Ethics and Algorithms Toolkit (Beta)

(Section) Part 2: Manage Algorithm Risk

Overview of toolkit beta release

Welcome to the beta release of our Ethics and Algorithms Toolkit! This toolkit is designed to help governments (and others) use algorithms responsibly.

Who is this toolkit for?	If you are building or acquiring algorithms in the government sector this toolkit is for you. Though we expect others will find it useful.
What is the toolkit?	The toolkit is really a process. It walks you through a series of questions to help you 1) understand the ethical risks posed by your use of an algorithm and then 2) identify what you can do to minimize those ethical risks.
What are the parts of the toolkit?	The toolkit comes in several parts: 1. The introduction to the toolkit (this document) 2. Part 1: Assess Algorithm Risk 3. Part 2: Manage Algorithm Risk 4. Appendices (including a handy worksheet)
Who made this toolkit?	The beta release was a collaboration between The Center for Government Excellence (GovEx) at Johns Hopkins, the City and County of San Francisco, Harvard DataSmart, and Data Community DC.
How can I give feedback?	Send feedback on our beta release at http://labs.centerforgov.org/toolkit/.

Part 2: Manage Algorithm Risk

In the first section of the toolkit, you assessed a variety of risk factors for a set of algorithms you plan to implement. Using the results from that assessment along with this section, you 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.

Instructions: Select and implement the mitigation mechanism that corresponds to each risk level you selected (return to *Ethics & Algorithms Toolkit: Part 1* for your selections from corresponding steps). You may underline, circle, or check mitigations that apply to your scenario. You will find this part of the toolkit

split into two pieces: **risk-to-mitigation matching** (a quick look at strategies) and **mitigations in detail** (in-depth explanations of those strategies).

Risk-to-mitigation matching

For Step 1.3 “scope estimate”...

If you selected **very narrow** or **limited/narrow**, engage impacted communities (mitigation 1).

If you selected **substantial**, use public performance monitoring (mitigation 2).

If you selected **broad/wide-ranging**, create an IRB[1] (mitigation 3) or some other public advisory group with decision-making authority for the program (mitigation 4).

For Step 1.4 “rank overall impact risk”...

If you selected **very low**, **low**, or **moderate**, engage impacted communities (mitigation 1).

If you selected **significant**, use public performance monitoring (mitigation 2).

If you selected **high** or **extreme**, create an IRB (mitigation 3) or some other public advisory group with decision-making authority for the program (mitigation 4).

For Step 2.3 “appropriate data use”...

If you selected **low** or **medium**, create a dialogue with the public about the new uses of the data as they are applied to algorithms (mitigation 5).

If you selected **high**, find or create alternate data sources to replace inappropriate ones (mitigation 6).

For Step 3.3 “accountability”...

If you selected **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).

If you selected **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).

---

[1] See the Oregon State University’s IRB definition and the U.S. Department of Health & Human Services’ IRB registration instructions for more information.

For Step 4.2 “third party”...

If you selected low or medium, transfer liability risk to contractor (mitigation 12) and implement independent monitoring through an internal or 2nd third party (mitigation 13).

If you selected high, include contractor incentives for desired outcomes (mitigation 14).

For Step 5.1 “historic bias”...

If you selected low or medium, tune the algorithm to systematically minimize bias impact / compensate for missing data (mitigation 15).

If you selected high, do not use the data (mitigation 6) and find alternate proxies with accurate biases (mitigation 16).

For Step 6.1.3 “bias and inaccuracy”...

If you selected high, run a data management improvement project (mitigation 16) or find another source of data (mitigation 6).

For Step 6.2.3 “training data”...

If you selected high, find a more appropriate source of data (mitigation 6).

For Step 6.3 “improper methodology”...

If you selected high, recruit algorithmic auditors to audit for influence of factors, variables, or covariates (mitigation 17).

For Step 6.4 “overall risk of bias”...

If you selected 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).

If you selected high, compare pre-existing bias to predicted bias (mitigation 19)..

Overall, if the majority of your selections were...

Low, ensure program managers understand and sign off on the risk profile (mitigation 20).

Medium, ensure system users (and impacted individuals) are aware that decisions are being made via automation (mitigation 21) and apply multiple algorithms to the same decision, favoring the decisions which lead to desired outcomes (mitigation 22).

High, delay implementation until risks can be reduced or benefits significantly outweigh the dangers (mitigation 23), create an IRB with decision-making authority for the program (mitigation 3), and have researchers periodically evaluate implementation (mitigation 11) and provide reports to the IRB (mitigation 3).

Mitigations in detail

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 13. Shifting the monitoring of an algorithm to an internal or additional third party adds one more level of subjectivity to an algorithm's methodology.

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 intentful 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?