
Imran Ahmed  Giles L. Colclough  Daniel First  QuantumBlack contributors  
QuantumBlack, 1 Pall Mall E, London, UK

Abstract

Developing and deploying machine learning models without due care can lead to unfair—or unlawful—decision making. As machine learning increasingly integrates into business decision processes with wide-ranging consequences, from hiring through to law enforcement, there is a need for models to be transparent, unbiased, and robust over time. There are as yet no well-adopted standard approaches to ensure that models meet these requirements. We introduce a risk management protocol and webapp platform for practitioners that highlight major risks around fairness, bias, and explainability at each stage of development. Because risks are embedded in this protocol, practitioners can understand risks and follow mitigation advice associated with the tasks they are currently completing. This promotes the mitigation of risk while models are in development, instead of ex post by checklist audits. In this workshop, we invite discussion on how to make the protocol and platform open-access, community-sourced, and an industry-standard approach to building models that are fair, accountable, and transparent. We also seek ideas on how to develop technical tooling, such as data and code linters, to automate some of the risk mitigation tasks and activities.

Topics: ‘Social welfare and justice’ and ‘Fairness and transparency in applied machine learning.’

1. Introduction

It is well recognized that bias at any point in the machine learning (ML) model lifecycle—whether around problem definition, data collection, model development, and model deployment—can translate into model predictions and recommendations that run contrary to intended design, including by discriminating against disadvantaged subgroups (Crawford, 2013). At the same time, ML models are increasingly pervasive and already affect many facets of our daily lives, such as whether one should be granted a loan, let out of prison, receive admission to a university, be labeled a high-risk driver, or be promoted (Berk et al., 2018; Chouldechova et al., 2018). Hidden from sight, algorithms also determine much of the culture that is consumed online, by recommending music, images, news articles or movies. As stories emerge in the media of prison recidivism algorithms biased against African-Americans, adverts discriminating by populations, and chatbots turned racist soon after deployment, there has been an increasing awareness of the risks of discrimination in deployed ML models (Garcia, 2016; Barocas & Selbst, 2016).

While much has been written about the risks associated with ML, little progress has been made around how to systematically address and manage said risks. To address this, over thirty colleagues across diverse roles at QuantumBlack have collaborated to build a risk management system that flags fairness-related issues that may emerge in the machine learning process (Ahmed et al., 2019). In this system, over one hundred risks are flagged to practitioners at the relevant stage of the ML model-design and building process (see Table 1 for a selection). These risks have been crowd-sourced internally from actual challenges faced by practitioners applying ML in the field. The system also highlights a number of other risks related to a model’s ability to achieve high performance or explainability (which may be required by regulatory compliance, e.g. GDPR, or other business constraints). Given the groundswell of interest in standardizing and regulating model development (European Commission, 2018; Benkler, 2019), including initiatives such as the Office for AI and other (UK Government, 2018) and an algorithmic accountability bill proposed in the United States (Booker & Warden, 2019), coming to

a consensus on industry standards for managing risk is a timely issue. We see our proposal, along with further work towards an open and widely adopted standard, as progress in that direction.

Our aspiration with this work is to establish an industry-standard protocol, along with an accessible library of fairness-related risks that teams should consider when building machine learning models. Teams could then follow this protocol as they design, develop, and deploy their models to understand which risks are relevant to them at each stage of development and how these can be mitigated. For transparency, they can also document in a standardized way which risks were found to be relevant (or not relevant) to the model-building effort at hand, and what mitigation actions were taken.

In this workshop, we would like feedback on our proposal for an industry standard risk management protocol, as well as suggestions on how to best develop the platform and open-source the library of risks.

### 2. Problem

What protocols should teams follow in order to ensure ML models sustain high performance over time, are sufficiently explainable, and are not biased against subpopulations? What should be the norm expected of teams in terms of transparency and documentation of issues found?

There is a rich wealth of literature on highly-specific algorithmic techniques and metrics for measuring and reducing bias (Corbett-Davies & Goel, 2018). Little, though, has been published on what protocols should be followed to assess for fairness across the model development lifecycle.

Previous approaches to risk management in machine learning (Breck et al., 2017; Holland et al., 2018; Mitchell et al., 2019; Gebru et al., 2018; Arnold et al., 2018; Varshney et al., 2018) take the form of pre-production checklists: lists of questions that are typically considered or answered after modelling is completed (see, for example, Breck et al.’s rubric for ML production readiness, or the Model Card framework (Mitchell et al., 2019)). Our user research indicated that this checklist approach was insufficient. Practitioners found standalone checklists not only arduous to fill-out at model completion, but also disconnected from their work on a day-to-day basis during model creation. Identifying risks after a model has been built can often be too late for development efforts to be repeated. Further, practitioners primarily wanted advice on how to overcome risks, rather than simply questions that prompt them to consider risks.

Given the absence of a consensus on an industry standard to these questions and user needs, comprehensiveness of any ad hoc approach to ML risk management is highly variable.

