Algorithmic Equity Checklist (Guiding Questions)

Potential Harms

Impact

The effect technology will have on the community members/users
Case study/example: i.e. from AI blindspot, theyarewatching

Could the technology have negative impacts on the users such as denial of benefits or services?

- Who is responsible if users experience negative impacts from the technology?
- What will the reporting process for negative impacts be?

Who is the technology benefiting or most likely to directly benefit?

- Who are your end-users or stakeholders?

https://www.fatml.org/resources/principles-for-accountable-algorithms

Could the technology reinforce or amplify existing bias that targets minority or low-income communities? --

- How will you monitor this when implementing the technology?

Surveillance of religious communities

____

Appropriate use

The extent to which the technology is appropriate for the community and purpose
Is the data used in the algorithm compatible for the purpose and use of the technology? For instance, predictive policing technologies which forecast criminal activities are designed with data based on historical context and racial bias thus reinforcing bias against minority communities.
Is the data used validated and representative of the community, real-world situations? For instance, if the technology is a pretrial risk assessment tool or facial recognition, was the training data representative of your community/jurisdiction. How accurate is the data? How timely is the data?
Transparency (& accountability)
The extent to which the algorithms (codes, data) used are available to users.
Does the tool or system provide information about data: Who was the sample population? When was it collected? How was it analyzed? Was the data public or private? Does the data has identifiable personal information?

What types of personal information are used?
Does the tool or system provide information about human involvement: Who designed the tool? Who should be held accountable? What was the purpose of the tool?
Is there any explanation about the assumptions, models, and algorithms used for the technology? What are the limitations, uncertainty of the model? What features (variables) are used as the input for the model? What are the accuracy rates? What is the margin of error? What were the tools used for the models?
Does the tool explicitly state limited intended purposes, and are the applications limited to the listed purposes? For instance, will the Automated License Plate Readers (ALPR) be used to determine patterns and target drivers who visit gun shops, immigration clinics, health centers, protests or places of religious worship?

Security and privacy
The extent to which data is protected from security breaches
Is there information about the measures taken to protect personally identifiable data such as name, address, or face? - Health privacy organization impacted by NSA surveillance - Protecting Location Data - individuals and businesses can get your information from the gov't -
Will the data collected be used only for the purpose intended and not be shared with other agencies, government or companies? For instance, will the data collected by ALPR also be shared with banks, auto recovery companies, or insurance companies? - Government web site leaks three years of vehicle location data - Officers exposed using databases for personal reasons -- risks increase once data is shared - -
Is there information about data security, for instance how information shared amongst law enforcement agencies from the automated license plate readers will be protected?
Interpretability

The extent to which the technology can be understood by users, government agencies, officials, stakeholders and community organizations
Does the technology provide clear documentation on how to interpret the models and outputs?
Are there policies or guidelines for proper use of the technology? Targeting License Plates, Targeting People //
Will the institutions be held responsible for decisions made by algorithms used, even if it is not feasible to explain in detail how the algorithms produce their results?
Operability
The extent to which the technology can be administered by officials or users.
Have you been trained how to operate the technology correctly? - ALPR reads incorrectly
Is there a straightforward and non-technical term that describes the technology, its use, inputs and outcomes?
Accountability

[methods]

While technological advancement has led to positive changes within our society, research shows that technology negatively impacts society and our culture. Some of the sociological and legal issues we should be concerned about when implementing technologies in our communities include impact and fairness, appropriate use, transparency, security and privacy, interpretability and operability, accountability (Bavitz et al 2018, Casacurbeta 2018, Moy 2019, Friedman & Nissenbaum 1996, Angwin et al 2018, Friedman B & Nissenbaum H. 1996, Diakopolous 2016)...

