Auditing Government AI: How to assess ethical vulnerability in machine learning

Alayna A. Kennedy
IBM US Federal
Washington, DC 20005
alayna.kennedy@ibm.com

Daphne Coates
IBM UK
London, UK
daphne.coates@ibm.com

Katelyn Lindquist
IBM US Federal
Washington, DC 20005
katelyn.lauren.lindquist@ibm.com

Abstract

Governments are increasingly using algorithmic systems to make important decisions that impact large populations. Governments struggle to thoroughly audit, monitor, regulate, and evaluate these systems, leaving them vulnerable to ethical infractions such as biased outcomes, disparate impacts on different populations, and solutions that are unexplainable and unaccountable to citizens. This paper outlines a tool that quickly assesses the vulnerability of a machine learning project to ethical infractions by translating data scientist's technical expertise into a high-level risk score. This score allows governments to quickly identify high risk projects across their portfolio. As a result, they can appropriately allocate resources, continuous monitoring, governance and audit cycles to regulate and mitigate ethical concerns.

1 Introduction

Machine learning (ML) systems are increasingly being used to make decisions within large scale government projects, including the deployment of humanitarian resources [23], who is granted bail [8], which citizens are subjected to increased police presence [15], whether or not reports of abuse are investigated [9], and who receives government funded welfare [3]. ML systems deployed in the private sector have already shown to contain gender bias [16] and racial bias [**?**]. Using similar ML algorithms to make large-scale decisions in government organizations [28] has the power to amplify societal inequities and propagate bias [13].

Rising concerns regarding the societal implications of biased ML algorithms has led many researchers to develop tools to ensure that large-scale algorithms perform ethically [1, 22].. However, research efforts in the field of ML Ethics have overwhelmingly focused on mitigating bias and unfair ML in the technical stages of development, such as detecting historically biased data and removing disparate impact from a model's output [22, 24, 19]. Much less work has been done on developing tools to audit and regulate ML tools from a high-level organizational perspective [9]. In the case of large-scale government models, statistical fairness is insufficient, since the environment in which the model is operating is constantly changing, and auditors need to periodically reassess model performance and outcomes with thorough, multidisciplinary auditing teams [11] to avoid unethical outcomes seeping into the models over time [17].

Navigating the Broader Impacts of AI Research Workshop at the 34th Conference on Neural Information Processing Systems (NeurIPS 2020).

Furthermore, the regulatory framing of AI Ethics challenges has overwhelmingly focused on high-level frameworks [29, 14]. Most global technology companies releasing high level ethical frameworks, such as Google's AI Principles and IBM's Everyday Ethics guidelines [18], and many governments have released similar high-level frameworks for developing and deploying ML [20]. However, both the commercial and governmental frameworks are limited. Commerical frameworks fail to provide specific definitions for algorithmic behavior within an operational context [18, 19] or actionable mechanisms for implementation, and government frameworks focus only on the legal definitions of ML and on defining the role of ML in government [25]. They too fail to provide specific frameworks to audit ML projects within government agencies [21, 27].

Ensuring that ML algorithms behave ethically requires regulation, measurement, and consistent auditing [17]. It requires both the technical capability and the governance framework. However, governments face significant challenges in regulating ML systems since, agencies envelop increasingly large ML project portfolios, no central agency exists to oversee or audit ML projects, and few international industry standards [29]. Despite lack of central oversight and regulation, government agencies continue to increase their investment in ML technology [30]. However, with no central oversight of ML projects, and few existing auditing frameworks to draw from, government agencies face significant challenges in regulating their own projects[3, 4]. Many examples of government misuse of AI systems have recently been in the public eye, including the UK's automated allocation of benefits funding, the UK's use of AI to determine student's test scores, and the U.S.'s use of facial recognition technology in policing [3, 6, 1].

As governments around the world scale up their investments in AI technology [26], they will also need to scale up their capability to assess, audit, and review those technologies for ethical concerns to avoid amplifying inequality. Large scale government enterprises require a systemized method to look across their portfolio of projects and quickly assess which are more vulnerable to becoming unethical. In this paper, we propose a tool that will address the lack of auditable ML systems by providing a quantitative vulnerability score for ethical concerns. The vulnerability score measures the risk present in an ML project. We define risk as the inherent likelihood that an ML project will negatively impact a population, either through statistical bias, unfair assumptions or impact, unintended outcomes, or lack of human oversight; and we define vulnerability as the number of areas in the algorithmic development process where unintended bias or negative outcomes can enter an algorithm.

