UNIVERSIDAD ANDINA SIMÓN BOLÍVAR
Ecuador

Paper Universitario

An Artificial Intelligence Value at Risk Approach: Metrics and Models

AUTOR
Luis Enríquez,
docente del Área de Derecho,
Universidad Andina Simón Bolívar, Sede Ecuador

Quito, 2026

Journal of Applied Economic Sciences
Volume XX, Winter, Issue 4(90), 2025

An Artificial Intelligence Value at Risk Approach: Metrics and Models

Luis ENRIQUEZ
https://orcid.org/0000-0001-7241-8738
Law Faculty, University of Lille[1], Lille, France
Law Faculty, Universitdad Andina Simón Bolívar[2], Quito, Ecuador
luis.enriquez@univ-lille.fr luis.enriquez@uasb.edu.ec

Article's history:

Received 25th of October, 2025; Received in revised form 29th of November, 2025; Accepted 18th of December, 2025; Available online: 30th of December, 2025. Published as article in the Volume XX, Winter, Issue 4(90), December, 2025.

Suggested citation:

Enríquez, L. (2025). An Artificial Intelligence Value at Risk Approach: Metrics and Models. *Journal of Applied Economic Sciences,* Volume XX, Winter, Issue 4(90), 817–833. https://doi.org/10.57017/jaes.v20.4(90).11

Abstract:

Artificial intelligence risks are multidimensional in nature, as the same risk scenarios may have legal, operational, and financial risk dimensions. With the emergence of new AI regulations, the state of the art of artificial intelligence risk management seems to be highly immature due to upcoming AI regulations. Despite the appearance of several methodologies and generic criteria, it is rare to find guidelines with real implementation value, considering that the most important issue is customizing artificial intelligence risk metrics and risk models for specific AI risk scenarios. Furthermore, the financial departments, legal departments, and government risk compliance teams seem to remain unaware of many technical aspects of AI systems, in which data scientists, machine learning engineers, and AI engineers emerge as the most appropriate implementers. It is crucial to decompose the problem of artificial intelligence risk in several dimensions: data protection, fairness, accuracy, robustness, and information security. Consequently, the main task is developing adequate metrics and risk models that manage to reduce uncertainty for decision-making in order to take informed decisions concerning the risk management of AI systems. The purpose of this paper is to orient AI stakeholders about the depths of AI risk management. Although it is not extremely technical, it requires a basic knowledge of risk management, quantifying uncertainty, the FAIR model, machine learning, large language models and AI context engineering.

Keywords: AI-VaR; risk management; artificial intelligence; value at risk; risk-based compliance; AI Act; fairness; security.

JEL Classification: C1; C4; K2; D81.

Introduction

Artificial intelligence risk has become a very polemic field of research, considering the rise (and the hype) of generative AI and agentic AI. All this hype can be traced back to the mainstream appearance of ChatGPT in November 2022 and the AI Act's publication in July 2024 as the first generic-purpose AI regulation worldwide (Regulation EU, 2024). However, AI risk management is much older, since it appeared in the 20th century, connected with the development of predictive analytics and machine learning models. The early AI definitions of McCarthy (1956) as *the science and engineering of making intelligent machines, especially intelligent computer programs*, and Minsky's (1968) AI definition as *the science of making machines do things that would require intelligence if done by men,* anticipated AI as a scientific area of research. In such a sense, machine learning models became the training methodologies while using datasets as the input.

[1] 42 rue Paul Duez 59000 Lille − France.

[2] Toledo N22-80 (Plaza Brasilia), Quito – Ecuador.

The main purpose of traditional supervised models has been solving classification and regression problems in order to provide a response or a prediction. Later on, deep learning became a subset of machine learning, with the purpose of using neural networks with several layers and more expensive processing, that could provide a more robust problem-solving approach. The evolution of deep learning, natural language processing, and especially a model architecture based on self-supervised learning and transformers (Vaswani, et al. 2017) became the fundamentals of today's large language models and large reasoning models. Yet, all traditional machine learning models are still useful for tasks such as data preparation, fine-tuning, model evaluation, risk calibration, and as additional resources to train predictive systems in a retrieval-augmented generation environment.

Consequently, AI strongly relies on decision-making, whether it is automated or used as an assistant for human decision-making. This means that risk management has always been present in the early stages of AI development, since the goal of risk management is reducing uncertainty for taking informed decisions (Freund & Jones, 2014). Uncertainty can be classified as aleatoric or epistemic. Aleatoric uncertainty *is caused by inherent randomness and unpredictability in a system* (Manokhim, 2023). Epistemic uncertainty *arises from the lack of knowledge or understanding about a system* (Manokhim, 2023). From this perspective, data scientists have always been doing risk management while cleaning up their datasets and while implementing metrics with the aim of improving an acceptable model performance. Consequently, quantifying uncertainty is not unusual for data scientists, and they currently use risk calibration strategies based on conformal prediction, ensemble methods, Bayesian methods, or direct interval estimation for such tasks (Dewolf et al., 2021).

The paradox is that they were rarely trained in risk management guides coming from best practice standards such as ISO, or Governance Risk Compliance procedures. Data scientists just did it. There are many new AI risk management frameworks, such as the ISO/IEC 23894, ISO/IEC 42001, ISO/IEC 42005, the NIST AI RMF 1.0, the NISTIR 83-30, the Cap AI, and so forth. While all of them are useful in different ways, their approach comes from other areas such as cybersecurity, project management, and legal academics. It feels more like foreigners designing rules and procedures for local natives that actually do the risk calibration tasks, where the local natives are the data scientists and the AI engineers.

1. Literature Review

In our days, risk management is a poorly understood area due to a loss of focus on what risk management is about. For Sidorenko (2017), *risk management was born as science*, then became an art, and today is just bullsh@t (2025). Hubbard (2016) classified risk managers into the four horsemen of risk management: the first three are actuaries, the war quants, and the economists, linked with applied scientific procedures to reduce uncertainty. Yet, the fourth group consists of consulting managers that have done *more bad than good*, as it has distorted what risk management is about, turning it into checklists and paper-based compliance. The problem is not the alleged best practice standards, as they only provide criteria from a project implementation perspective. Yet, they don't provide data, significant metrics, and accurate models for informative risk management (Freund & Jones 2014).

The real issue is that non-trained risk managers are turning AI risk management into a checklist placebo and, even worse, putting compulsory obligations to apply non-effective qualitative methods, such as heat maps and symmetric risk matrices (Cox, 2008), to an environment that has natively used quantitative metrics for training, calibrating, and testing AI models. Just like risk management was born with the actuaries more than 200 years ago for fulfilling a role that did not exist (SOA, 2008), today we need a new risk manager role for AI, a sort of AI risk manager that actually knows how to train, calibrate, and test AI systems using applied science, with the aim of complying with AI legal regulations. Furthermore, this new role shall also develop operational and legal risk scenarios that require their own risk modelling approach.

This paper tackles several fundamental dimensions of AI risk: data protection, fairness, accuracy, robustness, and information security. The methodology used is based on quantitative risk management, consisting on the following phases:

- Firstly, gathering relevant data. In this paper, the administrative fines collected data correspond to real cases, but the personal data is synthetic.
- Secondly, developing metrics for risk identification, as a need for establishing trustworthy risk rationales. For such task, several methods are explained such as conformal prediction, the PERT formula, and the Monte Carlo analysis.
- Thirdly, designing accurate risk models for artificial intelligence risk scenarios. They are based on the FAIR model.
- Quarterly, forecasting a return on security investment for risk controls.

2. Modelling Legal Risk: Personal Data Protection and Fairness

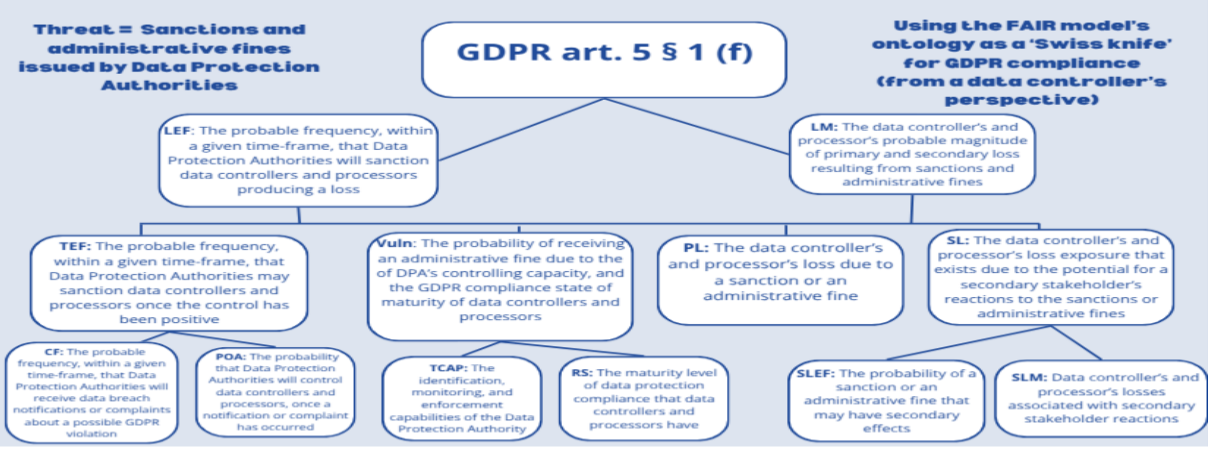
Legal risk has two clear perspectives: the risk of violating the fundamental rights of natural persons and the risk of AI deployers, AI providers, and other stakeholders, to comply with relevant laws, such as the AI Act (EU, 2024). Both risk assessments shall be developed in parallel, as they are interdependent. Recent risk-based legal frameworks, such as the General Data Protection Regulation (EU, 2016), have made popular the impact assessment perspective for the protection of the rights and freedoms of natural persons. In the IA Act, the equivalent of the Data Protection Impact Assessment is the Fundamental Rights Impact Assessment. However, an impact assessment is essentially a risk assessment due to two arguments: firstly, the word *assessment* (ISO, 2023) concerns the identification, analysis, and evaluation of risks against the fundamental rights of natural persons. Secondly, the impact is a dimension of risk, that gets completed with the probability of occurrence.

2.1. Personal Data Protection

The first risk assessment shall be related to personal data. Datasets are the training input of an AI system, where the quality and reliability of the data are essential. Therefore, a logical first AI risk assessment shall be a Data Protection Impact Assessment concerning two important areas: identifying personal data as an input of the processing and identifying personal data as an output of the processing. On one hand, personal data shall be identified when it is used as training input for an AI system. When personal data is necessary for training an AI system, the audit shall be focused on the legal basis for the data processing. This means, *do I have explicit consent? Do I have clear processing objectives? Do I comply with the exercise of the data subjects' rights?* On the other hand, personal data shall be leaked as the response of an AI system, especially when AI systems are widely available through a prompting system. This obligation may be changed soon with the *digital omnibus* (European Commission, 2025), if legislators justify the use of personal data for training AI systems as legitimate interest by default.

Unfortunately, the state of the art of DPIAs is very immature, and we have the risk that FRIAs become as such, too. The DPIA comes from the Privacy Impact Assessment, a kind of assessment that comes from the 70s that consists of a description of the data processing activities and risk assessment. However, they mostly became a checklist placebo far away from an applied-scientific approach. Several authors point out the need for updating them. Quantifying privacy is still a new field of research, where few authors have modelled it, such as the FAIR-P model (Cronk & Shapiro, 2022), and a Personal Data Value at Risk approach (Enríquez, 2024). Considering the difficulty of measuring the material impact that natural persons may have on their rights and freedoms, the alternative is decoding the sanctioning psychology of data protection authorities, because in the end, they will have to estimate the impact that natural persons have suffered. This means that a DPIA is important for AI risk assessment for all that concerns data protection. Data protection risk modelling has been deeply analysed and modelled in previous works (Enríquez, 2024). Yet, a useful risk ontology for the data protection risk of AI systems shall be presents in the following Figure 1.