Impact (fairness) -

Appropriate use -

Transparency which can be defined as the extent to which the algorithms, models and data used are available and visible to users. ProPublica's Machine Bias report (Angwin et al, 2016) highlight the impact of invisibile algorithms used to predict risk assessments within the criminal justice system. The impact of using technologies that users do not have the codes, data, and models of the technology not only accounts for unfairly predicted scores of defendants but also bring issues about interpretability and operability of the technology. Diakopolous (2016) argues to ensure transparency and accountability, technology developers should disclose human involvement (the goal, intent and purpose of the technology, who created the technology, who is responsible for the tool and should be held accountable), the data used (collection method, vetted, transformed, was the data private or public), the model (what variables were used, which tools were used to create the models, which training data was used for the models, what were the weights used and assumptions) and inferences (what is the margin or error, what is the accuracy rate),

Security and privacy -

Interpretability -

Operability -

Accountability -

[add links of cases/examples to each section: i.e. AI blindspot cards have case studies]

Work Cited:

AI Now. 2018. Algorithm accountability policy toolkit.
Bavitz, Christopher, Sam Bookman, Jonathan Eubank, Kira Hessekiel, and Vivek Krishnamurthy. 2018. Assessing the Assessments: Lessons from Early State Experiences in the Procurement and Implementation of Risk Assessment Tools. *Berkman Klein Center for Internet & Society research publication.*
Center for Government Excellence. Ethics and Algorithms Toolkit https://ethicstoolkit.ai/
Diakopoulos N. 2016. Accountability in Algorithmic Decision Making. *Communications of the ACM, Vol. 59(2).*
Accessed from: https://cacm.acm.org/magazines/2016/2/197421-accountability-in-algorithmic-decision-making/fulltext

https://www.eff.org/pages/automated-license-plate-readers-alpr

https://www.rand.org/pubs/research_reports/RR2708.html

https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf

Casacurbeta D. (2018). Bias in a Feedback Loop: Fuelling Algorithmic Injustice. Accessed from: http://lab.cccb.org/en/bias-in-a-feedback-loop-fuelling-algorithmic-injustice/

Friedman B & Nissenbaum H. (1996). Bias in Computer Systems. ACM Trans. Inf. Syst., 14(3), 330–347. https://doi.org/10.1145/230538.230561

Moy, L. (2019). How Police Technology Aggravates Racial Inequity: A Taxonomy of Problems and a Path Forward. *SSRN Electronic Journal.*

Buolamwini J, & Gebru T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of Machine Learning Research, 81, 15. New York, NY.

Angwin et al (2016). Machine Bias: There's Software Used Across the Country to Predict Future Criminals and it's Biased Against Blacks. ProPublica.

Diakopoulous et al. Principles for Accountable Algorithms and a Social Impact Statement for Algorithms. *FAT/ML* Accessed from: https://www.fatml.org/resources/principles-for-accountable-algorithms

https://www.fatml.org/resources

Additional Resources:

AI Blindspot
https://aiblindspot.media.mit.edu/

Ethical OS
https://ethicalos.org/wp-content/uploads/2018/08/Ethical-OS-Toolkit-2.pdf

ACM US Public Policy Council. Statement on Algorithmic Transparency and Accountability. 2017
https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf

----

Higher-level feedback from Aaron

Good start! I think that overall this checklist is good and these are good questions to ask. Some more work could be done to make this checklist easier to use/understand.

Understanding: To the user of the checklist it seems unclear to me what exactly the harms/impacts are that result from the tool or system not being secure, interpretable, operable, ect. Perhaps, you could put a description of the potential harms from a "no" answer at the top of each section or within each cell.

Additionally, what is the result of a yes, no, or N/A? Does it add up to a score to assess the degree of impact in a potential area to evaluate the tool as the Gov Ex toolkit does?
Are yes and no questions enough for people to understand the impacts of a tool or system? Or would it be more helpful to ask how, what, and why questions where people write in their responses? For instance, the King County Racial Equity Toolkit does this for their assessment worksheet.

Is this an internal or external assessment of the tool/system? What will people do with the assessment?

Consistency: create a consistent format for providing examples after questions within each cell

Tasks: