Algorithmic Equity Checklist

Potential Harms
Impact
The effect technology will have on community members **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 its 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 affected community members such as adverse impacts on civil rights and fair allocation of policing and justice? ● If so, what solutions will be implemented to resolve the negative impacts? ● Are there other technologies that can be implemented without having negative impacts on the community members? ● Who is responsible if community members experience negative impacts from the technology? ● What will the reporting process for negative impacts be? Could the technology have negative impacts on affected community members such as denial of service, a credit, bank loan? ● If so, what solutions will be implemented to avoid this? ● Are there other technologies that can be implemented without having negative impacts on the community members?

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

Could the technology have negative impacts on affected community members such as privacy violations and profiling?

- If so, what solutions will be implemented to avoid this?
- Are there other technologies that can be implemented without having negative impacts on the community members?
- Who is responsible if community members experience negative impacts from the technology?
- What will the reporting process for negative impacts be?

Are there any other negative impacts?

What biases can this technology reinforce or amplify?

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

Will the technology have positive impacts on the affected community members such as maintaining public safety and crime prevention?

- If so, will the technology have positive impacts to all community members regardless of their age, gender, socioeconomic status or race?
- Do the positive impacts outweigh the anticipated harms?
- Are there other positive impacts?

Who or what will be benefiting from the technology or most likely to benefit?

- Who are the users likely to benefit?
- Who are the community members likely to benefit?
- What are the benefits?

Appropriate use
The extent to which the technology and data is appropriate for the community and stated 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/*
What is the primary intended use of the technology? • Is the technology compatible for the intended purpose? • Who are the primary intended users? • Can the technology be used for other purposes? • Can the technology be used for all contexts? • If so, what are the other purposes that the technology can be used for? • Are there locations that using the technology can lead to negative impacts such as privacy violations • If so, what are the risks or effects?
Is the data used validated and representative of the community and real-world situations? • If so, how accurate does the data represent real-world situations? • What is the timeframe, or how timely is the data used to train the algorithm? • For what purpose was the training data originally collected? • Is the data sampled, if so how was its representativeness to the larger community verified?

Transparency & accountability

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

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 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 documentation about its design and functions, such as:

- When was the training data collected?
- How was the data collected?
- Who was included in the sample population?
- If sampled, what was the sampling strategy?
- How was it analyzed?
- Was the data public or private?
- What were the variables used?
- How were the weights of the variables determined?
- Was the data tested for bias?
- Does the data have identifiable personal information?
- Does the data identify any subpopulation, such as by age, gender or race?
- What types of personal information are used?
- Will there be a system maintenance?
- If so, how will the system maintenance and monitoring be conducted?

How is the tool or system accountable and answerable to individuals and communities affected by its use?

- Were members of affected communities involved in the design of the tool, feedback or consulted about its features and predicted effects?
- Who designed the tool?
- Who should be held accountable?
- What was the purpose of the tool?
- Why does the system exist?
- Who or what made the decisions?
- Can the decisions be reviewed or audited?
- How will institutions be held responsible for decisions made by the algorithm?
- Can the users of the algorithm explain how it produced its results?
- Will vendors of technologies be held accountable?

Is there any documentation 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?
- Are the accuracy rates consistent across sociodemographic groups or do they vary?
- What is the margin of error?
- How would you mitigate the effects of error?
- What cost was given to false positives or false negatives?
- What priority was given to precision and recall?
- What were the tools used for the models?
- Was the algorithm tested before it is put into use?

- How was the system evaluated for effectiveness?
- Will the algorithm be modified over time?
- Is the algorithm from a third-party developer?

Are there policies or guidelines for proper use of the technology?

- How much information can be disclosed?
- Are there any legal or regulatory restrictions on disclosing these policies, and if so, what are they?"

Data Security and Privacy

The extent to which data is protected from security breaches & privacy concerns

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

Are there privacy implications?

- If so, what steps are taken to mitigate privacy violations?

Is there information about data privacy, such as:

- Measures taken to protect personally identifiable data such as name, address, or face?
- How accessible the data is to interested parties such as users, data brokers, researchers, hackers?

- How does the system protect all its data subjects[1]?
- How will this information be made available to data subjects?
- Was the data legally collected?
- Were the data subjects informed and consented to data collection?
- Will the data collected be used only for the stated purpose?
- Will data subjects be informed before data is shared between agencies or institutions?

Is there information about data security, such as:

- How is the data stored?
- Is it encrypted?
- How will the data be disposed?
- Will the data be disposed after some time and when?
- How was the length of time for disposal decided?
- How will data subjects be informed if there is a security breach?

Interpretability

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

Example Case: *Amazon recommends a confidence 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".*

[1] Data subject is any person whose personal data is collected, held or processed. https://eugdprcompliant.com/what-is-data-subject/

https://gizmodo.com/defense-of-amazons-face-recognition-tool-undermined-by-1832238149

Does the technology provide clear documentation on how to interpret the systems models and expected outputs (prediction, results)?

- If so, where can users access this documentation?
- Will the community members have access to this documentation as well?
- Will it be available in multiple languages?
- How easy is it to interpret and explain the algorithm predictions, and how it works to a layperson?
- Is there information about the confidence threshold, accuracy rates?

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 users been trained how to operate the technology correctly?

- If so, what kind of training was offered?
- Who will or has written the training curriculum?
- Is the training curriculum available to the public?

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?
Do the users (government agencies/law enforcement) understand the technology's possible modes of failure such as when facial recognition software result in misclassification of women and darker skinned faces? ● If so, what are the modes of failure for the technology? ● What steps will be taken to remediate the anticipated modes of failure such as false positives and negatives?

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 - 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., Partnership on AI)

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). A 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 concerns to consider include how the data will be stored and accessibility measures implemented to protect the data.

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/

Partnership on AI
https://www.partnershiponai.org/about-ml-get-involved/#read

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/

https://privacyinternational.org/feature/2216/who-supplies-data-analysis-and-tech-infrastructure-us-immigration-authorities