2 Motivation

2.1 Problem Statement

Governmental use of ML often involves large scale operations which may enact significant negative consequence on the population [6, 28]. However, due to significant challenges faced in regulating, monitoring and auditing ML projects, government agencies often fail to audit their riskiest projects until the later stages of deployment [3] and lack a clear line of sight across their portfolio of ML projects.

In order to audit ML projects, government agencies need tools to quickly assess the project's vulnerability to unintended bias or negative outcomes. With an agile tool to understand the risk profile, they can then ensure critical monitoring and strict auditing structures are in place to lower the risk score where possible. In addition, because the stakeholders charged with overseeing, funding and managing ML projects often do not have deep technical knowledge, they need a way to easily understand the technical risks of ML projects.

2.2 Proposed Solution

This paper develops a questionnaire for data scientists and development teams in the government industry that provides a cursory vulnerability score for a project.

The questionnaire covers several areas deemed to have significant impact on determining risk evaluation such as: data collection, human interaction and the algorithms intended domain and use. The questionnaire assesses the risk present in each of those areas by asking data scientists a series of yes or no questions about their algorithm and the context in which it is used. Each question receives a score of 1 or 0: 1 if the answer indicates a high risk, and 0 if the answer indicates lower risk. After

all questions are answered, the framework produces a vulnerability score as a quantitative output, benchmarked into thirds for low, medium and high-risk vulnerability.

The questionnaire is directed at data scientists as inputters but produces an easy-to-understand vulnerability score that can be used by functional stakeholders with little technical knowledge. This tool can be utilized by governmental bodies to quickly look across their portfolio of projects and see which are vulnerable to ethical infractions. Then, appropriate resources, including social scientists, domain experts, and AI ethics boards, can be allocated to audit the highest vulnerability algorithms and enforce continual monitoring practices. Once high vulnerability projects have been appropriately reviewed, monitoring practices can be put in place to mitigate further risk and bias inception.

3 Methodology

3.1 User Identification

First, target inputters were identified as data scientists working in the government industry. These users they have both the technical ML knowledge and an understanding on the intended use and deployment of these tools, therefore arguably well positioned to accurately answer questions on algorithms development and context. However, as mentioned prior, the end users of the questionnaire outputs are functional stakeholders, commonly responsible for resourcing, budgets, management and allocation of effort.

3.2 Literature Review & Inputs

The second step was to identify the most common vulnerability areas for ML projects, completing an extensive literature review of ML ethics research. Research unveiled that whilst many best practices depend on the context in which the algorithm is used, there are some common high-level principles emerging in the field. For example, an algorithm that uses racial data is more vulnerable to racial bias than an algorithm with anonymized data that does not include race or other proxies to race [7].

Using this research, six "vulnerability areas" were defined: aspects of the ML creation and deployment cycle where unintended bias or negative outcomes could enter the tool [22, 10]. The vulnerability areas identified are:

- **Data:** The type and content of the data can have impacts on the risks for discriminatory bias.
- **Domain:** Risk varies based first on what the algorithm is intended to do. There are specific topics that have been found to be ethically challenging when it comes to ML projects.
- **Robustness:** The scope, scale, and security of the tool determine whether its baseline is safe and secure.
- **Bias/Fairness:** The potential of inherent bias against populations in the tool, or of unfair outputs.
- **Accountability:** The more a human is involved in the decision-making process, the less risky (at this moment in history).
- **Interpretability:** There are best practices that add explainability. This section is trying to score the innate explainability of your model.

Data scientists answer questions about their project (listed in Figure 1) pertaining to the different vulnerability areas.

3.3 Pilot Testing

After identifying these six vulnerability areas, specific questions were crafted to assess the risk present in each vulnerability area. The focus was to create simple, closed-ended, 'yes or no' questions that a data scientist could reasonably answer without outside research or manager input in a short space of time. Closed-ended questions increase response rate and decrease cognitive load while taking surveys [12], as well as being much easier to analyze and visualize [12, 2] since this questionnaire is intended to be a quick method of assessing ML risk, closed-ended questions help prioritize agility.

However, it must be noted that closed-ended items tend to be more difficult to curate since they must include an appropriate set of response options [12]. Therefore, the questionnaire was pilot tested

with a select group of data scientists within the government industry to ensure that the questionnaire provided enough options to identify risk, whilst remaining easy to complete. A small sample of approximately 20 users volunteered to take the questionnaire and provide feedback.