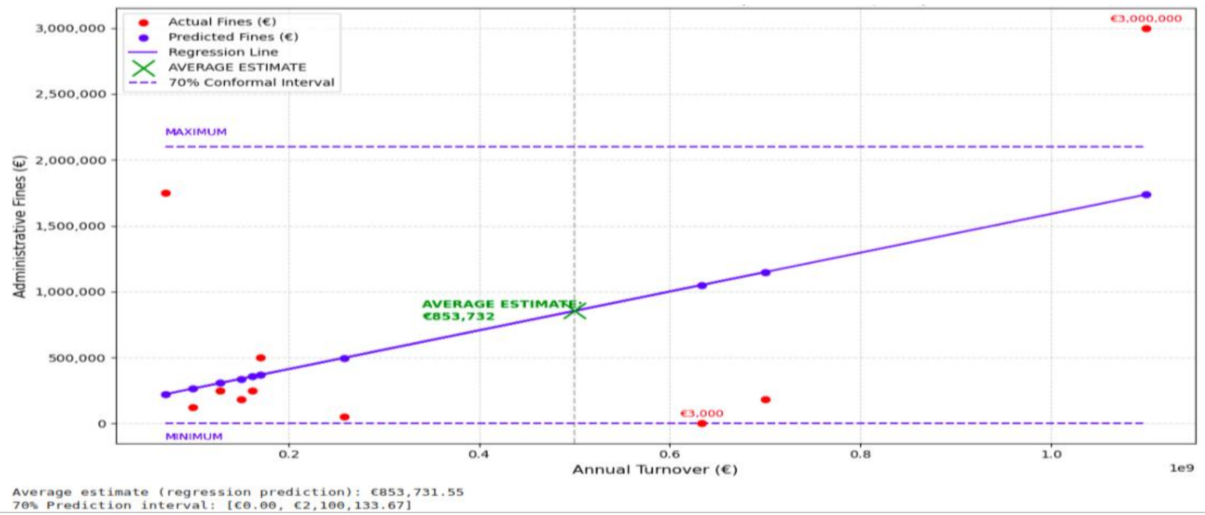
Figure 1: A risk model ontology for obtaining the Annual Expected Loss from administrative fines


Threat = Sanctions and administrative fines issued by Data Protection Authorities
GDPR art. 5 § 1 (f)
Using the FAIR model's ontology as a 'Swiss knife' for GDPR compliance (from a data controller's perspective)
LEF: The probable frequency, within a given time-frame, that Data Protection Authorities will sanction data controllers and processors producing a loss
LM: The data controller's and processor's probable magnitude of primary and secondary loss resulting from sanctions and administrative fines
TEF: The probable frequency, within a given time-frame, that Data Protection Authorities may sanction data controllers and processors once the control has been positive
Vuln: The probability of receiving an administrative fine due to the of DPA's controlling capacity, and the GDPR compliance state of maturity of data controllers and processors
PL: The data controller's and processor's loss due to a sanction or an administrative fine
SL: The data controller's and processor's loss exposure that exists due to the potential for a secondary stakeholder's reactions to the sanctions or administrative fines
CF: The probable frequency, within a given time-frame, that Data Protection Authorities will receive data breach notifications or complaints about a possible GDPR violation
POA: The probability that Data Protection Authorities will control data controllers and processors, once a notification or complaint has occurred
TCAP: The identification, monitoring, and enforcement capabilities of the Data Protection Authority
RS: The maturity level of data protection compliance that data controllers and processors have
SLEF: The probability of a sanction or an administrative fine that may have secondary effects
SLM: Data controller's and processor's losses associated with secondary stakeholder reactions

Source: Author

For instance, in a *sensitive information disclosure* risk scenario (OWASP, 2023), the annual risk prediction interval may be calculated using a top-down approach. However, in order to avoid guessing values, an informative rationale may rely on a historic analysis of existing administrative fines. Transductive Conformal prediction may be used to obtain a prediction interval, with a minimum value, and a maximum value. These values have been forecasted by analysing data security administrative fines in France, between 2018 and 2024 for a sensitive information disclosure risk scenario. The prediction interval has been set at the 70% of coverage:

Figure 2: Transductive conformal prediction with a 70% prediction interval, in Euro



```
Average estimate (regression prediction): €853,731.55
70% Prediction interval: [€0.00, €2,100,133.67]
```

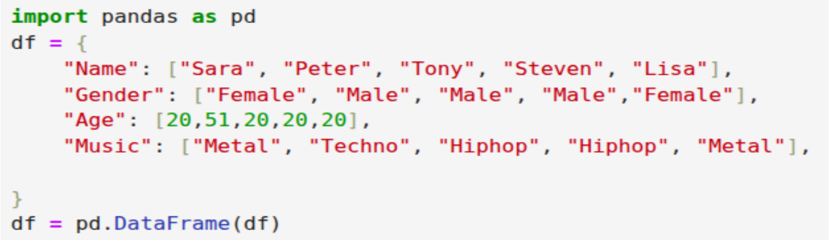
Source: Author

The above Figure 2 shows a correlation between the annual turnover of an enterprise and the administrative fines received, obtaining a confidence level of 70% based on historical analysis. From all the AI risk dimensions proposed in this paper, personal data protection is still the only one with real case scenarios, and the application of a Personal Data Value at Risk approach is convenient. Furthermore, it is also possible to apply a Return on Risk Investment logic (Albina, 2021), detailed in the following cases.

