TC Charter: Risk Card Standard

John Stockton
10/6/2022

Section 1: Technical Committee Charter

TC Name
Risk Card Standard

Statement of Purpose

Private and public organizations are often tasked with algorithmically assessing the risk posed by a named entity, such as a person or organization. For example, in fighting financial crimes, bank investigators must adhere to government imposed know-your-customer (KYC) and anti-money laundering (AML) regulations, which result in the filing of suspicious activity reports (SAR's) in the US. Law enforcement agencies, who receive these reports, compile this information with other intelligence to generate their own assessments. The supply-chain risk management industry faces similar challenges. These threat-detection ecosystems include tech companies who build models for identifying risk, NGO's who specialize as domain experts in particular categories of risk, investigative journalists who provide raw news content, regulators who oversee industry, and multiple allied defense intelligence organizations across the world who address risks from a global perspective.

This "Risk Card" standard is meant to create a common risk-centric language by which all of these organizations can effectively communicate. This language is not only for the benefit of human-to-human communication, but also for human-to-machine communication (e.g., training models to identify a particular risk), and machine-to-human communication (e.g., a user interface flagging and explaining a risk for a screened entity). In other words, the definitions of each risk need to be made "computable".

For example, assume that a bank is tasked with building models to identify which customers are likely involved with a risk like "Wildlife Trafficking". Or to identify which customers are connected to "Corruption". The problem is that, without further information, these risks are ill-defined. Does "Wildlife Trafficking" include illegal fishing? It depends on where you look. Does "Corruption" include accusations or only criminal convictions? Opinions vary.

When organizations define risk typologies too loosely or too inconsistently, the process of detecting and eliminating threats is negatively impacted in many ways. If definitions are too broad, false positives decrease the efficiency of investigators. If too narrow, then key signals may be missed, resulting in false negatives. If too inconsistent, then cases may be directed to the wrong investigators, who are assigned by threat. International differences in policy, terminology, and language will break assumptions. Models can't be trained or tested effectively. Ultimately, the resulting dissonance can lead to a lack of faith in the intelligence system, such that the power of automation is abandoned when it is needed the most.

This standard is meant to establish a framework where risk labels can be made objective and computable. With the resulting Risk Cards, technical model builders will know what to build, their algorithms will transparently explain results to build trust, and non-technical practitioners will know what risk labels mean (and don't mean) such that they can be effectively processed.

Business Benefits

Banks will be able to explain their entity risk labels to regulators and law enforcement.

Intelligence agencies, both within and across countries, will be able to share and integrate intelligence assets with consistent, well understood, and explainable threat labels.

NGO's and domain experts will be able to share particular signals and potential model features to data scientists who build models for automated risk detection.

Tech companies will have clean "specifications" for how to consistently build, train, and test risk models. In other words, the Risk Cards help "train the trainers".

Scope

The scope of this standard is limited to the structure and production guidelines for the Risk Cards. The structure will include elements such as definitions, training data examples, signals, relevant data sets, and more as described within the standard itself. The goal is to set a standard that is clarifying and helpful for both the model builder and the model consumer.

The following is not necessarily in scope:

- Particular Risk Cards. The standard, at least the base version, does not need to include any particular Risk Cards themselves. Certain risks may be discussed by way of example, but the goal is to define the shape of a Risk Card, not produce a complete set of content. Ultimately, we would like to set a standard where different organizations (public or private) can create their own Risk Cards, and those can be redundant with others. Any given model should reference a Risk Card that meets the standard stated here, wherever it is published and whoever it is authored by. Ideally, there would be a base set of Risk Cards that many organizations use and iterate over collectively, but there should also be healthy competition and freedom for companies to implement their own versions.
- Categories or Taxonomies of Risk. Quantifind has its own list of Risks, driven by clients and aggregated across a broad number of public and private taxonomies (e.g., Interpol website, DOJ categories, FinCEN guidance), and its own taxonomy for placing those risks into clusters (e.g., Financial Health, Financial Crimes, National Security, ESG). However, these also include subjective decisions, and the standard does not concern itself with hierarchies. For the purpose of this standard, each Risk can be considered independently with its own Risk Card. Also, the standard does not consider "severity" or

"strength" of risk, because this is also context dependent. The risk will be considered as a boolean label on an entity, which may be implemented in ways that express confidence in that label, but not how much risk it presents to the end client (regarding money, safety, or similar).

- Models that implement Risk Cards. The Risk Card standard is similar to and inspired by the Model Card standard, but they play a more supportive role. A Model Card, for example, could implement a Risk Card through a classifier model. The goal of the Risk Card then, is to define the objective of any model, but not the model itself, which may take many forms. However, the Risk Card approach is consistent with and supportive of the development of "explainable" models that transparently highlight why they returned with the results they did on any given example (e.g., what features from the Risk Card were triggered by input data). The Risk Card Guidelines will also distinguish between models that are explainable with direct evidence from one record, as opposed to multiple pieces of indirect evidence from multiple records. They will also advise model-builders on how to distinguish between those entities who deserve the risk label, versus "innocent bystanders" in data sets who either are on the good side of the risk (e.g., law enforcement), or it is unclear. In some cases the models may need to err on the side of caution, and tradeoff more false positives for less false negatives.
- Other extracted entity information. This standard could in principle be "generalized" to include the association of an entity with other information. E.g., we could have called them "Label Cards" to include not only inferred Risks, but also inferred metadata for an entity (ages, addresses, etc.) or inferred relationships between entities. While such a generalized approach is possible, we limit the scope of this standard to inferred Risks.
- Entity recognition or resolution. Although important for a fully functioning risk engine, this standard does not consider the problem of entity recognition (pulling names from unstructured text) or entity resolution (determining that two records refer to the same individual), and these can be treated as separate problems. However, assumed entity resolution may arise in the determination of some risks. For the sake of this standard, a model based on a Risk Card is assumed to apply to either a single record or a collection of records known to refer to the same real-world entity.

