Ethical Data Use at Code for America
Introduction

At Code for America, data is critical to our mission. We use data to:

1. **Show what’s possible.** Data helps us create and improve simple, accessible, dignified government services for people who need them most.
2. **Help others do it themselves.** Data helps us demonstrate the benefits of changes to processes or services, learnings that we then share externally.
3. **Build a movement.** Data helps us understand and support the interests of our community of partners, volunteers, donors, and friends.

To do our work ethically, we must continuously and rigorously evaluate how it affects people, and whether those effects are acceptable. Above all, we must do our best to ensure that the potential benefits to our users justify any possible risks.

Our vigilance extends in two directions:

1. **At the source:** We strive to safeguard our users’ personal data while maintaining the data resources that allow us to do our work.
2. **Downstream:** In all aspects of our data collection, experimentation, analysis, and modeling, we strive to strike an ethical balance between risks and benefits to users with the organization’s mission firmly in mind.
1. At the source: Protecting our users’ personal data

We understand and accept the responsibilities inherent in collecting and using people’s personal information. We further recognize the sensitivity of the data of our particular users, many of whom are interacting with the criminal justice and/or social service systems. Our users trust us with their personal information, and we commit to being worthy of that trust.

We collect a user’s personal information in order to assist their individual case as they seek and make use of government services. Secondarily, their information becomes part of the body of data that we use to identify barriers and opportunities as we work to improve user outcomes, our services, and the delivery of government programs.

We must weigh the risks to our users of retaining more detailed (and therefore more identifiable) data against the benefits that this data provides in our efforts to drive positive change.

In practice, in order to protect our users’ personal data\(^1\),\(^2\),\(^3\),\(^4\),\(^5\):

1. We work with the minimum amount of personal data necessary for a project or program.\(^6\)
2. We always seek informed consent to use personal data for a purpose beyond improving a program or service.
   a. For example, even after we remove identifying details, we obtain consent to use all quotes on our microsite.
3. We store people’s data with attention to security, guarding against unauthorized access and accidental exposure. In the case of a breach, we notify our affected users and/or government partners.\(^7\)
4. We periodically dispose of personally identifiable data that is no longer of potential utility for users seeking government services.
5. We conduct external security audits to evaluate our local and cloud data storage practices.

---

\(^1\) We recognize that privacy protection is an issue of social justice for low income and otherwise vulnerable populations. See for example Data privacy as a tax on the poor.

\(^2\) ACM Code of Ethics and Professional Conduct, section 1.6

\(^3\) Ethics and Data Science, Chapter 2

\(^4\) GetCalFresh Privacy Policy

\(^5\) GetCalFresh Security Policy

\(^6\) Data Ethics Framework gov.uk, section 3 “Use data that is proportionate to the user need”

\(^7\) ACM Code of Ethics and Professional Conduct, section 2.9
2. Downstream: Guarding against unintended consequences

Collected and used responsibly, data is one of our best resources to help our users get the services they need. We use data to inform our iterative development processes and to evaluate the impacts of our efforts. However, there are various points in the process of data collection, cleaning, analysis, and modeling where bias can enter the system. If we base our work on problematic data or methods, we risk inadvertently harming our users. Therefore, ethical data science practice requires us to attend scrupulously and with accountability to the validity of our work at each point in the process.

A. Data collection

We understand that data is not truth. Data is a form of information produced by humans for human use. As we have seen in countless examples in recent years, using data that is unexamined for bias and errors can lead to negative outcomes for certain communities, almost always groups with fewer avenues of recourse (such as those in poverty), or groups that have long faced prejudice (such as people of color).

Understanding this hazard, we approach our work with humility and curiosity, committed to building datasets with an ethic of equity.

In practice, in order to mitigate bias in data collection:

1. We examine our data for biases that, amplified by a model, might negatively affect our users. These biases could include:
   a. Historical (societal) bias in the world as it is;
   b. Representation bias emerging from limitations of sampling or collection methods;
   c. Measurement bias in which the quality or completeness of data varies across groups.

2. We consult published research, public datasets, subject matter experts, and our own diverse team.

3. We work in collaboration with qualitative researchers to build and corroborate our understanding of the data, ground-truthing our work through examination of the lived experiences of our users.

---

8 “Raw Data” is an Oxymoron
9 For example, predictive policing’s disparate racial impacts.
10 Data Ethics Framework gov.uk, section 4 “Understand the limitations of the data”
11 A Framework for Understanding Unintended Consequences of Machine Learning
12 Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Consumer Harms
13 Ethics and Data Science Chapter 4
ETHICAL DATA USE AT CODE FOR AMERICA

B. Analysis and modeling

We understand that no statistical model can capture the complexity of real people in the real world. Models are, by design and intent, simplified representations of the relationships we wish to understand. Furthermore, models only reflect what the analyst knew to include and could include. The ethical concern is that in simplifying the complexity of a particular situation, a model may cause harm by fitting some groups better than others, resulting in disparate (and unfair) outcomes.

In practice, to mitigate bias in analysis and modeling:

1. We test ideas arising from historical data using experiments, with the hope of exposing critical gaps in our understanding.
2. We test our products with a diversity of users.
3. When planning experiments, we explicitly consider the potential risks and benefits that we can anticipate, documenting these trade-offs in the write-up of the plan.

