Data/AI Ethics Framework

Part 1: Risk Assessment

Last Updated: 06.15.2020

# **Overview**

**AI Miti**

We define steps to determine the inherent data/AI risks. We can broadly classify four overarching sets of risks that must be evaluated when using algorithms:

1. **Impact** examines the algorithm in terms of the effects it will have on people and property.
2. **Appropriate Use** inspects the relationship between the data being used in the algorithm and the purpose for which the data was collected and perceptions of the anticipated use.
3. **Accountability** surfaces how much involvement people have in the ongoing use of the algorithm, including whether automated decisions can be clearly explained to anyone.
4. **Bias** explores the underlying influence of the data and the people who helped build the algorithm.

In this part of the toolkit, we will evaluate each of these risks through a series of steps. Each step explores an individual factor. These factors are then rolled up to help provide stakeholders with a comprehensive score to arrive at a suitable risk mitigation recommendation (Part 2 of the toolkit).

# **Step 1: Understanding and Assessing Impact**

In order to quantify Impact, we project it as a function of four dimensions:

1. **Type.** This is used to classify the impact and defines the nature of the impact.
2. **Degree.** This is the level of the impact from negligible to major. For example, deciding the bail of an incarcerated person would be considered major.
3. **Scale.** This is how many people, places, or things are affected.
4. **Direction.** This is whether the impact is positive or negative. Most algorithms will have both positive and negative impacts. For example, an algorithm meant to connect persons experiencing homelessness to affordable housing positively affects those selected for housing, but negatively impacts those excluded.

The below steps will walk us through:

* Identifying who or what will be impacted and
* An analysis of impact along the four dimensions of type, degree, scale and direction

## **1.A Describe the impact**

### 1.A.1 Identify who or what will be impacted

To identify who or what might be impacted, let us define the proximity of impact:

* **Primary.** These are the immediate objectives of the algorithm, that is the people, places or things the algorithm provides input into.
* **Secondary.** These are the people, places or things that may feel the results of the algorithm as a function of its impact on the primary impactees.
* **Unexpected/unintended.** These are the people, places or things that may feel unintended or unexpected impacts from the algorithm. While you may not know these, you can take time to brainstorm them.

The table below provides some examples on primary, secondary and unexpected/unintended impactees.

### 1.A.2 Identify the types of impact

1. The table below describes the different types of impact. One type of impact may implicate another.
2. We should identify the type of impact for each group we identified in Step 1.1.1.

|  |  |
| --- | --- |
| **Type** | **Description** |
| Access to goods, benefits or services | These types of algorithms inform who, what or where does or does not receive access to goods, benefits or services. This can include access to insurance, government benefits, housing opportunities, education, maintenance or prevention services, recreation etc. |
| Financial | These types of algorithms impact the financial health of individuals, groups, entities or areas. |
| Property or equipment | These types of algorithms impact the quality or value of property or equipment. |
| Reputation | These types of algorithms impact the reputation of an individual, group, entity, or location. |
| Emotional | These types of algorithms impact the emotional health and well-being of an individual or group of individuals. |
| Life / safety | These types of algorithms impact the life or safety of an individual, group, entity, or location. |
| Privacy | These types of algorithms impact the privacy of an individual or group. |
| Liberty / freedom | These types of algorithms impact the liberty / freedom of an individual, group, or entity. |
| Rights / intellectual Property | These types of algorithms impact the rights / intellectual property of an individual, group or entity. |

## **1.B Assess scope of impact**

Scope of impact is a function of both the degree and scale of impact.

