Imran Ahmed, Giles L. Colclough, Daniel First[1], QuantumBlack Contributors[2]

QUANTUMBLACK
A MCKINSEY COMPANY

Building Fair and Transparent Machine Learning via Operationalized Risk Management: Towards an Open-Access Standard Protocol

Background & Motivation

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. There are as yet no broadly-adopted standard approaches to ensure that models meet these requirements.

It is critical that models

Generalize well in production

Are sufficiently explainable

Are fair to all groups

Literature Review

Worked Examples

Qualitative assessments, such as Racist in the Machine[3], which detail the impact of unaddressed risks on models, and their societal implications.

Post-Hoc Checklists

Checklists, such as ML Test Score[4], covering potential risks across a range of axes from model performance to legal concerns.

These are used to audit and document models after they're deployed.

Specialized Technical Tools

Technical packages, such as What-If tool[5], which are used to debug various aspects of machine learning projects

Tools typically inspect for a particular or related set of issues.

User Research

Extensive interviews revealed that our users wanted a "one-stop shop" linked to the way they worked, that would help them identify and overcome the most relevant risks.

Previous approaches to risk management in machine learning 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.

Requirements

PRACTICAL
Practitioners can manage risk as they go through projects, and can know which risks are relevant at any point in time

ACTIONABLE
Each risk includes practical solutions to adress them

COMPREHENSIVE
Risks considered across model development lifecycle

UNIFIED
Risks are presented in standardized form across all risk categories

SCALABLE
Consistent format enables scaling

Personas

PRACTITIONER DS/DE

"Which risks are relevant to the tasks I am doing now?"

"How have other teams handled these challenges?"

TEAM MANAGER

"How can I help my team prioritize and scope for risks?"

"How I can be sure I've considered all risks, comprehensively"

COMPANY LEADERSHIP

"Where are our gaps?"

"How can we make sure we don't make the same mistake twice?"

Operationalized Risk Management

Key Contributions

i. Risks embedded in an ML Model-Building Protocol
Allows practitioners to manage risks as they are building models, rather than auditing for risks after the models have been created

Enables practitioners to quickly find the risks and mitigation materials that are most relevant to the tasks they are doing

ii. Mitigations That Capitalize on Expertise
This is the first approach to managing risk in machine learning that uses a scalable system to record mitigations along with risks, informed by historical experience and reviewed by experts

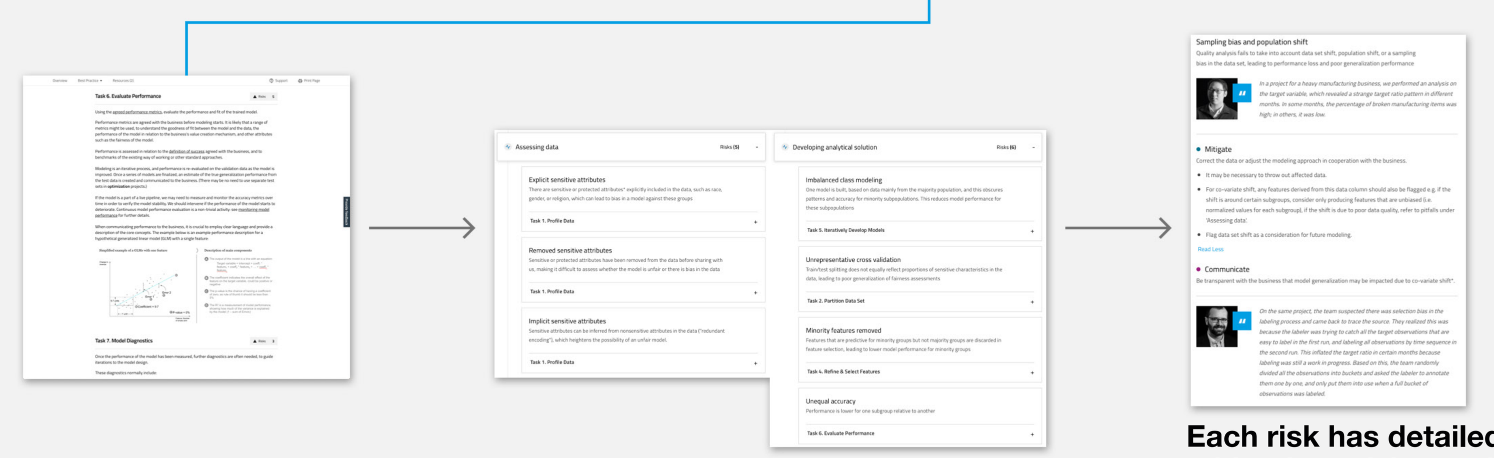
iii. Consistent Conceptual Structure
Risks, Mitigations, and War Stories are captured in a consistent conceptual structure, to facilitate scaling by adding risks and mitigations after each project

Our Risk Management Protocol & Webapp

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.

