Ethics and Algorithms Toolkit (Beta)

(Section) Appendices

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/.

Appendix A: Data Questions

1. Creation.
    - Failing to work with the most comprehensive or appropriate data
        - *Did the data come from a system prone to human error?*
        - *What technology facilitated the collection of the data?*
        - *Does the context of the collection match the context of your use?*
        - *Was your data collected by people or a system that was operating with quotas or a particular incentive structure?*
        - *Who is represented in the data? Who is under-represented or absent?*
        - *Would this use of the data surprise the data subjects?*
        - *Are there any fields that should be eliminated from your data?*
    - Failing to include select / train the proper features
        - *Can you describe the logic that connects the variables to the output of your equation?*
        - *How did you determine what weight to give each variable?*
        - *What assumptions are you relying on to determine the relevant variables and their weights?*

    - Failing to assign subject matter experts to the architecture of an algorithm
        - *Is there a person on your team tasked specifically with identifying and resolving bias and discrimination issues?*
    - Failing to monitor sensitive relationships between inputs and anticipated outcomes
        - *Are there sensitive characteristics you need to monitor in your data in order to observe their effect on your outputs?*
        - *Will your variables apply equally across race, gender, age, disability, ethnicity, socioeconomic status, education, etc.?*
    - Failing to ask a diverse audience if your outcome expectations make sense to them?
2. Operation.
    - Implementing an algorithm so that it monopolizes or becomes responsible for an extremely personal, sensitive process *(algorithms should augment or supplement humans, not override or replace)*
3. Maintenance.
    - Failing to periodically revisit your methodology and iterate
    - Failing to update or re-train models when new data are introduced
    - Failing to process new data and variables with the same inquiry as the original model
        - *Are unintended factors or variables correlated with sensitive characteristics?*
    - Failing to ensure your system is behaving the way you intend
        - *What amount and type of error did you expect? Is it performing on task?*

Appendix B: Background (for Bias)

We are naturally skeptical of algorithms, and we are not necessarily incorrect by harboring this skepticism. Currently, research suggests that people believe that algorithms are less fair than humans (Lee, 2018). As beneficial as algorithms can be, there remain possibilities for bias and inaccuracy to breed detrimental outcomes. In truth, algorithms can harbor biases against disadvantaged groups or reinforce structural discrimination. And when algorithmic designers ignore social nuance or inequality, they risk designing systems that create disparate impacts.

For example, within the fourth "Consider Potential Scenarios" example listed earlier, unforeseen or insidious issues with the data might exist. Implementing an algorithm to "better public health" could be extremely unethical without considering possible sources of bias. Is the data being used to train the algorithm as "progressive" as it should be? Evaluating your data with this mindset could help you determine whether or not the data are too unfair for use. (Do the data represent enough of the population's demographics, so that we could train a robust, comprehensive model on them? Or should we continue to search for additional data?)

Algorithms have the potential to "mask deep seated biases behind the promise that the numbers will speak for themselves" (MIT Center for Civic Media). Why? Because people bring their own preconceived notions to any task. Machine bias is human bias: Human biases can feed into data biases which then feed into algorithmic biases. Algorithm and data-driven products will always reflect the design choices of the humans who built them—algorithms are mirrors. Consider this simple example from Medium:

> *"Suppose two people are tasked with developing a system to sort a basket of fruit. They have to determine which pieces are "high quality" and will be sold at the market, and which will instead be used for making jam. Both people are given the exact same data — the fruit — and the same task… Given the same task and data, the two people are likely to have different results. Perhaps*

one person believes the primary indicator of a fruit's quality is brightness of color. That person may sort the fruit based on how vibrant it is, even though not all fruits are brightly colored; that person would send strawberries to the market and melons to the jam factory. Meanwhile, the other person might believe that unblemished fruit is the best quality, even though fruits with protective rinds might look scruffy on the outside, but are perfectly fine on the inside; that person could send unripe strawberries to the market and ripe melons or bananas to the jam factory. These different, yet similarly logical and evenly applied criteria, will result in two different outcomes for the same basket of fruit."