### 1.B.1 Rate the degree of impact

1. Now that we have identified the type(s) of impact our algorithm has; we can rank their relative impact. The table above describes impact levels for each type ranging from “No discernable” to “Major” impact.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Type** | **No discernable** | **Minor** | **Moderate** | **Major** |
| Access to goods, benefits or services | No differential access to goods, benefits or services | Minor differential access to goods, benefits or services | Moderate differential access to goods, benefits or services | Major differential access to goods, benefits or services |
| Financial | No financial impact | Minor financial impact | Moderate financial impact | Major financial impact |
| Property or equipment | No damage, improvement or change in value | Minor damage, improvement or change in value | Moderate damage, improvement or change in value | Major damage, improvement or change in value |
| Reputation | No change in reputation | Minor change in reputation | Moderate change in reputation | Major change in reputation |
| Emotional | No emotional impact | Minor emotional impact | Moderate emotional impact | Major emotional impact |
| Life / safety | No effect on life, physical well-being or safety | Minor effect on life, physical well-being or safety | Moderate effect on life, physical  well-being or safety | Major effect on life, physical well-being or safety |
| Privacy | No effect on privacy | Minor effect on privacy | Moderate effect on privacy | Major effect on privacy |
| Liberty / freedom | No change in liberty  / freedom | Minor change in liberty / freedom | Moderate change in liberty / freedom | Major change in liberty / freedom |
| Rights / intellectual property | No change in property or intellectual rights | Minor change in property or intellectual rights | Moderate change in property or intellectual rights | Major change in property or intellectual rights |

### Step 1.B.2 Get the scale of impact

Now you can assess the scale of impact. Is this a few people, things or places or many? Use the table below to estimate the scale of impact for each area of impact from Step 1.1.1.

|  |  |
| --- | --- |
| **Scale** | **Description** |
| Small | This algorithm impacts very few people, places or things in our jurisdiction. |
| Medium | This algorithm impacts a substantial number of people places or things in our jurisdiction. |
| Large | This algorithm impacts nearly every people, place or thing in our jurisdiction and may impact those outside. |

### Step 1.B.3 Assign scope estimate

Use the degree and scale of impact to assign a scope estimate.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Scope Estimate** | | **Scale of Impact** | | |
| Small | Medium | Large |
| **Degree of Impact** | No discernable | Very narrow | Very narrow | Limited/Narrow |
| Minor | Very narrow | Limited/Narrow | Substantial |
| Moderate | Limited/Narrow | Substantial | Broad/wide ranging |
|  | Major | Substantial | Broad/wide ranging | Broad/wide ranging |

## **Step 1.C Estimate the overall direction of impact**

Each type of impact may be either positive, negative or both, regardless of intensity.

For example, access to a benefit may be good for an individual whereas identifying areas to target for surveillance may be both bad and good. In either case, there are at least two groups affected differentially

* those that do or don’t receive the benefit or those that are or are not targeted. So your algorithm will often impact two or more groups and in two different directions.

Nonetheless, you should assess the overall direction of impact. This will help you in later sections as you weigh the steps you should take to improve the responsible and ethical use of your algorithm.

* + **Positive.** Provides an overall positive impact, does not result in differential access (e.g. some miss out) or negative changes or impacts, and does not take away from another group or area.
  + **Mostly positive.** Provides a positive impact to some but does not take away from another group or area. Some will not benefit, but no one will be harmed.
  + **Mostly negative.** Provides a negative impact to some and may remove or take away from another group or area.
  + **Negative.** Provides or allocates mostly negative impacts, removes or takes away from any groups or areas it applies to.

## **Step 1.D Assign overall impact risk**

Combine the scope of impact estimate from Step 1.B.3 with the overall direction estimate from Step 1.C to estimate the overall impact risk from the algorithm.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Overall Impact Risk** | | **Overall Direction** | | | |
| Positive | Mostly Positive | Mostly Negative | Negative |
| **Scope** | Very Narrow | Very low | Very low | Low | Moderate |
| Limited/Narrow | Very low | Low | Moderate | Significant |
| Substantial | Low | Moderate | Significant | High |
| Broad/wide ranging | Moderate | Significant | High | Extreme |

# **Step 2: Assess appropriate data use risk**

**AI Miti**

Appropriate use in this section focuses on the question: should you use the data for the purposes of this algorithm? Step 2 focuses on can you use the data from the perspective of representativeness and accuracy.