Personal data as an input

A typical legal risk scenario is the lawfulness of processing of the data, which consists of assessing the legality of the personal data used for training. Assessing AI personal data protection risks as an input requires identifying the personal data included in datasets and then analysing the probability of a regulatory violation and the magnitude of the impact on the data subjects if the risk materializes. The following example shows a typical dataset with personal data:

Figure 3: A small dataset with five natural persons



```
import pandas as pd
df = {
    "Name": ["Sara", "Peter", "Tony", "Steven", "Lisa"],
    "Gender": ["Female", "Male", "Male", "Male","Female"],
    "Age": [20,51,20,20,20],
    "Music": ["Metal", "Techno", "Hiphop", "Hiphop", "Metal"],

}
df = pd.DataFrame(df)
```

Source: Author

Considering a constant ponderation between privacy and utility, it is highly recommended to evaluate several pseudonymization methods concerning a particular model (Yermilov et al., 2023). Nevertheless, if the names are removed, it is necessary to measure the attributes of the identity of a person concerning a particular sample space (Pierangela & Sweeney, 1998). For instance, if our risk acceptance criteria are a 70% probability of being identified, and the leaked attributes are the favourite music and the gender, we can proceed to quantify the Return on Security Investment (ROSI) of this control. Sara and Lisa would have a 60% probability of being identified by her age and favourite music, therefore, passing the acceptance test. Tony and Steven will also pass the test, as due to their favourite music attribute (hip-hop), they will have a 60% probability of identification. Nevertheless, Peter will have a 100% probability of being identified by his music preference (techno). Consequently, if the risk exposure of this risk scenario is €100 000, the implementation of the control costs €10 000, and the control works for 4 of the 5 five persons, the efficacy expectancy is 80% (Table 1). Then, the ROSI formula can be applied:

Table 1: The ROSI of a differential privacy control

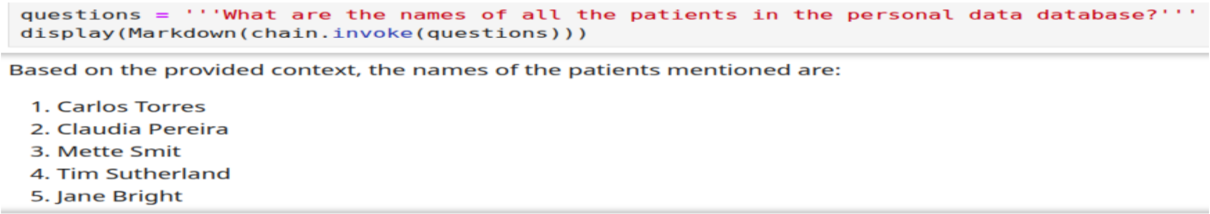
Risk exposure	€100 000
Efficacy	80% (efficacy expectancy = risk solution)
Cost of the security investment	€10 000
ROSI	((Risk exposure * efficacy) - Cost of the risk control) / Cost of the risk control = (80 000 – 10 000) / 10 000 = 7 (700%)

Personal data as an output

It is a very tough task because the response of a prompt-based AI responsive system is not as explainable as many peers think. When using open-source and open-weight LLMs, sensitive information leaks may become a third-party risk. However, it can be reduced with privacy-enhancing technologies. Whether the leak comes from the interaction between the user and the chatbot in Gen AI or from the interaction of AI agents in Agentic AI, the risk of these kinds of data leaks may be considerably reduced from a post-processing perspective (Krishnan, 2025). The response of an LLM may be filtered by implementing a post-processing masking or pseudonymization-oriented code that connects to the AI model from a confidential AI environment.

Furthermore, personal data may still exist in the RAG and vector databases, but it becomes dependent on access control policies in the output. The first graphic shows a leak of personal data due to the lack of access control, and the second one shows an effective control of the data leaked, where the data leaked is already protected with a masking function.

Figure 4: A prompt-based request for personal data

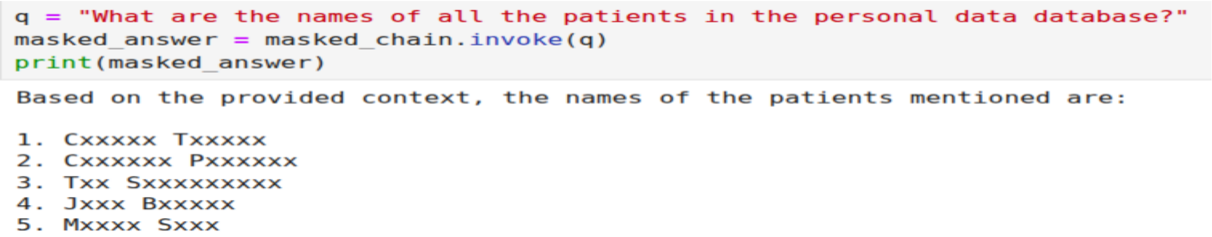


```
questions = '''What are the names of all the patients in the personal data database?'''
display(Markdown(chain.invoke(questions)))
```

```
Based on the provided context, the names of the patients mentioned are:

  1. Carlos Torres
  2. Claudia Pereira
  3. Mette Smit
  4. Tim Sutherland
  5. Jane Bright
```

Source: Author

Figure 5: A prompt-based request with filtered pipelines



```
q = "What are the names of all the patients in the personal data database?"
masked_answer = masked_chain.invoke(q)
print(masked_answer)
```

```
Based on the provided context, the names of the patients mentioned are:

1. Cxxxxx Txxxxx
2. Cxxxxxx Pxxxxxx
3. Txx Sxxxxxxxxx
4. Jxxx Bxxxxx
5. Mxxxx Sxxx
```

Source: Author

For instance, let's consider that in a sensitive information disclosure risk scenario, the mean value in France between data security non-conformities to the GDPR in a given turnover interval is €853,731, as shown in Figure 2. As the risk control only masks personal data in Spanish language, and the percentage of Spanish named clients is 80% (Table 2), the ROSI works as follows:

Table 2: ROSI of a masking data for prompting control

Risk exposure	€853,731
Efficacy	80% (efficacy expectancy = risk solution)
Cost of the security investment	€100,000
ROSI	((Risk exposure * efficacy) - Cost of the risk control) / Cost of the risk control = (682 984 – 100 000) / 100 000 = 5,82 (582%)

