Algorithmic Equity Checklist

Potential Harms
Impact
The effect technology will have on the community members/users **Example Case:** *The New York City Police Department uses various high and low tech methods to surveil and gather intelligence on Muslim neighborhoods and mosques. Having been questioned by the NYPD already many Muslim congregants are fearful and wary of attending religious services or engaging with their own religious community. Surveillance of the Muslim community became "so widespread that it interefere[d] with it's members' legitimate political and religious activities", ultimately denying them services which every American has a right to enjoy.* http://theyarewatching.org/issues/biased-targeting
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? • What are the negative impacts?
Will the technology have positive impacts on the users? • What are the positive impacts? •
Who or what will be benefiting from the technology or most likely to directly benefit?

- Who are your end-users or stakeholders?
- What are the benefits?

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

- How will you monitor this when implementing the technology?
- What strategies will be implemented to mitigate this issue?

Appropriate use

The extent to which the technology and data is appropriate for the community and purpose

Example Case: *A paper by several scholars found that gender classification software from IBM, Microsoft, and Face++ misidentified the gender of women and darker-skinned people at higher rates than it did men and lighter-skinned people. This confirms that there is a real risk that women and people of color will be misidentified by face recognition technology. Yet, such technology is used by police across the United States and is often unregulated.*
http://gendershades.org/overview.html
https://www.perpetuallineup.org/

Is the data used in the algorithm compatible for the purpose and use of the technology?

- Will there be transparency with the public on what data the algorithm employs?
- For what purpose was the data input originally collected?

Is the data used validated and representative of the community, real-world situations?

- How accurate does the data represent real-world situations?
- How timely is the data?

Transparency (& accountability)

The extent to which the algorithms (codes, data) used are available to community members/ users.

Example Case: *Residents of Idaho with developmental disabilities saw their Medicaid assistance slashed by thousands of dollars. The formula, data, and code that the Medicaid program was using to determine how much funding individuals were due was not available to the public or interested parties. Thus, there was no explanation for why an individual's Medicaid spending allowance was cut. A judge ruled that the state Medicaid program had to release this information, and upon further inspection a team of experts found serious flaws with the statistical methods and data used. These flaws ultimately resulted in individuals' medicaid allowances being incorrectly and baselessly cut.*
https://www.aclu.org/blog/privacy-technology/pitfalls-artificial-intelligence-decisionmaking-highlighted-idaho-aclu-case

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?

What were the variables used?

Does the data have 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?

Who or what made the decisions?

Can the decisions be reviewed or audited?

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?

- Will vendors of technologies be held accountable?

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?

How would you mitigate the effects of error?

What were the tools used for the models?

Was the algorithm tested before it is put into use?

Will the algorithm be modified over time?

Is the algorithm from a third-party developer?

Data Security and Privacy

The extent to which data is protected from security breaches

Example Case #1: *A health privacy organization that receives and pursues anonymous tips from whistleblowers about health privacy violations in industry and government was forced to change their privacy policy when they discovered the National Security Agency (NSA) was tracking calls. Instead of promising callers that the information they relay is confidential, the organization had to update their policy to inform callers that neither calls nor emails were secure. As a result of their inability to ensure whistleblowers' privacy and security, the organization saw a drop in reports of health privacy violations.*
http://theyarewatching.org/issues/affects-who-you-want-be-seen

Is there information about the measures taken to protect personally identifiable data such as name, address, or face?

- Does the system protect all its users?
- Where will this information be made available to users?
- How is it the data stored?
- Is it encrypted?
- Will the data be disposed after some time?

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?]

- Will the users consent to all data collected on their behalf?

Is there information about data security, for instance how information shared amongst law enforcement agencies from the automated license plate readers will be protected?

- Will users be informed before data is shared between law enforcement agencies?
- Will all law enforcement agencies be held to the same standards to ensure data privacy?

Interpretability

The extent to which the technology can be understood by users, government agencies, officials, stakeholders and community organizations

Example Case: *Amazon recommends a threshold of at least 95% or higher for clients who use their face recognition software (Rekognition) for "law enforcement activities". However, at least one "law enforcement agency Amazon has acknowledged as a client says it...does not use Rekognition in the way Amazon claims it recommends". More specifically, at the Washington County Sheriff's Office in Oregon "the software is deployed in cases ranging from theft to homicide" and the office "do[es] not set nor...utilize a confidence threshold". While Amazon provides documentation and "support on the software end, no direct training was given to the investigators who continue to use the suite".* https://gizmodo.com/defense-of-amazons-face-recognition-tool-undermined-by-1832238149

Does the technology provide clear documentation on how to interpret the models and outputs?

- Where can users access this documentation?
- Will it be available in multiple languages?
- Is it easy to explain the algorithm and how it works to a layperson?

Are there policies or guidelines for proper use of the technology?
How much information can be disclosed?

Operability

The extent to which the technology can be administered by officials or users.

