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

Imran Ahmed   Giles L. Colclough   Daniel First[1]   **QuantumBlack contributors**[2]
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 in the UK (UK Government, 2018) and an algorithmic accountability bill proposed in the United States (Booker & Warden, 2019), coming to

[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, 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, Ian Whalen, Chris Wigley, Didier Vila, Jiaju Yan, Jun Yoon, and Huilin Zeng. All authors are listed in alphabetical order.

Appearing at the International Conference on Machine Learning AI for Social Good Workshop, Long Beach, United States, 2019.

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 on the back of these findings.

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 developers 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 $\sim$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).

Activity	**Task**	**Risk**
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
Assess the data Does the data exhibit any qualities that should 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
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
	Refine and Select Features	**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
	Evaluate Model Performance	**Unequal performance** - Performance is lower for one subgroup relative to another

from practitioners' challenges and experiences in applied settings. This allows us to capitalize on the experience 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

Ahmed, I. S., Colclough, G. L., First, D., and QuantumBlack contributors. Operationalizing Risk Management for Machine Learning: Building a Protocol-Driven System for Performance, Explainability, and Fairness. In *Debugging Machine Learning Models Workshop, ICLR*, 2019.

Arnold, M., Bellamy, R. K. E., Hind, M., Houde, S., Mehta, S., Mojsilovic, A., Nair, R., Ramamurthy, K. N., Reimer, D., Olteanu, A., Piorkowski, D., Tsay, J., and Varshney, K. R. FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity. *IBM technical report*, arXiv:1808.07261, 2018.

Barocas, S. and Selbst, A. D. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

Benkler, Y. Don't let industry write the rules for AI. *Nature*, 569(7755):161–161, 2019. doi: 10.1038/d41586-019-01413-1.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, pp. 0049124118782533, 2018.

Booker, C. and Warden, R. Algorithmic Accountability Act of 2019, S.1108, 2019.

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. In *2017 IEEE International Conference on Big Data (Big Data)*, pp. 1123–1132, Dec 2017. doi: 10.1109/BigData.2017.8258038.

Chouldechova, A., Benavides-Prado, D., Fialko, O., and Vaithianathan, R. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*, pp. 134–148, 2018.

Corbett-Davies, S. and Goel, S. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

Crawford, K. The Hidden Biases in Big Data. *Harvard Business Review*, Apr 2013.

European Commission. Ethics guidelines for trustworthy AI, 2018.

Garcia, M. Racist in the machine: The disturbing implications of algorithmic bias. *World Policy Journal*, 33(4): 111–117, 2016.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., III, H. D., and Crawford, K. Datasheets for Datasets. In *Proceedings of the 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning, Stockholm, Sweden*, 2018.

Google PAIR Lab. The What-If Tool: Code-Free Probing of Machine Learning Models, 2018.

Holland, S., Hosny, A., Newman, S., Joseph, J., and Chmielinski, K. The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards. arXiv:1805.03677, 2018.

