QUANTPI

On unified aspects of conformity assessment of AI, from requirements to technical implementation

Dr. Antoine Gautier | November 29, 2023

Workshop: Testing frameworks and infrastructure as baseline for trustworthy AI

Agenda

1 Introduction to QuantPi

2 Multidimensional Risk Exposure

3 Meet QuantPi’s AI Trust Platform

4 Unified Aspects of Conformity Assessment

5 Open Challenges and Q&A

Since 2020, our mission...

...has been to help organizations understand their AI-system.

We aspire to bring **transparency** into AI models and systematically **identify risks** and **hidden value** across organizations complete AI landscape.

celonis
Google
shopify
SAP
KPMG
Deloitte.
PyTorch
Yandex
ICML
SIGNAVIO
DFKI
CISPA
NEURAL INFORMATION PROCESSING SYSTEMS
EPFL
UNIVERSITÄT DES SAARLANDES
nature
Federal Ministry of Education and Research
Federal Ministry of Economics and Technology
OXFORD
Saïd Business School
elsa

Company Profile

- ➔ Working on trustworthy AI for over 7 years
- ➔ 27 employees, including 7 PhDs
- ➔ 16 languages, 12 nationalities

Various roadblocks along the AI transformation

1

Governance Need

Without a clear overview of AI risk and performance metrics, foggy decisions are made – such as the uncertainty of where to deploy AI specialists or what projects to invest in.

2

Regulatory Pressure

AI regulations, such as the EU AI Act, or internal AI guidelines, are approaching fast without having a scalable, governance framework in place to operationalize them.

3

Individual Model Risk

Lack of tools to effectively measure and validate model risk and performance across the entire lifecycle, as well as AI experts to test and enhance models.

Many risk factors inherent in all AI models now and in the future, which can result in significant financial, reputational and legal damage.

Meet the QuantPi AI Trust Platform - Removing roadblocks along your AI Transformation

2 Trust Profiles

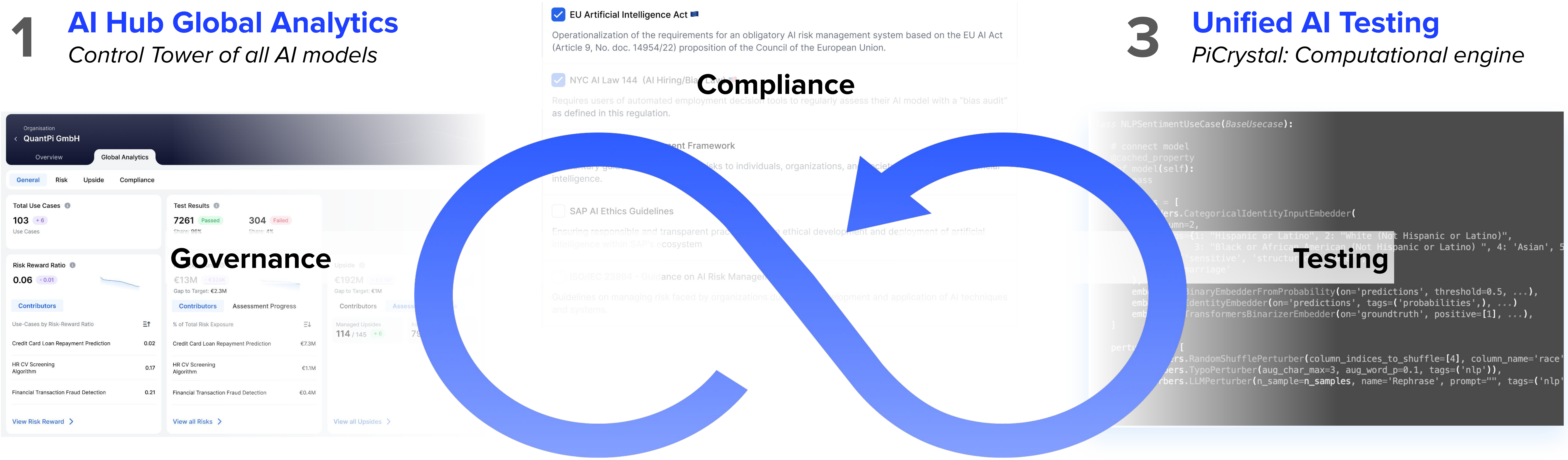
Operationalizing Requirements

1 AI Hub Global Analytics

Control Tower of all AI models

3 Unified AI Testing

PiCrystal: Computational engine



Organisation
‹ QuantPi GmbH

Overview | Global Analytics

General | Risk | Upside | Compliance

Total Use Cases ⓘ
103 +6
Use Cases

Test Results ⓘ
7261 Passed — Share: 96%
304 Failed — Share: 4%

Risk Reward Ratio ⓘ
0.06 -0.01

Contributors

Use-Cases by Risk-Reward Ratio	
Credit Card Loan Repayment Prediction	0.02
HR CV Screening Algorithm	0.17
Financial Transaction Fraud Detection	0.21

View Risk Reward ›

€13M -€324K
Gap to Target: €2.3M

Contributors | Assessment Progress

% of Total Risk Exposure	
Credit Card Loan Repayment Prediction	€7.3M
HR CV Screening Algorithm	€1.1M
Financial Transaction Fraud Detection	€0.4M

View all Risks ›

Upside ⓘ
€192M +€3.3M
Gap to Target: €1M

Contributors | Assess

Managed Upsides
114 / 145 +6

View all Upsides ›

☑ EU Artificial Intelligence Act 🇪🇺

Operationalization of the requirements for an obligatory AI risk management system based on the EU AI Act (Article 9, No. doc. 14954/22) proposition of the Council of the European Union.