3.4 Weighting

To ultimately generate a vulnerability score, the questionnaire assigns a score of 1 or 0 to an answer, depending on its risk level. Although each question receives the same numerical risk score (maximum of 1), most of the question focus on the "Domain" vulnerability area. ML Ethics literature emphasizes that the intended use of an algorithm is an important factor in determining its ethical impact [5]. Additionally, feedback from pilot testing indicated that the breadth of use cases for ML in the government industry required a higher number of Domain questions.

This binary method of scoring was chosen to make the questionnaire simpler, so that it can be completed and utilized quickly, while still recording important information about ML projects. For example, in the "Robustness" section, a question asks about the size of the dataset and gives 6 possible options for answers. Currently, the questionnaire assigns a risk score of 1 to any dataset smaller than 10,000, since smaller sample sizes have higher risks of bias [22]. This is a draft scoring system, but the binary weighting of the questions can be changed based on each organization's context. For example, a company may decide to score any ML project using weapons systems as high vulnerability. The binary scoring method has the flexibility to change weighting scores, while remaining simple and agile.

3.5 Outputs

After all questions are answered, the framework produces a vulnerability score as a quantitative output. The numerical scores from each question are summed and categorized as high, medium, or low risk. The risk boundaries are split into equal thirds of the maximum score. Since the maximum score possible on the questionnaire is 33, a score of <11 is low risk, a score <22 and >11 is medium risk, and a score >22 is high risk. Like the binary scoring system, the final score boundaries are simple to calculate and can be adjusted depending on organizational values. These risk scores will be visible across an entire portfolio, providing a quick visual of what projects are higher risk. Visualizations of these risk scores can be found in the Appendix.

Broader Impact

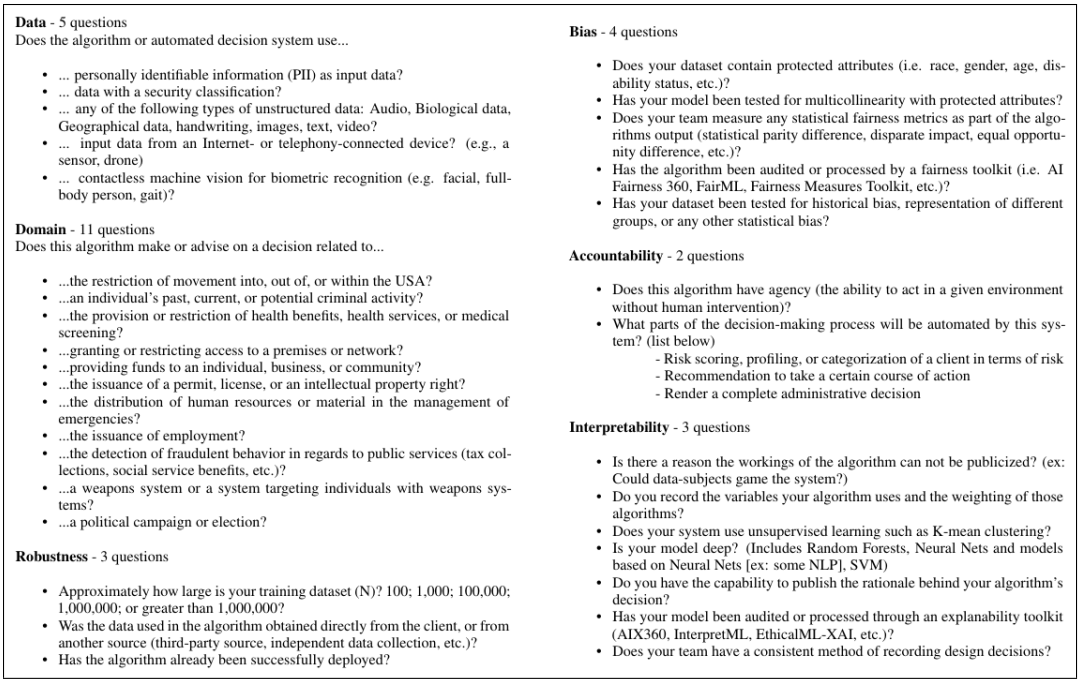
Governments are using ML in ways that could negatively impact large populations if they are not appropriately flagged, monitored and audited, since governments currently face substantial challenges in regulating their own ML projects [25, 28]. This paper proposes an agile tool to quickly assess the vulnerability of ML projects to ethical concerns like biased data or unintended harmful outcomes. The tool prioritizes agility and ease of use and can be modified to different organizations ethical ML approaches. Its agile nature means it easily be repeated over time to track a projects development. Using such a framework offers organizations and government agencies awareness of the scope of their decision-making tools.