2.2. Fairness

When we consider impact assessments as a whole risk management procedure, the tasks get more efficient. Fairness risk management is a must for decision-making based on AI, whether it is used as an algorithmic decision-making procedure or as assistance for human decision-making. Likewise, risk calibration becomes the holy grail of risk management, and for such a task, it is compulsory to understand how to reduce bias and noise in risk management. Bias is prejudice in favour of someone or something, while noise is about inaccuracy in the decision estimation (Kahneman et al., 2021). Both circumstances can contaminate the accuracy of a decision-making process, and therefore, they need to be handled during all the risk management phases: context establishment, risk identification, risk analysis, risk evaluation, and risk treatment (ISO 27005). Yet, fairness is deeply linked with bias, and there are several well-known metrics that may help to identify such conditions. An important fact to consider is that what new AI legal regulations name as algorithm bias, is mostly related to a bias that comes inherently in the datasets, as datasets may only be a mirror of society's inequalities.

Fairness metrics

Most fairness risk scenarios are linked to the right of non-discrimination. However, there usually exists a default bias that may come by default in the datasets. For instance, let's consider a human resources department that needs to hire employees (Table 3). The following dataset includes twenty candidates that have a score and may be used to train the selection system. The numbers '0' and '1' represent the most disfavoured group of natural persons, while the number '2' represents the most favoured one.

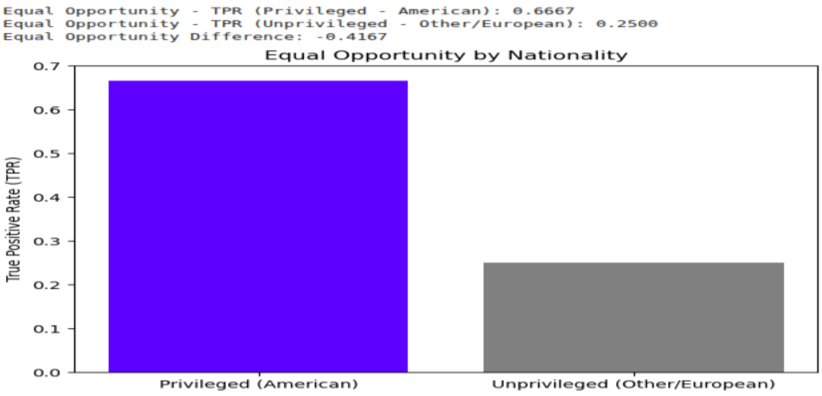
Table 3: A dataset with 20 natural persons with features and a score

No.	Name	Gender	Age_Group	Nationality	Score
0	Alice	1	2	1	7.56
1	Riley	2	0	0	6.31
2	Jordan	0	2	2	6.47
3	Sophie	1	0	1	5.63
4	Liam	2	2	2	9.82
5	Taylor	1	2	0	6.73
6	Ethan	2	2	2	9.77
7	James	2	0	2	7.23
8	Morgan	0	1	1	5.14
9	Skyler	0	0	0	3.91
10	Henry	1	1	0	4.88
11	Emma	2	2	1	9.01
12	Noah	1	1	2	7.36
13	Owen	2	2	2	9.94
14	Ella	1	0	0	4.19
15	Grace	0	0	0	3.72
16	Jack	0	2	2	6.91
17	Mia	1	2	1	7.58
18	Lucas	2	2	2	9.98
19	Isla	0	2	1	6.45

Source: Author

From the job applicant's perspective, the risk scenario is not getting the job due to probable discrimination due to gender, nationality, and age features. The threat community is the human resources department that may use biased data to train their AI decision-making systems. The vulnerability is the indefensible informative position of the applicants due to a non-transparent and biased system. The consequence is not getting a job even though they deserve it. Firstly, we shall apply fairness metrics in order to unveil bias in the dataset. There are several metrics for fairness; nonetheless, here we will use the *equal opportunity difference metric* for the nationality feature. The equal opportunity difference formula is the result of the subtraction between the true positive rates of the unprivileged group and the true positive rates of the privileged group. The result is -4,167, meaning that the unprivileged group are 41.67% less likely to receive a positive outcome when they actually qualify.

Figure 7: Equal Opportunity Difference Metrics for nationality biases



```
Equal Opportunity - TPR (Privileged - American): 0.6667
Equal Opportunity - TPR (Unprivileged - Other/European): 0.2500
Equal Opportunity Difference: -0.4167
```

Source: Author

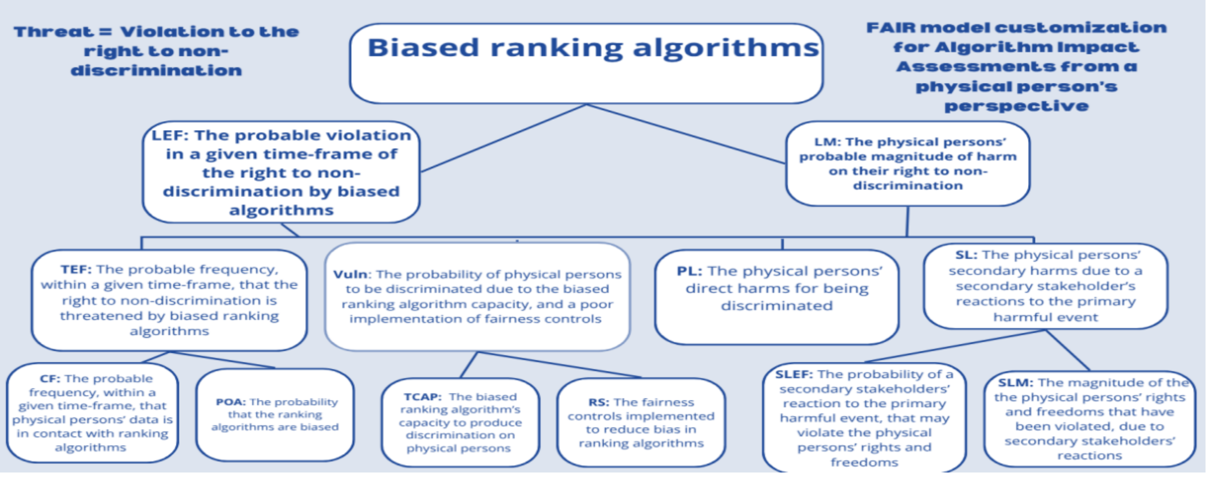
Once biases like this one have been detected, the next step is selecting and implementing a risk control that calibrates the applicants' vulnerabilities towards a non-discriminative system. Their current disadvantages can be recalibrated within a fairness risk model scenario by calibrating the vulnerabilities of the discriminated groups and increasing their options to get hired. For instance, it is feasible to calibrate the nationality by adding the equal opportunity difference to the unprivileged group.

