group performance? Ex) Does the model have more errors among women than men?
   * Yes
   * No
6. Are confidence intervals acceptable and understood? Ex) Especially if a sub-group population is under-represented, they may have a larger confidence interval around their predictions
   * Yes
   * No

Model productionisation and monitoring: deployment bias

1. Is the model a part of a complex sociotechnical system, e.g. inter-connected models or embedded in human processes? Ex) A CV-scoring algorithm may feed into a candidate’s evaluation system
   * Yes
   * No
2. Is there an appropriate human feedback mechanism for any errors? Ex) A human reviewer reads a sample of machine transcriptions to identify any errors and retrains the algorithm with the corrections
   * Yes
   * No
3. Is the model robust to any external changes, e.g. shifts in policy, dramatic changes in input data, etc.? Ex) There is a monitoring mechanism in place to alert the team if there is a significant change in the distribution in the input data beyond a pre-defined threshold
   * Yes
   * No
4. Can the feedback loop be reinforcing any existing biases? Ex) if loans predicted to default are denied. Is there any user interaction with the output? Ex) user clicking on links recommended by the algorithm
   * Yes
   * No