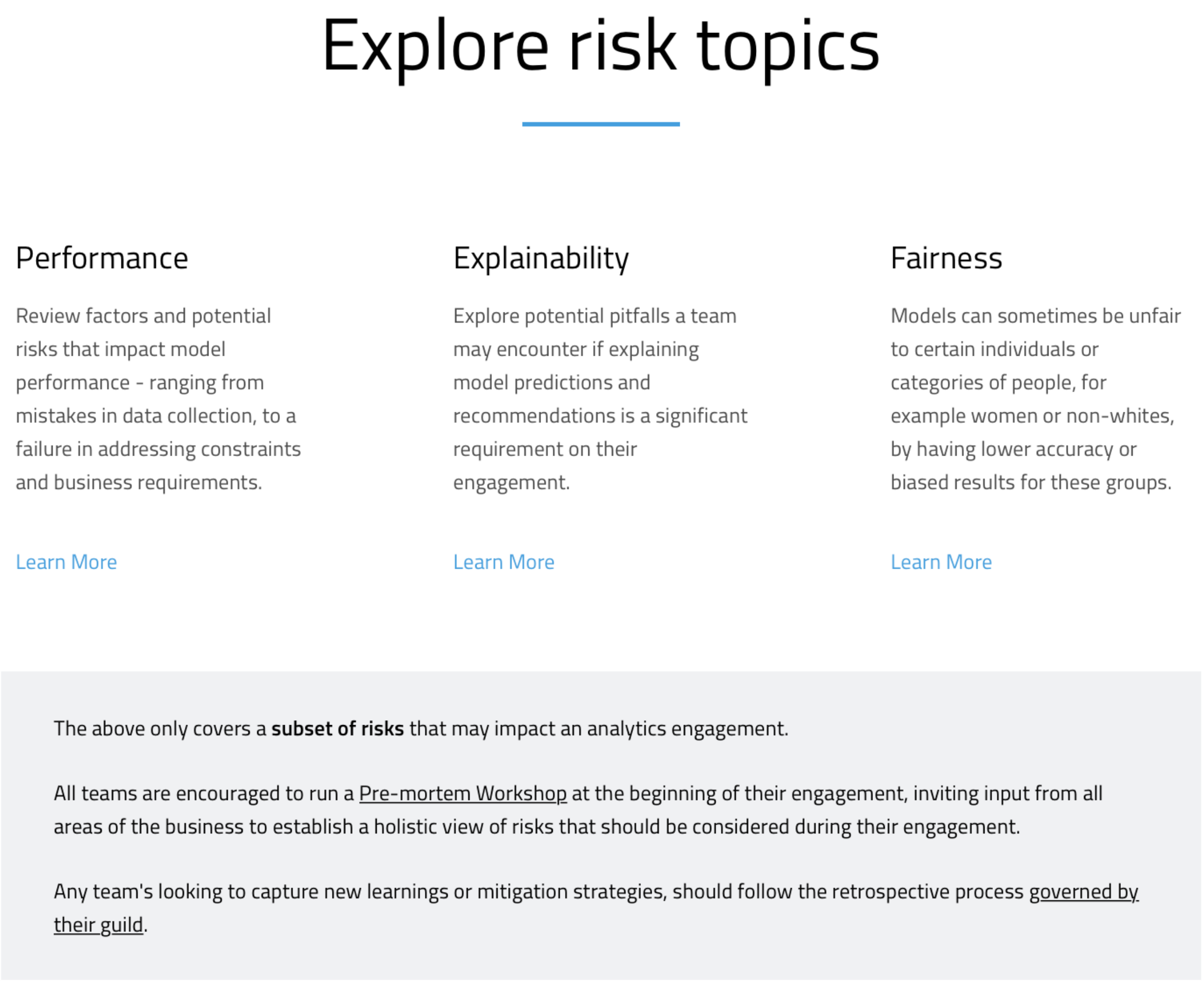
Hynes, N., Sculley, D., and Terry, M. The Data Linter: Lightweight, Automated Sanity Checking for ML Data Sets. Workshop on ML Systems, NeurIPS, 2017.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency—FAT* '19, Atlanta, GA, USA*. ACM Press, New York, NY, USA, 2019.

Schelter, S., Grafberger, S., Schmidt, P., Rukat, T., Kiessling, M., Taptunov, A., Biessmann, F., and Lange, D. Deequ-Data Quality Validation for Machine Learning Pipelines. Machine Learning Systems Workshop, NeurIPS, 2018.

UK Government. Office for Artificial Intelligence, 2018.

Varshney, K. R., Wei, D., Dhurandhar, A., Ramamurthy, K. N., and Tsay, J. Automatic Generation of Factsheets for Trusted AI in a Runtime Environment. Presentation at the NeurIPS Expo Demo, 2018.



Explore risk topics

Performance

Review factors and potential risks that impact model performance - ranging from mistakes in data collection, to a failure in addressing constraints and business requirements.

Learn More

Explainability

Explore potential pitfalls a team may encounter if explaining model predictions and recommendations is a significant requirement on their engagement.

Learn More

Fairness

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.

