Principles for Accountable Algorithms and a Social Impact Statement for Algorithms

Principles for Accountable Algorithms

Automated decision making algorithms are now used throughout industry and government, underpinning many processes from dynamic pricing to employment practices to criminal sentencing. Given that such algorithmically informed decisions have the potential for significant societal impact, the goal of this document is to help developers and product managers design and implement algorithmic systems in publicly accountable ways. Accountability in this context includes an obligation to report, explain, or justify algorithmic decision-making as well as mitigate any negative social impacts or potential harms.

We begin by outlining five equally important guiding principles that follow from this premise:

Algorithms and the data that drive them are designed and created by people -- There is always a human ultimately responsible for decisions made or informed by an algorithm. "The algorithm did it" is not an acceptable excuse if algorithmic systems make mistakes or have undesired consequences, including from machine-learning processes.

Responsibility

Make available externally visible avenues of redress for adverse individual or societal effects of an algorithmic decision system, and designate an internal role for the person who is responsible for the timely remedy of such issues.

Explainability

Ensure that algorithmic decisions as well as any data driving those decisions can be explained to end-users and other stakeholders in non-technical terms.

Accuracy

Identify, log, and articulate sources of error and uncertainty throughout the algorithm and its data sources so that expected and worst case implications can be understood and inform mitigation procedures.

Auditability

Enable interested third parties to probe, understand, and review the behavior of the algorithm through disclosure of information that enables monitoring, checking, or criticism, including through provision of detailed documentation, technically suitable APIs, and permissive terms of use.

Fairness

Ensure that algorithmic decisions do not create discriminatory or unjust impacts when comparing across different demographics (e.g. race, sex, etc).

We have left some of the terms above purposefully under-specified to allow these principles to be broadly applicable. Applying these principles well should include understanding them within a specific context. We also suggest that these issues be revisited and discussed throughout the design, implementation, and release phases of development. Two important principles for consideration were purposefully left off of this list as they are well-covered elsewhere: privacy (http://oecdprivacy.org/) and the impact of human experimentation (http://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/). We encourage you to incorporate those issues into your overall assessment of algorithmic accountability as well.

Social Impact Statement for Algorithms

In order to ensure their adherence to these principles and to publicly commit to associated best practices, we propose that algorithm creators develop a Social Impact Statement using the above principles as a guiding structure. This statement should be revisited and reassessed (at least) three times during the design and development process:

- design stage,
- pre-launch,
- and post-launch.

When the system is launched, the statement should be made public as a form of transparency so that the public has expectations for social impact of the system.

The Social Impact Statement should minimally answer the questions below. Included below are concrete steps that can be taken, and documented as part of the statement, to address these questions. These questions and steps make up an outline of such a social impact statement.

Responsibility

Guiding Questions

- Who is responsible if users are harmed by this product?
- What will the reporting process and process for recourse be?
- Who will have the power to decide on necessary changes to the algorithmic system during design stage, pre-launch, and post-launch?

Initial Steps to Take

- Determine and designate a person who will be responsible for the social impact of the algorithm.
- Make contact information available so that if there are issues it's clear to users how to proceed
- Develop a plan for what to do if the project has unintended consequences. This may be part of a maintenance plan and should involve post-launch monitoring plans.
- Develop a sunset plan for the system to manage algorithm or data risks after the product is no longer in active development.

Explainability

Guiding Questions

- Who are your end-users and stakeholders?
- How much of your system / algorithm can you explain to your users and stakeholders?
- How much of the data sources can you disclose?

Initial Steps to Take

- Have a plan for how decisions will be explained to users and subjects of those decisions. In some cases it may be appropriate to develop an automated explanation for each decision.
- Allow data subjects visibility into the data you store about them and access to a process in order to change it.
- If you are using a machine-learning model:
    - consider whether a directly interpretable or explainable model can be used.
    - describe the training data including how, when, and why it was collected and sampled.
    - describe how and when test data about an individual that is used to make a decision is collected or inferred.
- Disclose the sources of any data used and as much as possible about the specific attributes of the data. Explain how the data was cleaned or otherwise transformed.

Accuracy

Guiding Questions

- What sources of error do you have and how will you mitigate their effect?
- How confident are the decisions output by your algorithmic system?

- What are realistic worst case scenarios in terms of how errors might impact society, individuals, and stakeholders?
- Have you evaluated the provenance and veracity of data and considered alternative data sources?

Initial Steps to Take

- Assess the potential for errors in your system and the resulting potential for harm to users.
- Undertake a sensitivity analysis to assess how uncertainty in the output of the algorithm relates to uncertainty in the inputs.
- Develop a process by which people can correct errors in input data, training data, or in output decisions.
- Perform a validity check by randomly sampling a portion of your data (e.g., input and/or training data) and manually checking its correctness. This check should be performed early in your development process before derived information is used. Report the overall data error rate on this random sample publicly.
- Determine how to communicate the uncertainty / margin of error for each decision.

Auditability

Guiding Questions

- Can you provide for public auditing (i.e. probing, understanding, reviewing of system behavior) or is there sensitive information that would necessitate auditing by a designated 3rd party?
- How will you facilitate public or third-party auditing without opening the system to unwarranted manipulation?

Initial Steps to Take

- Document and make available an API that allows third parties to query the algorithmic system and assess its response.
- Make sure that if data is needed to properly audit your algorithm, such as in the case of a machine-learning algorithm, that sample (e.g., training) data is made available.
- Make sure your terms of service allow the research community to perform automated public audits.
- Have a plan for communication with outside parties that may be interested in auditing your algorithm, such as the research and development community.

Fairness

Guiding Questions

- Are there particular groups which may be advantaged or disadvantaged, in the context in which you are deploying, by the algorithm / system you are building?
- What is the potential damaging effect of uncertainty / errors to different groups?

Initial Steps to Take

- Talk to people who are familiar with the subtle social context in which you are deploying. For example, you should consider whether the following aspects of people's identities will have impacts on their equitable access to and results from your system:
- Race
- Sex
- Gender identity
- Ability status
- Socio-economic status
- Education level
- Religion
- Country of origin
- If you are building an automated decision-making tool, you should deploy a fairness-aware data mining algorithm. (See, e.g., the resources gathered at http://fatml.org).
- Calculate the error rates and types (e.g., false positives vs. false negatives) for different sub-populations and assess the potential differential impacts.

Authors

Nicholas Diakopoulos, University of Maryland, College Park (http://www.nickdiakopoulos.com/)

Sorelle Friedler, Haverford College (http://sorelle.friedler.net/)

Marcelo Arenas, Pontificia Universidad Catolica de Chile, CL (http://marenas.sitios.ing.uc.cl/)

Solon Barocas, Microsoft Research (http://solon.barocas.org/)

Michael Hay, Colgate University (http://cs.colgate.edu/~mhay/)

Bill Howe, University of Washington (https://faculty.washington.edu/billhowe/)

H. V. Jagadish, University of Michigan (https://web.eecs.umich.edu/~jag/)

Kris Unsworth, Drexel University (https://unsworthk.com/)

Arnaud Sahuguet, Cornell Tech (http://arnaud.sahuguet.com/)

Suresh Venkatasubramanian, University of Utah (http://www.cs.utah.edu/~suresh/web/)

Christo Wilson, Northeastern University (https://cbw.sh/)

Cong Yu, Google (https://sites.google.com/site/congyu/home)

Bendert Zevenbergen, University of Oxford (http://www.benzevenbergen.com/)