Example Case: *In 2009 a woman drove past a police car in San Francisco with an automatic license plate reader (ALPR). Unfortunately, the ALPR mis-scanned her license plate and identified her car as a stolen vehicle, though it was not. The woman was pulled over by several police officers who, "with guns drawn, handcuffed her, and conducted a field search of the car that exposed their mistake". The officers were not trained to "verify both the plate number and the model and color of the car, either of which would have clearly revealed the mistake".*

http://theyarewatching.org/issues/potential-mistakes

Have you been trained how to operate the technology correctly?

- Who will write the training curriculum?

Is there a straightforward and non-technical term that describes the technology, its use, inputs and outcomes?

- Will this term undergo user testing to ensure its comprehensibility?

Methodology:

The above questions designed for community leaders and members were drawn from the literature on Fairness, Accountability and Transparency in Machine Learning, AI Ethics and Governance. While algorithmic systems are efficient and effective in providing services to our communities, the systems do raise new ethical questions and concerns about equity and fairness within social institutions (Osoba et al 2019). Biased algorithms designed with biased data and assumptions may lead to unintended consequences. Hence, some of the sociological and legal issues we should be concerned about when implementing algorithmic systems in our communities include impact and fairness, appropriate use, transparency and accountability, security and privacy, interpretability and operability (Bavitz et al 2018, Casacurbeta 2018, Moy 2019, Friedman & Nissenbaum 1996, Angwin et al 2018, Friedman B & Nissenbaum H. 1996, Diakopoulos 2016, USACM 2017, Ekstand et al 2018, GovEx).

Impact (fairness) - the Center for Government Excellence (GovEX) Ethics and Algorithms Toolkit assesses algorithm bias, their toolkit provides risk management guidance to government leaders. Risks to be addressed include the impact of the system, appropriateness and bias. Questions to ask about the societal and technological impacts of the system include who or what will be impacted by the technology, the types of impact (access to goods and benefits), financial, privacy and the direction of impact, whether it is positive or negative. In terms of appropriate use, questions to ask are data compatibility, the purpose of the data when it was originally collected or obtained.

Transparency & accountability, 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. In the case of the risk assessment tool, the effect of using technologies that users do not have the codes, data, and models used to design the technology not only accounts for unfairly predicted scores of defendants as well as making it harder for a defendant to challenge the use of algorithms to predict risk scores (Osoba et al

2019). 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), and whether an algorithm is being used in the technology (Diakopoulos 2016, GovEx.)

Security and privacy - according to the scholars at the Berkman Klein Center, there should be a system for protecting data from breaches (Greene, 2018). Both privacy and fairness seek to protect people from the effects of social, legal and technical systems. The concepts of fairness and privacy intersect and should be considered as one issue instead of separate entities (Ekstrand et al. 2018). Major question we can ask about the technology is whether the system provides privacy protections to all its users or only protects some groups. Other aspects to consider are the attack capabilities of the systems on different users/groups.

Interpretability & Operability - when considering adopting technologies for an organization or service, interpreting and operating these technological systems might be challenging to non-tech experts and users. Bavitz and colleagues (2018) argue that transparency not only should be considered for the algorithms and companies that develop them but also to the government and organizations that implement these technologies. States and organizations should establish guidelines to ensure users can administer the technologies correctly. Trainings on interpreting results and documentations should be provided for the organizations using the technologies. In addition, algorithmic decisions and data should be explained to end-users in non-technical terms (Diakopoulos et al. ??).

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.*

Eckstrand M, Joshaghani R, Hoda Mehrpouyan. (2018). Privacy for All: Ensuring Fair and Equitable Privacy Protections. Proceedings of Machine Learning Research 81:1–13, 2018

Osoba A. Osonde, Benjamin Boudreaux, Jessica Saunders, J. Luke Irwin, Pam A. Mueller, Samantha Cherney. (2019). Algorithmic Equity: a framework for social applications. RAND Corporation. https://www.rand.org/pubs/research_reports/RR2708.html

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

ACM US Public Policy Council. (2017). Statement on Algorithmic Transparency and Accountability. Accessed from: 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.

Diakopoulos 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

Greene K. G. (2018). Buying your first AI or "Never Trust a Used Algorithm Salesman". Accessed from: https://medium.com/berkman-klein-center/buying-your-first-ai-136cd2e6dd2

Additional Resources:

For more information about ethics and algorithmic biases, please look at the following 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

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

Fairness, Accountability and Transparency in Machine Learning
https://www.fatml.org/resources

AI Ethics Guidelines Global Inventory. Algorithm Watch
https://algorithmwatch.org/en/project/ai-ethics-guidelines-global-inventory/

Ethics and Algorithm Toolkit
http://ethicstoolkit.ai/

Links to other case studies/examples:
https://www.newscientist.com/article/mg23631464-300-biased-policing-is-made-worse-by-errors-in-pre-crime-algorithms/

https://www.nyulawreview.org/wp-content/uploads/2019/04/NYULawReview-94-Richardson-Schultz-Crawford.pdf

http://lab.cccb.org/en/bias-in-a-feedback-loop-fuelling-algorithmic-injustice/