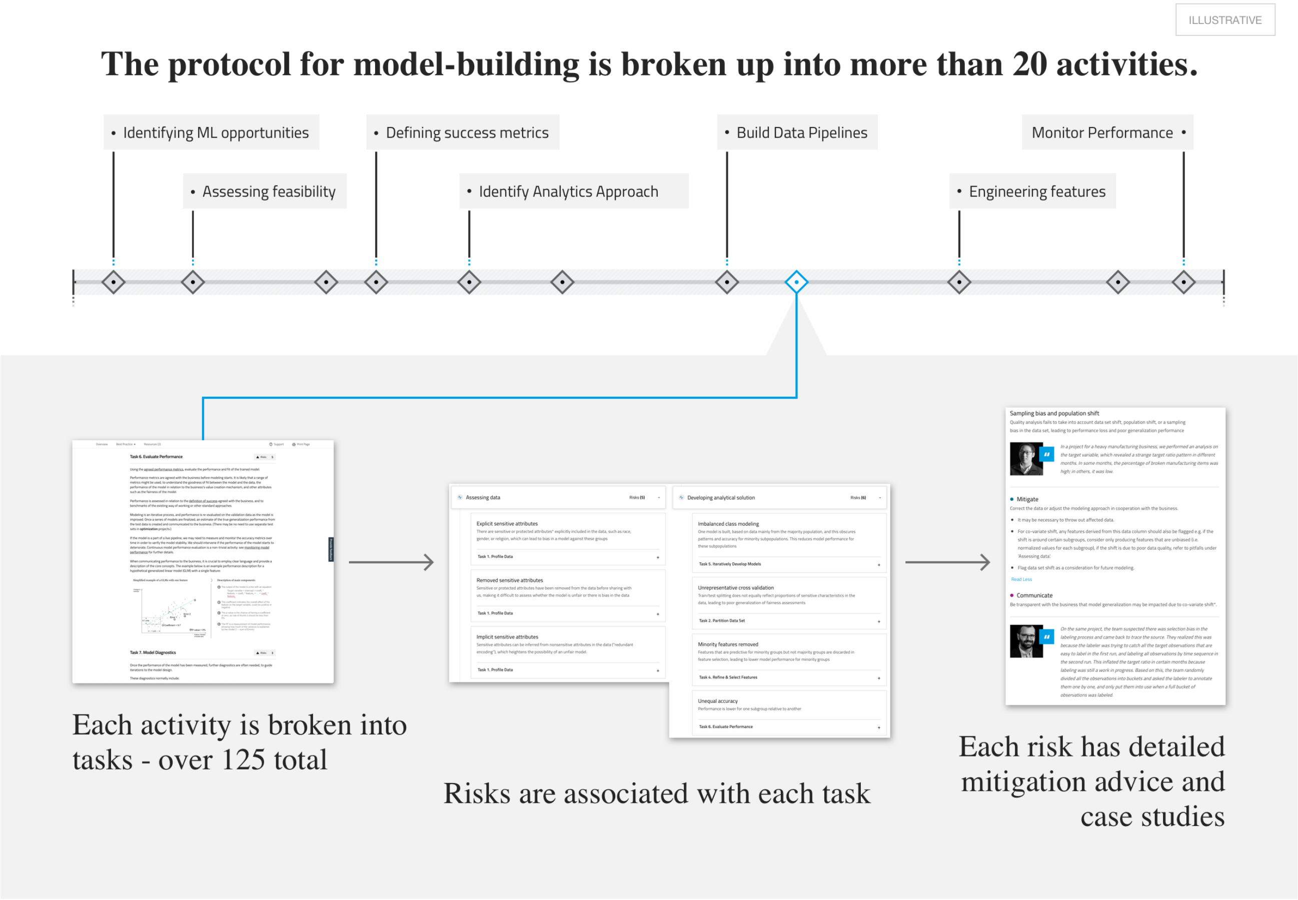
Learn More

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.



ILLUSTRATIVE

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

- Identifying ML opportunities
- Assessing feasibility
- Defining success metrics
- Identify Analytics Approach
- Build Data Pipelines
- Engineering features
- Monitor Performance

Task 6. Evaluate Performance ▲ Risks 5

Using the agreed performance metrics, evaluate the performance and fit of the trained model.

Performance metrics are agreed with the business before modeling starts. It is likely that a range of metrics might be used, to understand the goodness of fit between the model and the data, the performance of the model in relation to the business's value creation mechanism, and other attributes such as the fairness of the model.

Performance is assessed in relation to the definition of success agreed with the business, and to benchmarks of the existing way of working or other standard approaches.

Modeling is an iterative process, and performance is re-evaluated on the validation data as the model is improved. Once a series of models are finalized, an estimate of the true generalization performance from the test data is created and communicated to the business. (There may be no need to use separate test sets in **optimization** projects.)

If the model is a part of a live pipeline, we may need to measure and monitor the accuracy metrics over time in order to verify the model stability. We should intervene if the performance of the model starts to deteriorate. Continuous model performance evaluation is a non-trivial activity: see monitoring model performance for further details.