### 3. Proposal

We propose that teams building ML models should follow an industry standard protocol of checks for risks at each step of the model creation lifecycle they complete. Further, especially in high-risk settings or in industries with regulation, teams should document for each risk whether it was found to be relevant or not, and what actions were taken to address it. We have developed a protocol for managing risks associated with fairness, explainability, and model performance (figure 1), while models are in development, based on our experience of risks encountered in the field. At the moment, this protocol is made available on an internal webapp.

Our approach could extend to a comprehensive risk platform, capturing and consolidating feedback from practitioners at large, as well as to the inclusion of technical tooling to facilitate frictionless and automated risk controls.

We are seeking to discuss (i) how to open this platform for practitioners to use and contribute to, leading towards an accepted industry standard on managing these risks; and (ii) how to develop technical tooling to automatically identify and mitigate risks during model development.

### 4. Current State

Our risk platform (Ahmed et al., 2019) currently comprises:

- An accessible introduction to each risk topic, that highlights key concepts, links to relevant code packages and resources, and summarizes the major steps that can be taken to reduce each risk.

- A risk library with over 100 risks (see table 1 for sample fairness-related risks), embedded within a model development protocol, which organizes the process of building machine learning models into high-level ‘activities’ and more detailed ‘tasks’ (figure 2). Practitioners can understand risks associated with the tasks they are currently completing (figures 2 and 3), rather than consulting checklists once the model is complete.

- Each risk is provided with advice on mitigating actions, crowd-sourced from practitioners within our company and reviewed by experts in the respective topic. This goes beyond previous approaches in the field of ML risk management, which typically prompt modellers to ask questions or consider risks, without consistently providing processes to deal with the risks.

- The content on the platform is structured in a standardized format (see figure 3), enabling it to grow cleanly through crowd-sourcing users’ contributions. The risk library contains risk and mitigation material developed
Table 1. Extract of sample risks from the library of ~100 risks, showcasing a selection of fairness-related risks to illustrate the content. Each risk has detailed mitigation material associated with it as well as stories from the field (figure 3).

<table>
<thead>
<tr>
<th>Activity</th>
<th>Task</th>
<th>Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Define success metrics</td>
<td>Define Model Evaluation Metrics</td>
<td>Missing fairness metrics - Fairness metrics are not defined, when they may be useful for the use-case</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fairness/Performance imbalance - The trade-off between performance and fairness metrics is not defined, resulting in a model with poor performance or insufficient emphasis on fairness</td>
</tr>
<tr>
<td>Assess the data</td>
<td>Profile the Data</td>
<td>Explicit sensitive attributes - There are sensitive or protected attributes explicitly included in the data, such as race, gender, or religion, which can lead to bias in a model against these groups</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Removed sensitive attributes - Sensitive attributes can be inferred from nonsensitive attributes in the data ('redundant encoding'), which heightens the possibility of an unfair model.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Imbalanced data - If most of the data comes from one subgroup, then the model may be inaccurate for other subgroups, leading to lower performance as well as risk of discrimination</td>
</tr>
<tr>
<td>Assess Data Quality</td>
<td></td>
<td>Inferior data quality - Data for a subgroup is missing, inaccurate, or otherwise biased, which can lead to unfairness and discrimination</td>
</tr>
<tr>
<td>Developing the analytical solution</td>
<td>Partition Data Set</td>
<td>Unrepresentative train/test split - Train/test splitting does not equally reflect proportions of sensitive characteristics in the data, leading to poor generalization of fairness assessments</td>
</tr>
<tr>
<td></td>
<td>Refine and Select Features</td>
<td>Minority features removed - Features that are predictive for subgroups but not majority groups are discarded in feature selection, leading to lower model performance for subgroups</td>
</tr>
<tr>
<td></td>
<td>Evaluate Model Performance</td>
<td>Unequal performance - Performance is lower for one subgroup relative to another</td>
</tr>
</tbody>
</table>

from practitioners’ challenges and experiences in applied settings. This allows us to capitalize on the experiences of others over many projects, to create a consistent and reliable approach. After every project that our company runs, teams can upload new risks encountered in their work, as well as mitigation suggestions based on their experiences.

5. Future Directions

To develop this risk platform into a broader protocol for the industry, it needs to become open-access, open to contributions from practitioners outside the company, and the scope of risks considered may need to be expanded or specialized. This can involve:

1. Creating an open-source site on which practitioners could submit risks, stories from their experience, and risk mitigation strategies;

2. Extending the scope of risks included to problems beyond machine learning, including causality analyses or optimization models;

3. Broadening to other categories of risk, including information security or regulatory risks;

4. Building industry-specific content, including risk mitigation libraries for regulated industries such as healthcare and banking; and

5. Using a practitioner community to stress-test the risk platform in real-world cases, especially where careful consideration of fairness and discrimination is particularly important.
We envisage the library of crowd-sourced risks in the platform eventually extending into technical tooling which, when appropriate, will complement existing debugging tools that are used to inspect ML models, such as Google’s What-If tool (Google PAIR Lab, 2018). This could include a data linter that flags potential biases within data sources, or an open-source model pipelining framework, that is able to assess risks at defined stage-gates.