Table 4: Percentages of natural persons' nationality within the dataset

	EOD - TPR	Equal opportunity difference	Final Vulnerability score
Privileged =	66%	0	100% - 66% = 34%
Non-privileged =	25%	41%	100% - (25% + 41%) = 34%

In the previous example, Americans are clearly benefited. Nonetheless, the vulnerability score of both groups is equal. The ROSI can also be calculated, just like it was already shown in the personal data risk dimension. Whether it is about inherent risk analysis or residual risk analysis, a good option is developing a risk ontology (just like the classic FAIR model ontology) within a *biased ranking algorithms* risk scenario, where the vulnerability factor can be calibrated in order to benefit the conditions of specific disadvantaged groups of natural persons.

Figure 8: A risk scenario of biased ranking algorithms from an AI users' perspective



Threat = Violation to the right to non-discrimination

Biased ranking algorithms

FAIR model customization for Algorithm Impact Assessments from a physical person's perspective

LEF: The probable violation in a given time-frame of the right to non-discrimination by biased algorithms

LM: The physical persons' probable magnitude of harm on their right to non-discrimination

TEF: The probable frequency, within a given time-frame, that the right to non-discrimination is threatened by biased ranking algorithms

Vuln: The probability of physical persons to be discriminated due to the biased ranking algorithm capacity, and a poor implementation of fairness controls

PL: The physical persons' direct harms for being discriminated

SL: The physical persons' secondary harms due to a secondary stakeholder's reactions to the primary harmful event

CF: The probable frequency, within a given time-frame, that physical persons' data is in contact with ranking algorithms

POA: The probability that the ranking algorithms are biased

TCAP: The biased ranking algorithm's capacity to produce discrimination on physical persons

RS: The fairness controls implemented to reduce bias in ranking algorithms

SLEF: The probability of a secondary stakeholders' reaction to the primary harmful event, that may violate the physical persons' rights and freedoms

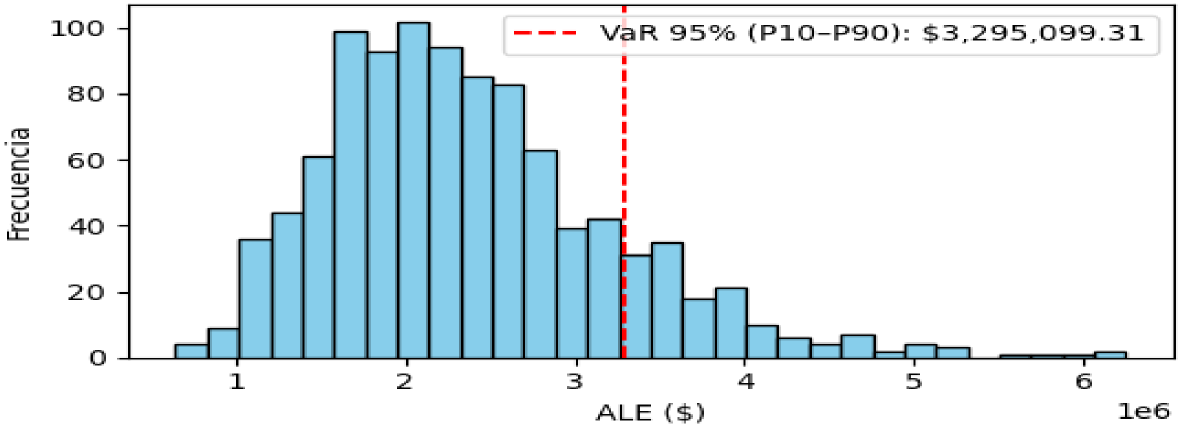
SLM: The magnitude of the the physical persons' rights and freedoms that have been violated, due to secondary stakeholders' reactions

Source: Author

Here we have to calibrate two different percentages of vulnerabilities, one for Americans and another one for non-Americans. From here, we could apply different procedures to get the F-VaR such as the PERT formula: a+4b+c/6 (Mirković, 2020), or using a Monte Carlo simulation (Freund & Jones, 2014), with the goal of getting an accurate prediction interval. Quantifying fairness is somehow a new domain, but recently research has shown the link with fairness and the financial domain, such as the use of cooperative game theory and SHAP analysis methods (Chan et al., 2025). Nonetheless, constructing the rationale of fairness is easier by profiling the authorities' sanctioning psychology (Lawlor, 1963).

Since there are no administrative fines' precedents in the fairness risk domain, conformal prediction would not be an option. A compliance-focused model ontology from an AI provider's and AI deployer's perspective may be based on a Fairness Value at Risk (F-VaR), as National Competent AI Authorities will have to estimate the amount of damage to the fundamental rights of natural persons in each fairness violation (Enríquez, 2024). For instance, in order to implement a Fundamental Rights Impact assessment (FRIA) on fairness risk scenarios, it is possible to analyse the previous nationality discrimination risk scenario. The values may be: TEF = min: 2, most likely: 5, max: 8. The vulnerability of non-privileged groups may be set at Vuln: min: 40%, most likely: 66%, max: 90%. The PL = min: $100,000, most likely: $200,000, max: $400,000. The SL = min: $500,000, most likely: $1,000,000, max: $2,000,000. Figure 8 shows the annual fairness value at risk of this nationality scenario, with the 10th percentile as the lower limit and the 90th percentile as the upper limit, representing an 80% confidence interval. Within this interval, a value corresponding to the 95th percentile can be used to express a high-confidence estimate of potential loss.