Risks are embedded in an ML model creation protocol

The protocol for model-building is broken up into more than 20 activities.



Each activity is broken into tasks - over 125 total

Risks are associated with each tast

Each risk has detailed mitigation advice and case studies

Sample Fairness Risks Selected from Library of 100+ Risks

Activity	Task	Risks
03 **Define success metrics** What metrics are appropriate, and how should they be defined?	Define Model Evaluation Metrics	**Missing fairness metrics** - Fairness metrics are not defined, when they may be useful for the use-case
		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
05 **Assess the data** Does the data exhibit any qualitiesthat s hould inform the modelling approach?	Profile the Data	**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
		Removed sensitive attributes - Sensitive attributes can be inferred from nonsensitive attributes in the data ('redundant encoding'), which heightens the possibility of an unfair model.
		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
	Assess Data Quality	**Inferior data quality** - Data for a subgroup is missing, inaccurate, or otherwise biased, which can lead to unfairness and discrimination
08 **Developing the analytical solution** What models should be built to solve the problem?	Partition Data Set	**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
	Partition Data Set	**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
	Partition Data Set	**Unequal performance** - Performance is lower for one subgroup relative to another

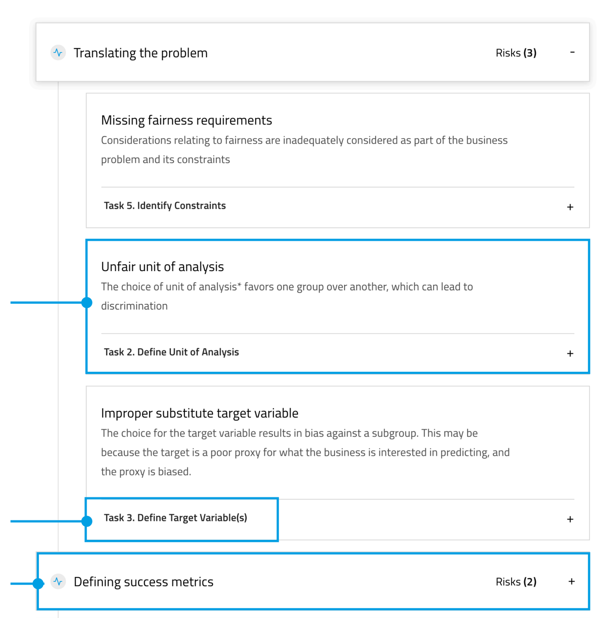
Structure of Risks and Mitigations:

Risks

Risks are embedded in tasks and have a consistent conceptual structure

Risks are formulated in one sentence with an impact clause, enabling consistency and scalability as new risks are added

Risks are nested within high-level "activities" and granular "tasks", so users can quickly identify the risk relevant to them

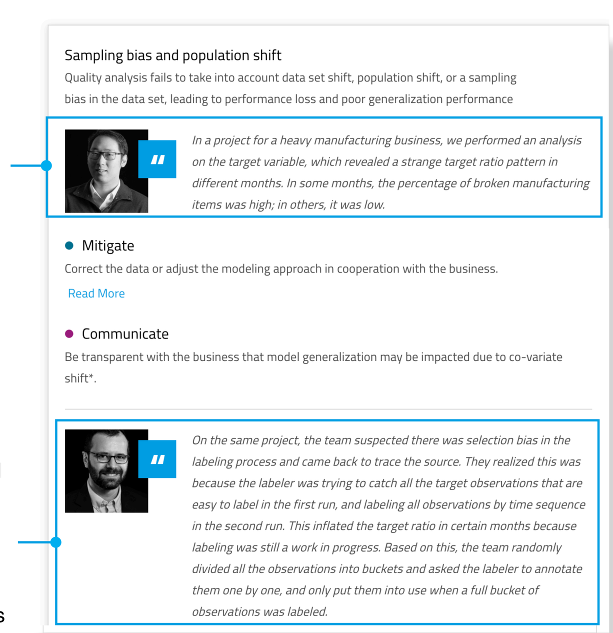


Mitigations

Each Risk comes along with a Mitigation. Mitigations include war stories and are broken up into: Assess, Mitigate, Communicate

"Risk War Stories" highlight challenges teams have faced in the past, to help bring risks to life and illustrate impact

"Mitigation War Stories" help teams learn how to overcome challenges



Future State: for Discussion

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.