Data inputs need to be evaluated in the contexts of consistency and compatibility and reputation and perception. This will help us understand the “ethical risk” inherent in using sources of information for the intended algorithm.

## **Step 2.A Consistency and Compatibility of use**

For what purpose were the data inputs originally created, collected or obtained? How compatible is the new use with the original reason for data collection? We use the table below to score the consistency and compatibility of our intended use.

|  |  |
| --- | --- |
| Consistency and Compatibility | Description |
| Yes | Our use of the data for this algorithm is consistent and compatible with the purposes and context under which the data was obtained. This includes applicable laws and regulations. |
| Somewhat | Our use of the data for this algorithm is somewhat consistent and compatible with the purposes and context under which the data was obtained. |
| Unknown | We are not familiar with the purpose and context for how this data was obtained. Can we trust the data because we don’t know how it was collected? |
| No | Our use of the data for this algorithm is not consistent and compatible with (or prohibited by) the purposes and context under which the data was obtained. |

## **Step 2.B Reputation and Perception from use**

What are the reputational and perceptions risks from our use of this data for the purposes of this algorithm? If this is public knowledge, how will people react? We Use the table below to classify the expected response. In general, use of data about individuals will have greater reputation and perception risks.

|  |  |
| --- | --- |
| Reputation and Perception | Description |
| Supportive | Most people would agree with our use of this data for the intended purposes of the algorithm. Though, as with any public endeavor, some will disagree with this use. Usage of the data for this specific purpose is defensible with precedence. Open Data |
| Mixed | We expect several groups of people would be concerned with our use of this data for the intended purposes of the algorithm. The is a common practice that has not been legally challenged. Defensible without precedence. |
| Not supportive | We expect most people would object to our use of this data for the intended purposes of the algorithm. Arguably defensible to achieve goals. |

## **Step 2.3 Assign appropriate use risk score**

Using the prior two steps to assign an appropriate use risk score.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Appropriate Use Risk Score | | Reputation and Perception | | |
| Supportive | Mixed | Not Supportive |
| Consistency and Compatibility | Yes | Low | Low | Medium |
| Somewhat | Low | Medium | High |
| Unknown | Medium | Medium | High |
|  |  | No | Medium | High |

# **Step 3: Assess accountability risk**

**AI Miti**

Accountability in the use of algorithms can be addressed by exploring the following questions:

1. Who or what made what decisions?
2. How were those decisions made?
3. How do we explain those decisions? Or can we explain those decisions?
4. How can we review or audit those decisions?
5. How can we modify those decisions if there is disagreement?
6. Specifically, to the algorithm:
   1. How did we test the algorithm before we put it in use?
   2. How do we ensure the algorithm is working as intended?
   3. How do we track performance of the algorithm?
   4. How do we modify the algorithm over time?

In the sections below, we address the first four questions by rating the accountability risk of the algorithm. Part 2 of this toolkit provides best practices to follow during the development of an algorithm to address questions 5 and 6.

## **Step 3.1 Determine automation score**

Using the table below to identify the level of automation in the decision making or action that the algorithm informs.

|  |  |
| --- | --- |
| **Score** | **Description** |
| Low - human mediated | The algorithm is being used to inform an individual or group of individuals. Ultimately, a human is making the final assessment. The algorithm does not include strong recommendations or make conclusions (e.g. policy decisions, risk factors, etc.). |
| Medium - algorithm mediated | The algorithm structures, constrains or otherwise makes recommendations for actions or decisions. The action or decision is ultimately made by an individual or group of individuals (e.g. sentencing, bail, etc.). |
| High - algorithmically determined | The algorithm automatically takes actions or makes decisions with no interference by a person or group (e.g. red light cameras, traffic flow management, inspection prioritization, etc.). |

## **Step 3.2 Determine accessibility score**

The accessibility score is a function of how easy it is to:

* + Explain the algorithm and
  + Audit and review it

### Step 3.2.1 Determine explainability score

Using the table below to rate how easy it is to explain the algorithm and how it works. (] “How well can I explain to a layperson?”)

|  |  |
| --- | --- |
| **Explainability** | **Description** |
| Easy | The algorithm is straightforward to explain and does not require sophisticated understanding of statistics and modeling techniques. |
| Medium | The algorithm can be explained but does require more understanding or careful explanation of statistical and modeling techniques. |
| Hard | The algorithm is challenging or even impossible to explain even to sophisticated users (e.g. “black box”). |

### Step 3.2.2 Determine auditability score

Using the table below to describe how easy it is to review or audit the algorithm function and inputs / outputs. How will the algorithm produce each / any specific result?

|  |  |
| --- | --- |
| **Auditability** | **Description** |
| Easy | We can access to audit and review the algorithm as needed and have a means to do so. |
| Medium | If we need to, we can access the algorithm to audit and review it. We need to figure out what a meaningful audit and review would look like. |
| Hard | We have no access to the algorithm or how it functions. We have no feasible means for determining how we would audit and review it. |

### Step 3.2.3 Assign accessibility score

Using the explainability and auditability scores to assign an accessibility score.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Accessibility Score** | | **Explainability** | | |
| Easy | Medium | Hard |
| **Auditability** | Easy | Accessible | Accessible | Some concerns |
| Medium | Accessible | Some concerns | Major concerns |
| Hard | Some concerns | Major concerns | Major concerns |

## **Step 3.3 Assign accountability risk**

In this step, you we the automation and accessibility scores to identify the level of accountability risk posed by the use of the algorithm.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Accountability Risk** | | **Automation Score** | | |
| Low - human mediated | Medium - algorithm mediated | High - algorithmically determined |
| **Accessibility Score** | Accessible | Low | Low | Medium |
| Some concerns | Low | Medium | High |
|  | Major concerns | Medium | High | High |

# **Step 4: Assess Algorithm Methodology risk**

**AI Miti**

## **Step 4.1 Answer third party methodology questions**

Below is a list of questions to help assess our algorithmic process. Each question occurs at one of the following stages: design, monitoring, or incorporation. Each of the questions below should be answered with a “yes” or a “no”.

|  |  |  |
| --- | --- | --- |
| **Stage** | **Question** | **Point if yes** |
| Design | We are the direct owners of the algorithm (it was developed in-house rather than through a third party). | 1 |
| Design | We involved subject matter experts in the design of the algorithm. | 1 |
| Design | Assumptions made by the creators were outwardly explained. We know the motives of the developer or vendor. | 1 |
| Design, Monitor | We have discussed the proposed outcomes of the algorithm with a diverse audience. | 1 |
| Design, Monitor | We periodically review decisions the algorithm has made and revise it to meet changing needs. | 1 |
| Design, Monitor | We have a way to rebuild and/or re-train the algorithm from the ground up when new variables are introduced. | 1 |
| Monitor | We monitor the algorithm on a regular basis to ensure it is operating the way we intend. | 1 |
| Incorporate | We have piloted/tested the algorithm against a subset of real-world decisions before fully deploying it to influence all decisions. | 1 |

## **Step 4.2 Assign methodology risk level**

|  |  |
| --- | --- |
| **Total Points** | **Methodology Risk** |
| 6-8 | Low |
| 3-5 | Medium |
| 0-2 | High |

1 point for each statement where we answered “yes”. Using the table below to determine our methodology risk.

# **Step 5: Historic Bias Risks**

In this step, we identify the level of risk due to historic framing of the data used in our algorithm. Inherent Data Bias.