This work begins to build a concrete, actionable way for large organizations to create and sustain ethical AI systems, moving beyond high-level principles to a framework that can actually be used by data scientists and managers creating algorithms. This framework represents an important first step in creating a robust ML ethics ecosystem within government agencies or other large enterprises. However, prioritizing simplicity in this tool leaves out necessary contextual information and open-ended assessments of ML project's ethical impact. Therefore, more work must be done on regulating and auditing ML projects after this initial assessment of their inherent risk.

Vulnerability scores could be fed into an internal ML ethics ecosystem consisting of internal organizational auditing teams and explainable algorithm scorecards. Furthermore, this questionnaire focuses on the government industry domain; additional domain questions could be generated by subject matter experts and used for different industries. As ML projects are increasingly used by governments to make large-scale decisions, the development of further ethical auditing tools will be necessary to prevent unintended negative impacts on the wider population.

In the next wave of AI Ethics development, we need to pry our focus away from high-level principles and bias-only concerns, and develop the mundane, practical tools to allow organizations to audit AI. Concrete tools, grounded in business concepts and language, will allow for organizations and government agencies to move beyond "AI Ethics washing" and create real change in the way they create and deploy their algorithms.

4 Figures



Data - 5 questions
Does the algorithm or automated decision system use...

- ... personally identifiable information (PII) as input data?
- ... data with a security classification?
- ... any of the following types of unstructured data: Audio, Biological data, Geographical data, handwriting, images, text, video?
- ... input data from an Internet- or telephony-connected device? (e.g., a sensor, drone)
- ... contactless machine vision for biometric recognition (e.g. facial, full-body person, gait)?

Domain - 11 questions
Does this algorithm make or advise on a decision related to...

- ...the restriction of movement into, out of, or within the USA?
- ...an individual's past, current, or potential criminal activity?
- ...the provision or restriction of health benefits, health services, or medical screening?
- ...granting or restricting access to a premises or network?
- ...providing funds to an individual, business, or community?
- ...the issuance of a permit, license, or an intellectual property right?
- ...the distribution of human resources or material in the management of emergencies?
- ...the issuance of employment?
- ...the detection of fraudulent behavior in regards to public services (tax collections, social service benefits, etc.)?
- ...a weapons system or a system targeting individuals with weapons systems?
- ...a political campaign or election?

Robustness - 3 questions

- Approximately how large is your training dataset (N)? 100; 1,000; 100,000; 1,000,000; or greater than 1,000,000?
- Was the data used in the algorithm obtained directly from the client, or from another source (third-party source, independent data collection, etc.)?
- Has the algorithm already been successfully deployed?

Bias - 4 questions

- Does your dataset contain protected attributes (i.e. race, gender, age, disability status, etc.)?
- Has your model been tested for multicollinearity with protected attributes?
- Does your team measure any statistical fairness metrics as part of the algorithms output (statistical parity difference, disparate impact, equal opportunity difference, etc.)?
- Has the algorithm been audited or processed by a fairness toolkit (i.e. AI Fairness 360, FairML, Fairness Measures Toolkit, etc.)?
- Has your dataset been tested for historical bias, representation of different groups, or any other statistical bias?

Accountability - 2 questions

- Does this algorithm have agency (the ability to act in a given environment without human intervention)?
- What parts of the decision-making process will be automated by this system? (list below)
  - Risk scoring, profiling, or categorization of a client in terms of risk
  - Recommendation to take a certain course of action
  - Render a complete administrative decision

Interpretability - 3 questions

- Is there a reason the workings of the algorithm can not be publicized? (ex: Could data-subjects game the system?)
- Do you record the variables your algorithm uses and the weighting of those algorithms?
- Does your system use unsupervised learning such as K-mean clustering?
- Is your model deep? (Includes Random Forests, Neural Nets and models based on Neural Nets [ex: some NLP], SVM)
- Do you have the capability to publish the rationale behind your algorithm's decision?
- Has your model been audited or processed through an explanability toolkit (AIX360, InterpretML, EthicalML-XAI, etc.)?
- Does your team have a consistent method of recording design decisions?

Figure 1: The caption text

References

[1] Muhammad Ali, Piotr Sapiezynski, Mirana Bogen, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. Discrimination through optimization. *Proceedings of the ACM on Human-Computer Interaction*, page 1–30, 2019.