Future directions

01 **Expand Scope of Risks/Mitigations**

a. Technique-specific Risks i.e. Deep learning, casuality analysis, or optimisation

b. Domain-specific Risks i.e. Health-care, banking, or insurance

c. Risk Themes i.e. Information security or regulatory risks

d. Open-Sourcing i.e. Make the risk and mitigation library public

02 **Develop technical tools**

A data linter that flags potential biases within data sources ...

An open-source model pipelining framework, that is able to assess risks at defined stage-gates

03 **Stress test risk protocol on applied ML studies across industries**

Impact

ADOPTION
Businesses will be quicker to adopt ML, as it will be less risky

TRANSLATION
Researchers will be able to access the latest techniques from academic literature more easily

TRANSPARENCY
Team leaders and business users will have more visibility into the risks that ML carries

Potential Solution for Practitioners: a Risk Mitigation Worksheet to record and communicate project risks

Risk Library	Finding	Mitigation
Assess the Data		
Inconsistencies in data collection or recording	Dates were sometimes recorded in US format (MMDD), sometimes in UK format (DDMM)	We used public injury data and ad-hoc examination to correct the dates
Data for a subpopulation is missing, inaccurate, or otherwise biased, which can lead to unfairness and discrimination	"Star players" had more accurate injury data than players typically on the bench	We tested whether the model performed better when we excluded players who did not play often
Align on Modelling Requirements		
Difficulties forecasting of the amount of time it would take to implement XAI	LIME was not straightforward to apply due to missing values	We brought a dedicated expert in LIME to build a custom solution. LIME did not work for four players who do not play often due to missing values. This has been flagged to management.
Define the analytical problem		
Target variable is an imperfect proxy	Data for the target variable comes from actual injuries, which is a proxy for what the model aims to predict: 'likelihood for injury'	We went in and removed all contact injuries. We also removed days where the players were not playing
Evaluation Metric does not match business use case	ROC vs. APS: does the high-precision model matter most?	We used Average Precision Score instead

Practitioners can create transparency about the risks that emerge on a study and the actions that were taken to mitigate them by filling out a risk mitigation worksheet and flagging to team leadership

Questions for Discussion:

01 **Would you use a risk management approach in your work?**

02 **What is your company's approach to ensuring performance, explainability, and fairness?**

03 **Would you contribute to an open-source risk library?**

04 **Do you have any suggestions for technical tooling?**

1 Corresponding author: daniel.first@quantumblack.com
2 The risk management system was created through a collaboration between over thirty colleagues, including data engineers, data scientists, product managers, management consultants, lawyers, and information security experts. Contributors included: Shubham Agrawal, Roger Burkhardt, Jacomo Corbo, Rupam Das, Marco Diciolla, Mohammed ElNabawy, Konstantinos Georgatzis, Carlo Giovine, Stephanie Kaiser, Matej Macak, George Mathews, Ines Marusic, Helen Mayhew, James Mulligan, Alejandra Parra-Orlandoni, Erik Pazos, Antenor Rizo-Patron, Joel Schwartzmann, Vasiliki Stergiou, Andrew Saunders, Suraj Subrahmanyan, Toby Sykes, Stavros Tsalides, Julian Waton, Ian Whalen, Chris Wigley, Didier Vila, Jiaju Yan, Jun Yoon, and Huilin Zeng.

3 Garcia, M. Racist in the machine: The disturbing implications of algorithmic bias. World Policy Journal, 33(4): 111-117, 2016
4 Breck, E., Cai, S., Nielsen, E., Slib, M., and Sculley, D. The ML test score: A rubric for ML production readiness and technical debt reduction. In 2017 IEEE International Conference on Big Data (Big Data), pp. 1123-1132, Dec 2017. doi: 10.1109/BigData.20178258038.
5 Google PAIR Lab. The What-If Tool: Code-Free Probing of Machine Learning Models, 2018