Consider both training data (i.e. data used when originally training the algorithm) and data used to feed the algorithm when it is in use (i.e. ongoing and future data).

|  |  |
| --- | --- |
| **Historic Bias Risk** | **Description** |
| Low | We have thoroughly researched context. Data is completely separate from any documented or well-known societal strife or controversial social topic. For example: an algorithm used by a content-streaming service to decide only a user’s potential movie preference is likely not historically biased. Data is recent (0-10 years-old). |
| Medium | We have moderately researched context. Data is slightly connected to documented or well-known societal strife or controversial social topic. For example: a natural language processing algorithm trained on older marital survey data would likely be historically biased against same-sex couples due to the unfair, discriminatory, and formerly legal practices baked within older data collection strategies. Data is fairly recent (11-25 years-old). |
| High | We have not researched context. There are negative historical connotations associated with the data. Data is deeply connected to documented or well-known societal strife or controversial social topic. For example: a housing placement algorithm that has been trained on decades-old housing data would likely be historically biased against black people due to reflections of discriminatory redlining present in older data. Data is old (26-50+ years-old). |

# **Step 6: Technical Bias Risks**

In this section, technical bias represents only bias surrounding data accuracy and data representativeness (or lack of). Technical bias in the use of algorithms can be addressed by exploring the following questions:

1. What is the quality of the data to be used?
2. How accurately does the data represent real-world conditions?
3. During development, was the algorithm’s methodology closely monitored, and by whom?
4. Who was involved in the development, and how were they able to contribute?
5. Where did the training/tuning data come from? Is this source appropriate for the context in which the algorithm will be used?

## **Step 6.1 Sampling Bias Risks and Data Quality risks**

Source data includes both training data (i.e. data used when originally training the algorithm) and that which is used to feed the algorithm when it is in use (e.g. real, live data). In both cases, you need to assess sample bias and data quality.

Step 6.1.1 Assess representativeness risk

In this step, identify the level of risk due to representativeness for the data used in your algorithm. Does the sample data represent your population?

|  |  |
| --- | --- |
| Sample Representativeness Risk | Description |
| Low | Data is “[progressive](http://methods.cochrane.org/equity/projects/evidence-equity/progress-plus)” as it represents the population as a whole regardless of subgroup. |
| Medium | Data over or under represents some subgroups and we have a sense of who/what/where is over/underrepresented. We may be using variables that are not direct measures of the what we care about (proxies). |
| High | Data is not representative (for example: 311 data is biased to those who call 311). Use of data may lead to circular results, i.e. self-fulfilling prophecy or can only be used to study a particular subgroup. We are using variables that are poor proxies of what we are trying to measure. Any results should not be extrapolated or applied to the larger population. |

Step 6.1.2 Data Quality risk

In this step, identify the level of risk due to quality for the data used in your algorithm. Think about how the data were collected or acquired, and identify potential sources of error from training, validation, data inconsistency, subpar collection methods, etc.

|  |  |
| --- | --- |
| Quality Risk | Description |
| Low | Data is highly structured, with strong validation, training and consistency of collection. Data collection is automated, highly structured, and easily validated. |
| Medium | Some of the data collection is automated and some is input manually or based on other human input. Validation is difficult or is used but errors can happen. |
| High | Data is not well structured, validation is not used, lack of training or inconsistent data collection methods. |

Step 6.1.3 Assigning representativeness and inaccuracy risk score

Combine your risk scores for each second to choose an overall risk score for both bias and inaccuracy.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Representativeness and Inaccuracy Risk Score | | Representativeness Risk | | |
| Low | Medium | High |
| Inaccuracy Risk | Low | Low | Low | Medium |
| Medium | Low | Medium | High |
| High | Medium | High | High |

## **Step 6.2 Arriving at the Overall Technical Bias Risk**

Combining the methodology risk matrix with the Technical Bias Risk matrix to arrive at the Overall technical Bias matrix.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Overall Technical Bias Risk | | Methodology Risk | | |
| Low | Medium | High |
| Representativeness and Inaccuracy Risk Score | Low | Low | Low | Medium |
| Medium | Low | Medium | High |
| High | Medium | High | High |