4. In situations where we can imagine the potential for harm, we take steps to minimize that harm by planning ahead how we can measure it, and by testing first with a small group with plans to remediate any adverse effects.

5. We prioritize transparency and interpretability in our research work using models. In the context of a production model that operates in real-time, if a black box method produces measurably better outcomes than an easily interpretable model, we review the results for equitable outcomes both during development and after deployment.

6. Any new application of our work is subject to the same rigorous ethical review as the original.
3. Implications for the organization

People who are drawn to Code for America’s work care deeply about helping others; we trust the intentions of all our colleagues. The greatest risk for causing harm lies in failing to take the time to thoughtfully weigh potential risks against potential benefits. Certain organizational practices can help ensure that we have the time and space we need for critical attention to ethical considerations in our data work. These include:

Making time to think\textsuperscript{18,19}

We are committed to working lean and iterating quickly. That said, we can’t compromise ethics for speed. We emphasize the “learning” component of the build-measure-learn cycle, including continually refining our understanding of potential risks and benefits to our users.

In this context of continual learning, anyone on a project team needs to be empowered to pause a process because of an ethical concern, without fear of retaliation or favoritism, even when it is inconvenient. The purpose of this metaphorical “Andon cord”\textsuperscript{20} is to create the space for an ethical discussion. Preferably, concerns will be aired, discussed, and resolved publicly in team meetings or other forums. The expectation is that the vast majority of concerns will be raised and addressed in this fashion.

Alternatively, someone might bring a sensitive issue privately or even anonymously to a member of the data team who can then raise it with the rest of the team.

The “ethical veto”

Rarely, if ever, the team may not be able to resolve an ethical question. In this situation, the question will be reviewed by the data team, in consultation with researchers, engineers, leadership, and/or others with relevant expertise. The perspectives of the data team hold particular weight, both because they have a deep level of visibility into the work that others may lack, and moreover because their professional reputations depend on the organization’s adherence to its ethical data use principles.

If, after review and discussion, two or more members of the data team believe that a project is unethical, then the team will make a recommendation to our executive team (known at Code for America as CxO) that the project stop, documenting their reasons in a brief memo. Importantly, individual data team members must feel empowered to make this “ethical veto” without fear of retaliation from management or leadership.\textsuperscript{21,22,23}

\textsuperscript{18} Ethics and Data Science Chapter 1
\textsuperscript{19} Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Consumer Harms
\textsuperscript{20} The Andon Cord
\textsuperscript{21} ACM Code of Ethics and Professional Conduct, sections 2.5, 3.4
\textsuperscript{22} Ethics and Data Science Chapter 4
\textsuperscript{23} Whistleblowing: Reaping the Benefits. Miceli, Marcia P and Near, Janet P. The Academy of Management Executive; Aug 1994

codeforamerica.org
ETHICAL DATA USE AT CODE FOR AMERICA

CxO will meet with the data team at their next scheduled “office hours” (i.e., within a week) to discuss the recommendation. If CxO agrees, the project is discontinued. If CxO disagrees, the question should be brought to one or more experts from outside the organization for review and guidance. CxO will document their final decision and reasons in writing.

Note that, should any person in the organization feel that the “ethical veto” process has failed to address a problem, the organization’s whistleblower policy applies.

**Hiring and references**[24]

When making hiring decisions for the data team, we explicitly seek out evidence of a candidate’s attention to the ethical use of data, including protecting sensitive or personal data. When providing references for our colleagues, we commit to including assessments of their commitment to ethical data practices.

Our intention here is to signal the potentially important role of professional reputation in shaping the behavior of organizations. Too many data ethics discussions assume that data professionals, because they are not empowered to make decisions about the direction of work, are not at all responsible for what their organizations do.

We believe that, like medical and scientific researchers, data professionals have a particular ethical responsibility in their organizations. A data practitioner’s professional reputation should reflect their ethical as well as their analytical work.

**Big picture**[25]

Everyone associated with Code for America must appreciate that the ethical use of data is critical to the organization’s success. Code for America’s status in the civic tech ecosystem depends on its reputation as a trustworthy example of ethical application of the tools of the digital age.

---

24 *Ethics and Data Science* Chapter 4
25 *Ethics and Data Science* Chapter 1
Further reading

- **Weapons of Math Destruction** by Cathy O'Neil
- **A hippocratic oath with teeth**
- **Automating Inequality** by Virginia Eubanks
- **A hippocratic oath for data science**
- **Ethics and Data Science** by DJ Patil, Hilary Mason, Mike Loukides
- **“Raw Data” is an Oxymoron** ed. Lisa Gitelman
- **Association for Computing Machinery code of ethics**
- **American Statistical Association ethical guidelines**
- **Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Consumer Harms**
- **Data Ethics Framework** gov.uk
- **Bias in Computer Systems**
- **A Framework for Understanding Unintended Consequences of Machine Learning**
- **Datasheets for Datasets**
- **Data Science Institute at Columbia: FATES**
- **AI Now Institute 2018 report**
- **Data & Society: Supporting Ethical Data Research**
- **Ethics is a commitment to struggle with the murk**
- **Building an ethical data strategy from the ground up**
- **Data privacy as a tax on the poor**
- **Ethics Will Shape the Customer Experience of AI**
- **It's Time for Data Ethics Conversations at Your Dinner Table**