Deliverables

The committee will align on the design of the following deliverables: **Risk Card Template**, **Risk Card Example**, **Risk Card Guidelines, and Risk Card Registry**.

This **Risk Card Template** will consider the following elements in its construction:

- Definition
- Inclusions/Exclusions
- Alternate Definitions
- Relevance
- Terms
- Representative Entities

- Data Sources
- Signals and Features
- Related Risks
- Related Industries
- References

The template will adequately describe what is intended for each section, in terms of acceptable content and length. Certain sections, e.g., Representative Entities, may refer to supplementary material with much more information.

The **Risk Card Example** will include one or more specific implementations of the Risk Card Template. These are not meant to be standards in themselves, but to give more educational context to the Risk Card Template.

The **Risk Card Guidelines** will cover suggested approaches for implementing existing Risk Cards in models, creating new Risk Cards, versioning of updates, quality control, and auditing.

The **Risk Card Registry** will give notes on what Risk Cards exist, where they can be found, and what other risks may be developed

IPR Mode

Non-Assertion – requires all Obligated Parties to provide an OASIS Non-Assertion Covenant as described in Section 10.3.

Quantifind intends to own the rights to those particular Risk Cards, and the models that implement those Risk Cards, that Quantifind develops with certain partners. However, the Risk Card Standard will not be owned by Quantifind.

Audience

The audiences that are relevant to the construction of Risk Cards include:

- Financial Institutions
    - Banks operating under BSA regulations
    - Model builders within banks
    - Investigators within banks
- Regulators
    - E.g., FinCEN and OCC
- Law Enforcement
    - Domestic and International
- Intelligence and Defense Agencies
- Financial Crime Consortia
    - E.g., GCFFC, ACFCS, ACAMS
- FFRDC
    - E.g. MITRE

- UARC
    - E.g., ARLIS
- NGO (Non-Governmental Organizations)
    - Mission-driven organizations focused on particular risks (e.g., Polaris for Human Trafficking, United for Wildlife for Wildlife Trafficking, …)
- Data Providers
    - Unstructured News Providers, Investigative Journalists
    - Structured Data Providers (e.g., S&P for corporate data)
- AI Governance Organizations
    - Responsible AI, Ethical AI
    - E.g., Chief Digital and Artificial Intelligence Office (CDAO), Defense Innovation Unit (DIU)
- Tech Companies
    - Third party software providers who build models and platforms to help organizations manage risk.

Language

The primary language to be used for work products will be English. However, many risks have strong regional biases and certain non-English language experts will be needed to provide guidance and content to make the Risk Cards truly complete. E.g., Wildlife Trafficking has a regional focus in Africa and Asia, necessitating language development for the translation of terms and signals in those regions.

References

None, but see “Identification of Similar Work” below.

Section 2: Additional Information

Identification of Similar Work

The Risk Card standard is partially inspired by Model Cards as discussed here and in other works: https://modelcards.withgoogle.com/about

The Quantifind blog has multiple entries that introduce the notion of Risk Cards (https://www.quantifind.com/resources/a-grand-unified-approach-to-entity-risk-risk-cards-ontologies-and-knowledge-graphs/, https://www.quantifind.com/resources/risk-cards-to-make-risk-labels-standardized/) and related concepts of Responsible AI (https://www.quantifind.com/resources/putting-responsible-ai-into-practice/)

This whitepaper introduces the concept of a Joint Common Knowledge Graph, wherein open source information is fused into an entity-resolved graph, and each entity tagged with risks in accordance to a Risk Card Standard. (https://www.afcea.org/signal-media/intelligence/battling-malign-influence-open).

Quantifind has produced a large number of example Risk Cards (not yet aligned to any final standard), on a private wiki. (https://graphyte.atlassian.net/wiki/spaces/KC/pages/851247082/Risk+Card+Overview) Email john@quantifind.com for access.

First TC Meeting

The first TC Meeting will be a virtual meeting held Thursday 12/1/2022 at 2pm ET.

Ongoing Meeting Schedule

Quantifind will be responsible for scheduling virtual meetings every quarter.

TC Proposers

John Stockton
Co-founder, Quantifind
john@quantifind.com

Others TBD: MITRE, CDAO, …

Primary Representatives' Support

"I, John Stockton, john@quantifind.com, as OASIS primary representative for Quantifind, confirm our support for this proposed Charter and endorse our participants listed above as named co-proposers."

"I, [Name-of-Primary-Representative, Personal-Email-Address], as OASIS primary representative for [OASIS-Organizational-Member-Name], confirm our support for this proposed Charter and endorse our participants listed above as named co-proposers."

TC Convener

John Stockton
Co-founder, Quantifind
john@quantifind.com
626-590-8426

Anticipated Contributions

A draft specification for the Risk Card Template may be provided at the time of submission.

FAQ Document

TBD

Work Product Titles and Acronyms

The work product will be the same as the deliverables: Risk Card Template (RCT), Risk Card Example (RCE), Risk Card Guidelines (RCG), and Risk Card Registry (RCR). Additional RCE may be provided in an ongoing manner.

Appendix

Example Charter
https://docs.google.com/document/d/1uLi6e5FGAjMigR5mF5htx9wKqlteVa_CGK_Zfr5k0J8/edit#heading=h.2dlolyb