When communicating performance to the business, it is crucial to employ clear language and provide a description of the core concepts. The example below is an example performance description for a hypothetical generalized linear model (GLM) with a single feature:

Simplified example of a GLMs with one feature

Description of main components

1. The output of the model is a line with an equation: Target variable = intercept + coeff₁ * feature₁ + coeff₂ * feature₂ + ... + coeffₙ * featureₙ
2. The coefficient indicates the overall effect of the feature on the target variable, could be positive or negative
3. The p-value is the chance of having a coefficient of zero, as rule of thumb it should be less than 5%
4. The R² is a measurement of model performance, showing how much of the variance is explained by the model (1 – sum of Errors)

Task 7. Model Diagnostics ▲ Risks 3

Once the performance of the model has been measured, further diagnostics are often needed, to guide iterations to the model design.

These diagnostics normally include:

Each activity is broken into tasks - over 125 total

Assessing data Risks (5)

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
Task 1. Profile Data +

Removed sensitive attributes
Sensitive or protected attributes have been removed from the data before sharing with us, making it difficult to assess whether the model is unfair or there is bias in the data
Task 1. Profile Data +

Implicit sensitive attributes
Sensitive attributes can be inferred from nonsensitive attributes in the data ("redundant encoding"), which heightens the possibility of an unfair model.
Task 1. Profile Data +

Developing analytical solution Risks (6)

Imbalanced class modeling
One model is built, based on data mainly from the majority population, and this obscures patterns and accuracy for minority subpopulations. This reduces model performance for these subpopulations
Task 5. Iteratively Develop Models +

Unrepresentative cross validation
Train/test splitting does not equally reflect proportions of sensitive characteristics in the data, leading to poor generalization of fairness assessments
Task 2. Partition Data Set +

Minority features removed
Features that are predictive for minority groups but not majority groups are discarded in feature selection, leading to lower model performance for minority groups
Task 4. Refine & Select Features +

Unequal accuracy
Performance is lower for one subgroup relative to another
Task 6. Evaluate Performance +

Risks are associated with each task

Sampling bias and population shift
Quality analysis fails to take into account data set shift, population shift, or a sampling bias in the data set, leading to performance loss and poor generalization performance

In a project for a heavy manufacturing business, we performed an analysis on the target variable, which revealed a strange target ratio pattern in different months. In some months, the percentage of broken manufacturing items was high; in others, it was low.

- Mitigate

Correct the data or adjust the modeling approach in cooperation with the business.

- It may be necessary to throw out affected data.
- For co-variate shift, any features derived from this data column should also be flagged e.g. if the shift is around certain subgroups, consider only producing features that are unbiased (i.e. normalized values for each subgroup), if the shift is due to poor data quality, refer to pitfalls under 'Assessing data'.
- Flag data set shift as a consideration for future modeling.