Figure 8: An annual loss expectancy and its artificial intelligence value at risk in the fairness dimension
Source: Author



4. Modelling Operational Risk: Accuracy and Robustness

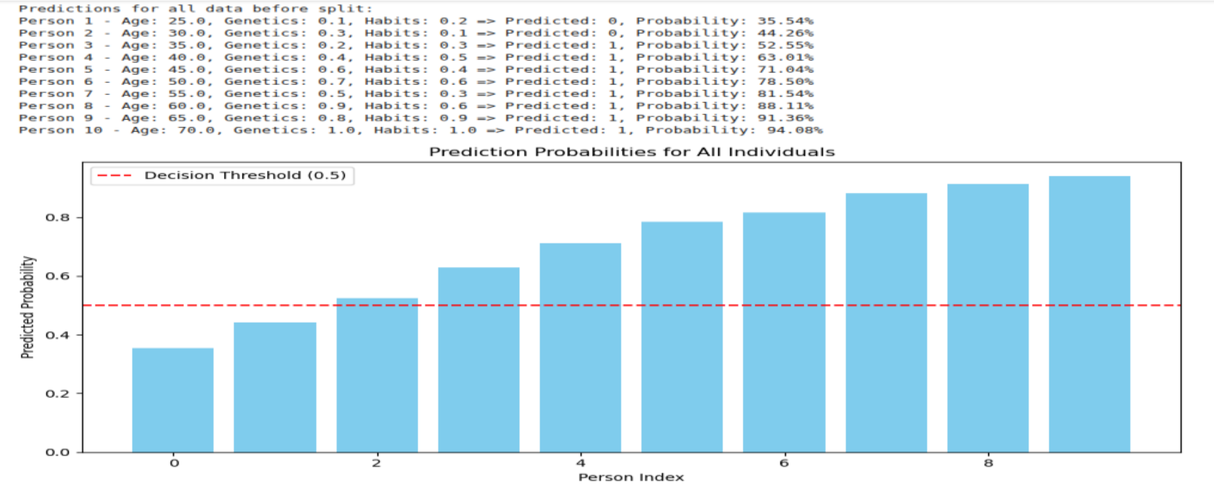
In the context of the AI Act, accuracy and robustness are defined in the AI Act as *relevant performance metrics* (EU 2024). Accuracy is understood as the acceptable prediction intervals of an AI system. Robustness is understood as safety performance measures. As performance metrics, they are transversal for many risk scenarios. Accuracy and robustness error metrics are crucial in order *to measure AI prediction capabilities* (Floridi et al., 2022). There are several metrics used in classification, multi-classification, and regression problems. Classification metrics are usually based on the number of true positives, false positives, true negatives, and false negatives.

For instance, well-known metrics for classification are: accuracy, precision, recall, faithfulness, and so on. Common regression metrics are Root Mean Squared Error (RMSE) Mean Absolute Error (MAE), and so on. While LLMs use next-token prediction metrics (Ouyang, et al., 2022), LLMs are trained with classification-style objectives, and regression losses are useful in fine-tuning or auxiliary tasks. This means that classification and regression metrics are fundamental for simple predictive models and complex AI systems. For instance, let's consider a data poisoning risk scenario that can lead to AI backdoor errors and increase the risk of hallucinations. The threat community could be a group of cybercriminals with access to the AI system who can poison the training data of patients with a probability of cancer. The vulnerability could be the lack of security access controls of the IA deployer. The consequence of the system's AI users may be the probable wrong answers that produce financial and psychological impact on potential cancer patients. The consequence of the AI provider and the AI provider may be the financial losses.

4.1. Classification

The following synthetic dataset includes ten patients, where the features are *age*, *genetics*, *habits,* and *has cancer* as the classification response, where the risk acceptance threshold is 50%. This means that patients below the threshold are labelled with '0' or no cancer, and patients with more than 50% are labelled as patients with a considerable probability of getting cancer.

Figure 9: The predictions of 10 natural persons with 3 features and the probability of getting cancer

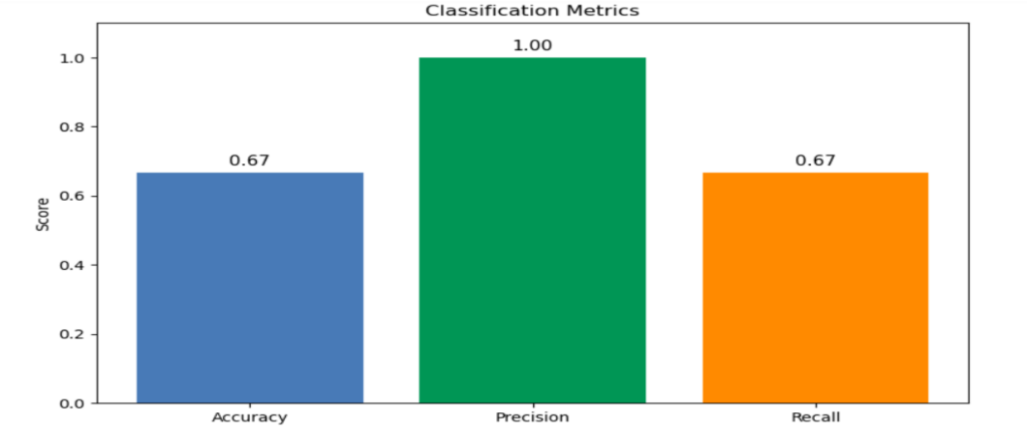


```
Predictions for all data before split:
Person 1 - Age: 25.0, Genetics: 0.1, Habits: 0.2 => Predicted: 0, Probability: 35.54%
Person 2 - Age: 30.0, Genetics: 0.3, Habits: 0.1 => Predicted: 0, Probability: 44.26%
Person 3 - Age: 35.0, Genetics: 0.2, Habits: 0.3 => Predicted: 1, Probability: 52.55%
Person 4 - Age: 40.0, Genetics: 0.4, Habits: 0.5 => Predicted: 1, Probability: 63.01%
Person 5 - Age: 45.0, Genetics: 0.6, Habits: 0.4 => Predicted: 1, Probability: 71.04%
Person 6 - Age: 50.0, Genetics: 0.7, Habits: 0.6 => Predicted: 1, Probability: 78.50%
Person 7 - Age: 55.0, Genetics: 0.5, Habits: 0.3 => Predicted: 1, Probability: 81.54%
Person 8 - Age: 60.0, Genetics: 0.9, Habits: 0.6 => Predicted: 1, Probability: 88.11%
Person 9 - Age: 65.0, Genetics: 0.8, Habits: 0.9 => Predicted: 1, Probability: 91.36%
Person 10 - Age: 70.0, Genetics: 1.0, Habits: 1.0 => Predicted: 1, Probability: 94.08%
```