6. Impact

Our goal is to ensure that machine learning models carry far less risk to under-served subgroups. We hope that this combination of risk identification and mitigation advice will promote the adoption of the latest and most-effective packages for improving fairness in algorithms. Longer term, software tooling will enable frictionless risk management, much as bugs and programming risks are caught by code linters today.

By standardizing ML risk management within an organisation, team leaders will be able to better understand and prioritize the risks associated with their projects; this is especially helpful in companies where senior stakeholders are not technical. Finally, by standardizing ML risk management across organisations, we believe we can build public trust and confidence in the models that these organisations deploy.

7. Risks

Our solution is “crowd-sourced” in spirit and hinges on capturing collective experiences and expertise from many model development projects. Being open-source enables us to accelerate growth and improvement of the platform. However, there are two key challenges with this approach:

1. Recognizing that this risk library is not—and never will be—an exhaustive solution.

2. Ensuring continual contributions to risk content, and to any technical tooling.

This first point needs to be well understood by development teams adopting the system as their standard approach for mitigating risks in model development. There may always be new situations that are not yet codified within the risk library. To address this, the platform should inform teams of this caveat and encourage ad hoc examinations of risks and protocols pertinent to new use-cases.

To the second point, we plan to continue contributing to this platform internally. We would like to invite a discussion about how we can facilitate open-source contributions.

8. Social System

Our interdisciplinary team includes those with backgrounds in data science, data engineering, machine learning engineering, risk management, management consulting, product management, applied ethics, user experience, and design, as well as lawyers and information security experts. They have come from a wide array of disciplines, ranging from healthcare and computer science to philosophy. The team also exhibits diversity in terms of gender and ethnicity as well as cultural and socioeconomic background.

To successfully open source our framework, we would need to complement our team with collaborators skilled in managing online communities of open-source platforms (e.g. Wikipedia), who can administrate open-source submissions. To support automated tooling of the risk framework, we would look to collaborate with experienced software engineers.

9. Technical System

1. Open-sourcing the risk library: Our system is currently hosted on an internal web app and would need to be moved to an open website, accessible and open to contributions from the wider ML community. This would require significant development effort.

2. Technical Tooling: In parallel, we can start to build the risks identified by practitioners into software toolkits. This could take the form of a linter for an end-to-end modeling pipeline.

A baseline approach for a data linter exists in the literature (Schelter et al., 2018; Hynes et al., 2017), and we would aim to improve on this by incorporating monitoring for risks from the library we have developed.

To develop a tool that flags unnoticed risks during model building, the risk system could be integrated with model pipelining software.

References


Building Fair and Transparent Machine Learning via Operationalized Risk Management

Barocas, S. and Selbst, A. D. Big data’s disparate impact. 

Benkler, Y. Don’t let industry write the rules for AI. 

Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. Fairness in criminal justice risk assessments: The state of the art. 


Breck, E., Cai, S., Nielsen, E., Salib, M., and Sculley, D. 
The ML test score: A rubric for ML production readiness and technical debt reduction. 

Chouldechova, A., Benavides-Prado, D., Fialko, O., and Vaithianathan, R. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. 


Crawford, K. The Hidden Biases in Big Data. 


Garcia, M. Racist in the machine: The disturbing implications of algorithmic bias. 


Presentation at the NeurIPS Expo Demo, 2018.
Explore risk topics

<table>
<thead>
<tr>
<th>Performance</th>
<th>Explainability</th>
<th>Fairness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Review factors and potential risks that impact model performance – ranging from mistakes in data collection, to a failure in addressing constraints and business requirements.</td>
<td>Explore potential pitfalls a team may encounter if explaining model predictions and recommendations is a significant requirement on their engagement.</td>
<td>Models can sometimes be unfair to certain individuals or categories of people, for example women or non-whites, by having lower accuracy or biased results for these groups.</td>
</tr>
</tbody>
</table>

Learn More | Learn More | Learn More

The above only covers a subset of risks that may impact an analytics engagement.

All teams are encouraged to run a Pre-mortem Workshop at the beginning of their engagement, inviting input from all areas of the business to establish a holistic view of risks that should be considered during their engagement.

Any team’s looking to capture new learnings or mitigation strategies, should follow the retrospective process governed by their guild.

---

*Figure 1. The risk mitigation system covers three categories of risk.* Users explore the risk mitigation system through a webapp interface.
The protocol for model-building is broken up into more than 20 activities.

Each activity and task in the model-building process has risks linked to it. Each risk is related to a specific task within each activity. Risks and their corresponding mitigations are recorded in a standardized way (see figure 3).
Figure 3. Each risk has stories from the field and mitigation suggestions attached to it. The stories either highlight the impact of the risk or help a team see how to overcome a challenging situation. Each risk has associated reactions to take in response, that are categorised into actions that Assess, Mitigate, or Communicate the risk.