A close up of a logo

Description automatically generated Data/AI Ethics Framework

Part 2: Risk Mitigation

Last Updated: 06.15.2020

# **Overview**

**AI Miti**

In the first section of the toolkit, we assessed a variety of risk factors for a set of algorithms we plan to implement. Using the results from that assessment along with this section, we will identify the appropriate mechanisms to help mitigate some of the risks.

Please note:

* Individual mitigations may be useful for multiple risks.
* Not all risks or levels have specific mitigations.
* Some risk subcomponents are included, but others are not.
* Each risk level builds on the previous mitigations. So, if you have a ​**high** **​**risk

factor, you should apply the mitigations for ​**low** **​**and ​**medium** **​**as well.

# **Risk Mitigation Recommendations**

**AI Miti**

1. **Step 1.3 “Impact scope estimate”**
2. If the scope of the impact is very low​ **​**or​**limited/narrow​​**, engage impacted communities (​Recommendation 1).
3. If the scope of the impact is **substantial**​**​**, use public performance monitoring (​mitigation 2​).
4. If the scope of the impact is very high create an IRB (​mitigation 3​) or some other public advisory group with decision-making authority for the program (​mitigation 4​).
5. **Step 1.4 “Overall impact risk”**
6. If the overall impact is Very low **​**, **low**​**​**, or **moderate**​**​**, engage impacted communities (​mitigation 1​).
7. If the overall impact is **significant**​**​**, use public performance monitoring (​mitigation 2​).
8. If the overall impact is **Very HIgh**​**​**, create an IRB (​mitigation 3​) or some other ethics advisory group with decision-making authority for the program (​mitigation 4​).
9. **Step 2.3 “Appropriate data use”**
10. If **low**​**​**or **medium**​**​**, create a dialogue with the customers about the new uses of the data as they are applied to algorithms (​mitigation 5​).
11. If ​**high​​**, find or create alternate data sources to replace inappropriate ones (​mitigation 6​).
12. **Step 3.3 “Accountability”**
13. If **low**​**​**or **medium**​**​**, use automated testing tools to periodically evaluate algorithm performance (​mitigation 7​), ensure there is a human adjudication mechanism (​mitigation 8​), and require human intervention before executing each algorithmic decision (​mitigation 9​).
14. If ​**high​​**, ensure human adjudication mechanism results feed into algorithm tuning (​mitigation 8​), ensure the relevant inputs and machine state(s) are captured in perpetuity for each decision (​mitigation 10​), and evaluate human-intervened decisions periodically (​mitigation 11​).
15. **Step 5.1 “Historic bias”**
16. If **low**​**​**or **medium**​**​**, tune the algorithm to systematically minimize bias impact / compensate for missing data (​mitigation 15​).
17. If ​**high​​**, ​do not​ use the data (​mitigation 6​) and find alternate proxies with accurate biases (mitigation 16).
18. **Step 6.1.3 “bias and inaccuracy”**
19. If ​**high​​**, run a data management improvement project (​mitigation 16​) or find another source of data (​mitigation 6​).
20. **Step 6.4 “overall bias”...**
21. If **low**​**​**or **medium**​**​**, define clear measures of bias and monitor your program over time to ensure that it / they does / do not increase (​mitigation 18​).
22. If ​**high​​**, compare pre-existing bias to predicted bias (​mitigation 19​).

**Mitigations in detail**

**AI Miti**

Mitigation 1. ​Effective community engagement is people-centered, partnerships-driven, and power-aware. Engagement with the community should be social (using existing social networks and connections), technical (skills, tools, and digital spaces), physical (commons), and on equal terms (aware of and accounting for power). An example of engaging impacted communities around open data could look like: the co-production of a policy and open data prioritization, the public creating innovative tools from raw data, and the public then interacting and engaging with data apps and visualization tools.

Mitigation 2. ​The purpose of public performance monitoring is to identify areas of good performance and areas where performance can be improved. Performance information should be focused (on the agency’s objectives and services), appropriate (to, and useful for, the stakeholders who are likely to use it), balanced, (giving a picture of what the agency is doing, covering all significant areas of work), robust (in order to withstand organizational changes or individuals leaving), integrated (into the organization), and cost-effective (balancing the benefits of the information against the costs).

Mitigation 3.​An institutional review board (IRB) is a traditional committee established to review and approve applications for research projects. An IRB can also exist in non-academic circles, and its committee members can serve as a necessary step before an algorithm is implemented.

Mitigation 4. ​Public advisory groups are typically comprised to key stakeholders related to a project as well as representatives of the general public, selected to inform the development of a project.

Mitigation 5. ​Starting a dialogue with the public about new uses of data could be as simple as creating a survey and surveying residents, sending out a weekly or biweekly memo or newsletter to inform residents of new uses, holding town hall meetings in order to discuss the data, publishing open data online, and/or maintaining a public Github.



Mitigation 6. ​Stop the controversy before it starts: Do not start a project with data that has the potential to be harmful. Find or create new data sources by completing a data inventory to locate more appropriate data, researching your topic online to find new data, or collecting new data.

Mitigation 7.​Automating testing tools (i.e. confusion matrices when evaluating classification models) to evaluate an algorithm’s performance can be a way to integrate systematic checks into the lifecycle of an algorithm. If the aforementioned classification model is falsely classifying 70% of cases, the automated testing tool can be programmed to produce “STOP” in red letters.

Mitigation 8. ​A human adjudication mechanism, being a process through which a person can introduce his or her own discernment, can be a great addition to a project involving an algorithm or algorithms. Ensuring that this mechanism can then feed into the tuning of an algorithm can be a great addition to the project.

Mitigation 11. ​Evaluate human-intervened decisions periodically to control for unintended rater bias.

Mitigation 15. ​Missing data can be a source of statistical inaccuracy in any project. Missing data has the potential to greatly exacerbate harmful bias in the context of algorithms. If you are aware that your algorithm is using data that is largely comprised of missing values, make sure that your algorithm has a way to systematically account for these values. For example: If your dataset is small, you might elect to weight certain demographics within the data in order to more accurately reflect a general population.

Mitigation 16. ​Running a data management improvement project could consist of: Creating a data governance structure, creating an open data policy, running a new data inventory, constructing an open data portal, committing to a new data publication process or data standards, systematically testing data quality, adhering to a new data retention policy or privacy and security policy, engaging with the community around the data, or hiring new staff and talent.

Mitigation 17. ​Recruiting algorithmic auditors (statisticians, data analysts, data science professionals, computer scientists, etc.) to audit for the influence of certain factors, variables, or covariates might be very helpful. You might find it helpful to have these auditors routinely return to your algorithm and run a systems check. After rigorously auditing your algorithm, what have they concluded?

Mitigation 18. ​Being clear and full of intent is highly important, regardless of context. In the context of an algorithm, define clear measures of bias and then decidedly monitor your program over time to ensure that it or they does or do not increase.

Mitigation 20. ​Ensure that program managers can understand and are able to sign off on the risk profile that is reflected by your algorithm. Can they explain each risk, and do they understand who or what this risk or these risks might affect?