[2] Vijay Arya, Rachel K. E. Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Q. Vera Liao, Ronny Luss, Aleksandra Mojsilović, Sami Mourad, Pablo Pedemonte, Ramya Raghavendra, John Richards, Prasanna Sattigeri, Karthikeyan Shanmugam, Moninder Singh, Kush R. Varshney, Dennis Wei, and Yunfeng Zhang. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques.

[3] Dan Bloom. Rise of dwp welfare robots - ai helps decide if universal credit claims are true, 2019.

[4] Robert Booth. Benefits system automation could plunge claimants deeper into poverty, 2019.

[5] Nick Bostrom. The ethics of artificial intelligence.

[6] Joy Buolamwini. Actionable auditing: Investigating the impact of publicly naming biased performace results of commercial ai products. *Conference on Artifical Intelligence, Ethics, and Society*, 2019.

[7] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classificaiton. *Proceedings of Machine Learning Research*, 81:1–15, 2018.

[8] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *FATML 2016 Conference*, pages 153–163, 2017.

[9] Cathy Cobey. Forbes insights: Ai regulation: It's time for training wheels, 2019.

[10] Henriette Cramer, Jean Garcia-Gathright, Aaron Springer, and Sravana Reddy. Assessing and addressing algorithmic bias in practice. *ACM Digitial Library: Interactions*, 25:58–63, 2018.

[11] Chris DeBrusk. The riks of machine learning bias (and how to prevent it). *Risk Journal - Rethinking Tactics*, 8, 2018.

[12] Susan Farrell. Open-ended vs. closed-ended questions in user research, 2016.

[13] Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Lütge, Robert Madelin, Ugo Pagallo, Francesca Rossi, Burkhard Schafer, Peggy Valcke, and Effy Vayena. Ai4people—an ethical framework for a good ai society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28, 11 2018.

[14] Benjamin Freed. To be fair, new york city assembles algorithm task force. Online; Accessed 10-June-2020.

[15] Clare Garvie, Katie Evans, and Alvaro Bedoya. The perpetual line-up. Available at `https://www.perpetuallineup.org/`, 2019. Online; accessed 10-May-2020.

[16] Rachel Goodman. Why amazon's automated hiring tool discriminated against women. Online; Accessed 01-June-2020.

[17] James Guszcza, Iyad Rahwan, Will Bible, Manuel Cebrian, and Vic Kaytal. Why we need to audit algorithms. *Harvard Business Review*, 2018.

[18] Thilo Hagendorff. The ethics of ai ethics: An evaluation of guidelines. *Minds and Machines*, 30:99–120, 2020.

[19] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Dume III, Miro Dudik, and Hanna Wallach. Improving fairness in machine learning systems: What do industry practitioners need? *ACM CHI Conference on Human Factors in Computing Systems*, 2019.

[20] Jeremy Kahn. In a.i., what would jesus do? Available at `https://fortune.com/2020/02/28/ai-ethics-vatican-microsoft-ibm/`. Online; Accessed 11-May-2020.

[21] Mark MacCarthy. How to address new privacy issues raised by artificial inteligence and machine learning. Technical report, Brookings Institute, 2019.

[22] Trisha Mahoney, Kush Varshney, and Michael Hind. *AI Fairness: How to Measure and Reduce Unwanted Bias in Machine Learning*. O'Reilly Media Inc., 1005 Gravenstein Highway North, Sebastopol CA, 1 edition, 2020.

[23] Billy Mitchell. Pentagon's jaic needs industry help for humanitarian assistance, disaster relief, 2019. Online; Accessed 15-May-2020.

[24] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. proceedings of the conference on fairness, accountability, and transparency. *FAT* '19: Conference on Fairness, Accountability, and Transparency*, pages 1–10, 2019.

[25] Joel Nantais. Federal government regulation of ai. Online; Accessed 31-May-2020.

[26] Andrew Nicklin and Miriam McKinney. Bloomber cities: The promise and peril of algorithms in local government, 2018.

[27] Ellen Perlman, Daniel Chenok, Jaimie Winters, and Jolito Rivera. Assessing the impact of artificial intelligence on the work of government. Technical report, The Partnership for Public Service, Washington, DC, 2019.

[28] Dillon Reisman, Meredith Whittaker, and Kate Crawford. Algorithms are making government decisions. the public needs to have a say. *AI Now Institute*, 2018.

[29] Jonathan Vanian. White house increase in ai spending. Available at `https://fortune.com/2020/02/11/white-house-a-i-funding/`, 2019. Online; Accessed 21-May-2020.

[30] James Vincent. White house encourages hands-off approach to ai regulation. Online; Accessed 01-June-2020.