☑ NYC AI Law 144 (AI Hiring/Bias Law) 🇺🇸

Requires users of automated employment decision tools to regularly assess their AI model with a "bias audit" as defined in this regulation.

ent Framework

isks to individuals, organizations, and society intelligence.

☐ SAP AI Ethics Guidelines

Ensuring responsible and transparent prac e ethical development and deployment of artificial intelligence within SAP's ecosystem

☐ ISO/IEC 23894 - Guidance on AI Risk Managem

Guidelines on managing risk faced by organizations du elopment and application of AI techniques and systems.

```
class NLPSentimentUseCase(BaseUsecase):

    # connect model
    @cached_property
    def model(self):
        pass

        s = [
        ers.CategoricalIdentityInputEmbedder(
            umn=2,
            os={1: "Hispanic or Latino", 2: "White (Not Hispanic or Latino)",
                3: "Black or African American (Not Hispanic or Latino) ", 4: 'Asian', 5
               'sensitive', 'structur
               marriage'
        ),
        emb        BinaryEmbedderFromProbability(on='predictions', threshold=0.5, ...),
        emb        IdentityEmbedder(on='predictions', tags=('probabilities',), ...)
        emb        TransformersBinarizerEmbedder(on='groundtruth', positive=[1], ...),
    ]

    pertu        [
        ers.RandomShufflePerturber(column_indices_to_shuffle=[4], column_name='race'
        bers.TypoPerturber(aug_char_max=3, aug_word_p=0.1, tags=('nlp')),
        rbers.LLMPerturber(n_sample=n_samples, name='Rephrase', prompt="", tags=('nlp'
```

Fast Time-to-Value

Fast tracks and significantly **improves AI procurement and deployment decisions** along the entire AI lifecycle.

Fast Time-to-Compliance

Scalable, auditable procedure ensures to **adhere to regulatory requirements** and internal AI compliance guidelines.

Fast Time-to-Insight

Proprietary AI Testing tool speeds up testing process, delivers **unified metrics** and allows benchmarking across AI models.

NYC Local Law 144 | Exemplary Regulation

§ 5-301 Bias Audit.

(a) An employer or employment agency may not use or continue to use an AEDT if more than one year has passed since the most recent bias audit of the AEDT.

(b) … [Compute the following according to § 5-300] …

Race/Ethnicity Categories	# of Applicants	# Selected	Selection Rate	Impact Ratio
Hispanic or Latino	408	204	50%	0.97
White (Not Hispanic or Latino)	797	412	52%	1.00
Black or African American (Not Hispanic or Latino)	390	170	44%	0.84
Native Hawaiian or Pacific Islander (Not Hispanic or Latino)	119	52	44%	0.85
Asian (Not Hispanic or Latino)	616	302	49%	0.95
Native American or Alaska Native (Not Hispanic or Latino)	41	18	44%	0.85
Two or More Races (Not Hispanic or Latino)	213	96	45%	0.87

Applicants: The number of applicants in the subgroup.
# Selected: The number of applicants in the subgroup with positive prediction.

$$\text{Selection Rate} = \frac{\text{\# Selected}}{\text{\# Applicants}} \qquad \text{Impact Ratio} = \frac{\text{Selection rate of the subgroup}}{\text{Selection rate of the most selected subgroup}}$$

Let $f : \mathcal{X} \to \{0, 1\}$ be an automated employment decision tool ("AEDT").

Suppose the applicants are partitioned into subgroups $A_1, A_2, \ldots, A_m$

$$\text{Selection Rate } A_i \approx \Pr(\, f(X) = 1 \mid X \in A_i \,)$$

$$\text{Impact Ratio of } A_i \approx \frac{\Pr(\, f(X) = 1 \mid X \in A_i \,)}{\max\limits_{j=1,\ldots,m} \Pr(\, f(X) = 1 \mid X \in A_j \,)}$$

The model is fair with respect to $A_1, A_2, \ldots, A_m$ if

Impact Ratio of $A_i = 1$ for all $i = 1, \ldots, m$

equivalently

$\Pr(\, f(X) = 1 \mid X \in A_i \,) = \Pr(\, f(X) = 1 \mid X \in A_j \,)$ for all $i, j = 1, \ldots, m$.

Equivalent metric appears under the name "Demographic parity" in

- ISO/IEC TR 24027, Section 7.5
- AIC 4, BI-02
- ...

Challenges towards a scalable, horizontal and transparent testing infrastructure for AI applications

QUANTPI

Open challenges in technical testing:

- How should the applicability and parametrization of testing algorithms be validated?
- For tested entities and use cases, which regulatory frameworks apply and how can their requirements be translated into technical tests?
- What are the concrete requirements for scalability of testing frameworks?
- ...

Open challenges in reporting on assessments:

- What are concrete, horizontal and application agnostic requirements for transparent reporting on AI system evaluations?
- How should the (numeric) results of technical assessments be visualized for different audiences (e.g. internal risk management, external auditors)?
- ...

Your contact at QuantPi

Dr. Antoine Gautier
Chief Scientist & Co-Founder

antoine.gautier@quantpi.com