Read Less

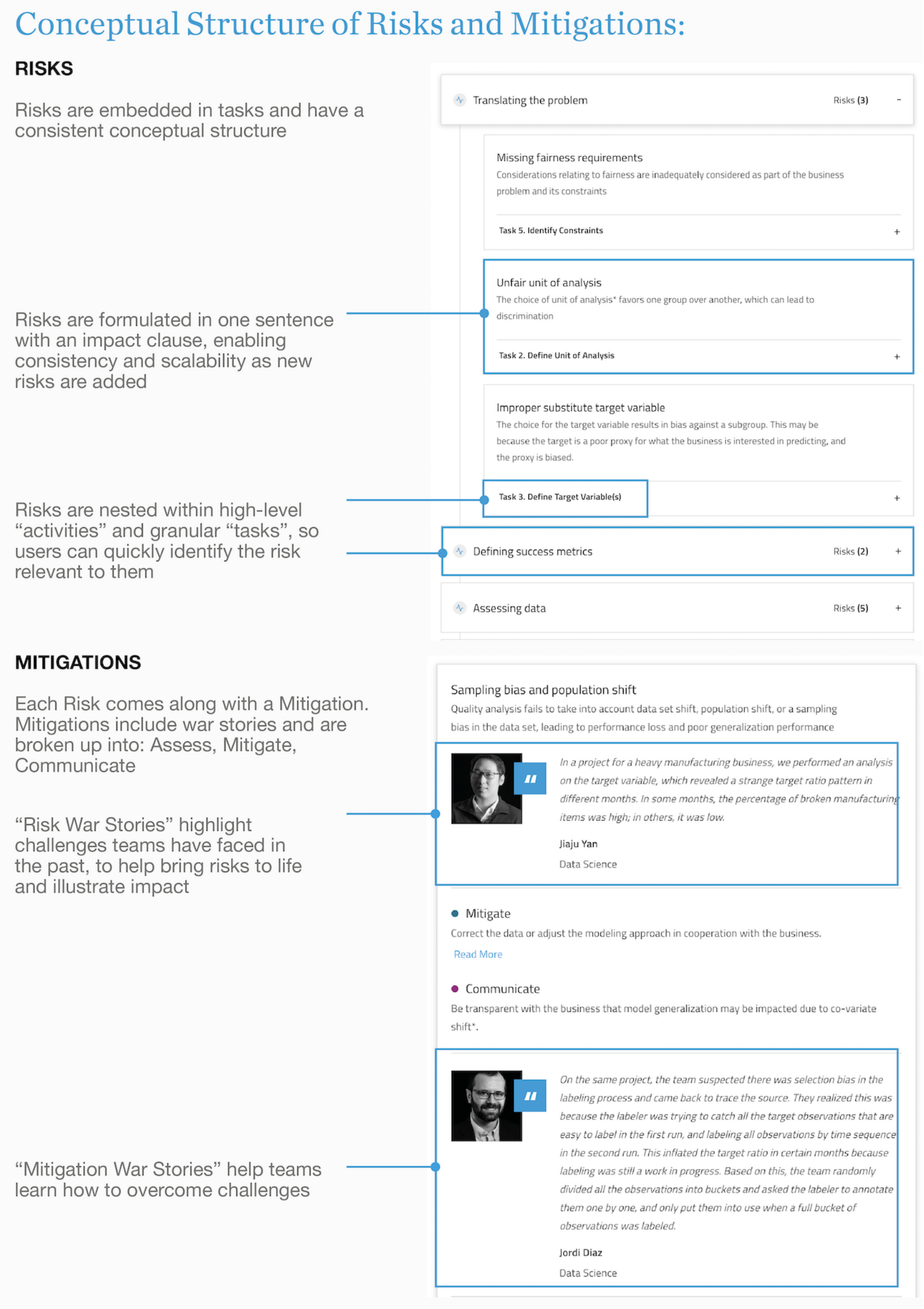
- Communicate

Be transparent with the business that model generalization may be impacted due to co-variate shift*.

On the same project, the team suspected there was selection bias in the labeling process and came back to trace the source. They realized this was because the labeler was trying to catch all the target observations that are easy to label in the first run, and labeling all observations by time sequence in the second run. This inflated the target ratio in certain months because labeling was still a work in progress. Based on this, the team randomly divided all the observations into buckets and asked the labeler to annotate them one by one, and only put them into use when a full bucket of observations was labeled.

Each risk has detailed mitigation advice and case studies

Figure 2. **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).



Conceptual Structure of Risks and Mitigations:

RISKS

Risks are embedded in tasks and have a consistent conceptual structure

Translating the problem Risks (3) -

Missing fairness requirements
Considerations relating to fairness are inadequately considered as part of the business problem and its constraints

Task 5. Identify Constraints +

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

Unfair unit of analysis
The choice of unit of analysis* favors one group over another, which can lead to discrimination

Task 2. Define Unit of Analysis +

Improper substitute target variable
The choice for the target variable results in bias against a subgroup. This may be because the target is a poor proxy for what the business is interested in predicting, and the proxy is biased.

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

Task 3. Define Target Variable(s) +

Defining success metrics Risks (2) +

Assessing data Risks (5) +

MITIGATIONS

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

Sampling bias and population shift
Quality analysis fails to take into account data set shift, population shift, or a sampling bias in the data set, leading to performance loss and poor generalization performance

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

In a project for a heavy manufacturing business, we performed an analysis on the target variable, which revealed a strange target ratio pattern in different months. In some months, the percentage of broken manufacturing items was high; in others, it was low.

Jiaju Yan
Data Science

● Mitigate
Correct the data or adjust the modeling approach in cooperation with the business.
Read More

● Communicate
Be transparent with the business that model generalization may be impacted due to co-variate shift*.

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

On the same project, the team suspected there was selection bias in the labeling process and came back to trace the source. They realized this was because the labeler was trying to catch all the target observations that are easy to label in the first run, and labeling all observations by time sequence in the second run. This inflated the target ratio in certain months because labeling was still a work in progress. Based on this, the team randomly divided all the observations into buckets and asked the labeler to annotate them one by one, and only put them into use when a full bucket of observations was labeled.

Jordi Diaz
Data Science

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.