Ethics and Algorithms Toolkit (Beta)

(Section) Introduction to the Toolkit

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/.

Introduction to the Toolkit

When you should use this toolkit

In short, whenever you are using an algorithm to inform a decision in the public sector. Below are just a handful of real-life scenarios where algorithms have been used:

- Working to make the restaurant inspections pipeline in Chicago more efficient
- Predicting the engagement of African American families in city services
- Determining a demographic to target for pre-K enrollment outreach
- Automating public assistance eligibility in Indiana

In most situations, the creators of algorithms intend for them to be additive and useful. Algorithms are used in our criminal justice system, employment arenas, higher education processes, and even social media networks. They are used to evaluate our teachers, rank our credit, insure our cars, and more. They

have the potential to classify, associate, or filter information using human-crafted and/or data-induced rules that allow for consistent treatment across populations.

All of these circumstances sound positive, so what is the problem? For one, algorithmic bias could have significant impact when government organizations decide how to distribute services or dole out justice based on the output of data-driven algorithms.

When do algorithms come into play? In the context of civic processes, example use cases include:

1. **Segmenting and targeting residents for individual-level results.** For example, a classification algorithm to determine which constituent base to contact for a new city initiative.
2. **Optimizing civic processes.** For example, using a processing software to automate the application process to the Supplemental Nutrition Assistance Program.
3. **Personal quantification and performance optimization.** For example, implementing a new regression algorithm for employees to measure their own performance.
4. **Improving healthcare and public health.** For example, creating a cancer risk scoring algorithm for constituents using their genetic data.
5. **Improving science and research.** For example, using a natural language processing algorithm to analyze large amounts of research literature.
6. **Optimizing machine performance.** For example, consistently updating an algorithm with new training data.
7. **Improving security and law enforcement.** For example, creating an algorithm that determines at-risk neighborhoods each month.

Where should you remain alert? The prior examples could pop up during in-house development, adoption, or acquisition. Algorithms can be developed internally or adopted through a third-party or developer. Typically, the in-house development of algorithms allows more control than acquired algorithms. The toolkit covers some tools to help implement control over acquired algorithms.

How can this toolkit help? This toolkit walks you through a practical set of questions that will help any person or organization be intentionally considerate and cautious in their use of algorithms.

High-level concepts and terms

We recognize that the topic of algorithms includes a bunch of technical jargon. Below are some high-level concepts and terms used in the toolkit that can help level-set you and your fellow toolkit users. There are many extensive explanations (and diagrams) online. This is just a quick set of ideas to get you started.

“Black box” vs. explainable artificial intelligence (xAI)

- “Black box” algorithms: Viewed in terms of its inputs and outputs without any knowledge of its internal workings.
- xAI: Aim to produce more explainable models while maintaining high prediction accuracy. Essentially, xAI systems work by simplifying information (analyzing various inputs used by a decision-making algorithm and reporting only on the set of inputs that had the biggest impact on the final decision).

Third-party built vs. in-house built

- Third-party built: Constructed by an outside, potentially impersonal source
- In-house built: Constructed within your organization

Automation vs. augmentation algorithms

- Automation algorithms: Replace and accelerate existing processes.
- Augmentation algorithms: Derived from subject matter experts' interaction with datasets.

Machine Learning

Machine learning is algorithms make predictions or calculated suggestions based on large amounts of data without being explicitly programmed by a person. Machine learning comes in a variety of flavors:

1. **Supervised learning.** In this case, the machine trains on labeled data (known inputs and outputs) and describes the relationship between the input and output. It then applies what it learns to new, unlabeled data -- i.e. linear or logistic regression, random forest classification.
2. **Unsupervised learning.** In this case, the machine does not have known inputs and outputs -- the dependent variable is unknown and there is no expected outcome. Instead, the machine surfaces what is important in the structure of the data -- i.e. clustering and association.
3. **Semi-supervised learning.** In this case, the machine akes use of both unlabeled and labeled training data and uses both types of techniques -- i.e. speech recognition.
4. **Reinforcement learning.** In this case, the machine's learning is "reinforced" through trial and error where each trial is either "rewarded" or not.

Functional Categories

There are many categories and classifications that breakdown the mechanics of algorithms, but the application of these algorithms is of primary concern to this toolkit. Secondarily, "how" an algorithm processes and executes its code also merits scrutiny because as ethical data practitioners, we need to understand and explain where bias can be introduced into an algorithm.

There are high-level concepts such as Artificial Intelligence (AI) and Data Science, each of which can be broken down by function and technological implementation in the table below.

Table: Functional Categories of Algorithms

High Level	Specific	Task	Input / Output	Risk / Ontology
Data science	Data collection	Convert real world into data	Surveys, sensors, studies, 3rd parties	Bias
Data science	Data munging	Normalization and clean up	Raw data → statistically useful data	Privacy

Data science	Data mining	Extract findings	Analysis, segmentation, personas	Inaccuracy
Data science	Data visualization	Inform insights	Clustering, trends	Bias
Data science	Dashboards & BI	Real time - inform decisions	Actionable trends & alerts	Bias
AI / data science	Natural language processing (NLP)	Interpret text	Text, voice	Inaccuracy
AI / data science	Machine learning	Determine correct action	Well-specified problem, big data	
AI / data science	Deep learning	Prediction models	Big data → output verification	