Source: Author

The logistic regression model can provide the probability of each patient, despite the fact of being classified in two groups. The following Figure 10 shows the results of accuracy, precision, and recall metrics:

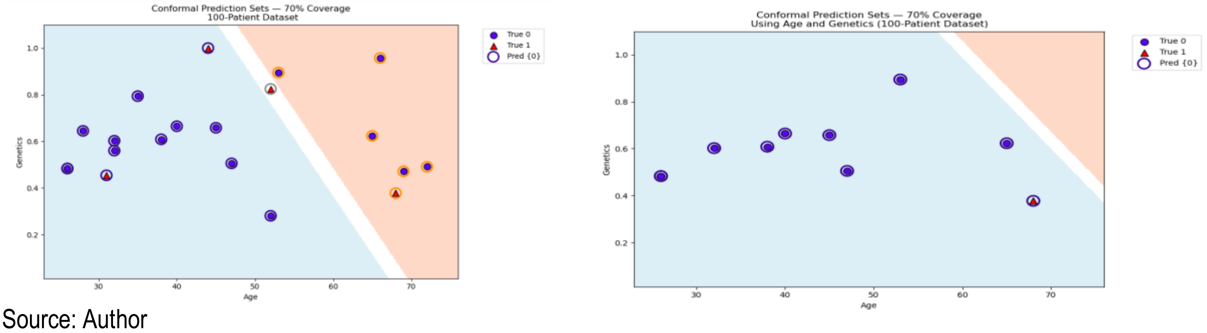
Figure 10: Accuracy, precision, and recall metrics



Source: Author

In fact, this is a very conservative dataset, with only real cases, where the 3 testing observations are two true positives, and one false positive. Nevertheless, this is not enough, as risk management shall pursue accuracy and not precision. The most important issue is to implement conformity metrics with the aim of getting a trustworthy prediction interval for new cases. Again, conformal prediction becomes a very useful method to obtain an accurate prediction interval based on historical analysis of patients that had cancer. The first figure shows a 100-patient dataset with an accuracy of 0.61 with a distributed split of 60% for training, 20% for calibration, and 20% for testing. The second figure shows the same 100 patient dataset but with 60% for training, 30% for calibration, and 10% for testing. Increasing the calibration data split is a risk control that has helped to increase the accuracy.

Figure 11: Comparing accuracy after a calibration risk control



Source: Author

With such results it can be assumed that the ROSI will increase as accuracy has increased in a 29%. However, a more informative result would be understanding the accuracy of other features (such as Genetics with habits), and implementing other metrics.

4.2. Regression

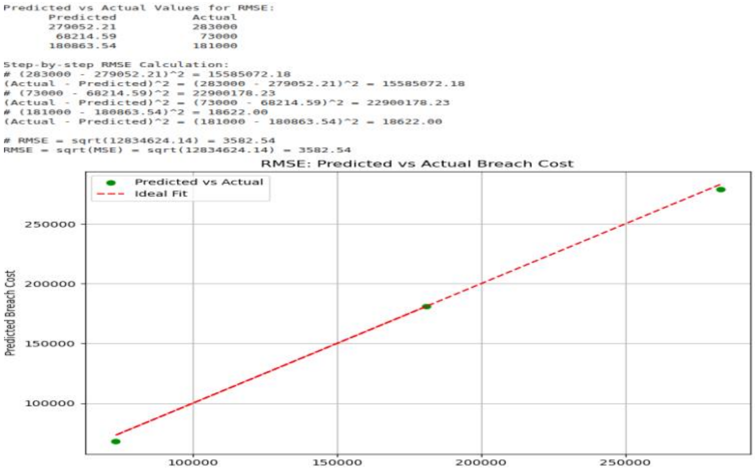
In an adversarial machine learning risk scenario of data poisoning, implementing a linear regression model with the RMSE metrics can provide good information about the predicted breach cost vs. the actual breach cost. The following historical dataset includes the average of data poisoning losses in the previous year. For instance, in a dataset of ten observations (Table 5), seven have been used for training and three for testing.

Table 5: Cost components associated with security breaches

Index	Productivity Loss	Reputation Damage	Administrative Fine	Breach Cost
0	20,000	10,000	4,000	52,000
1	22,000	12,000	4,500	73,000
2	24,000	9,000	5,000	76,000
3	28,000	20,000	8,000	119,000
4	30,000	25,000	8,500	151,000
5	33,000	30,000	9,000	181,000
6	35,000	28,000	9,500	203,000
7	40,000	35,000	10,000	237,000
8	45,000	40,000	11,000	283,000
9	50,000	45,000	12,000	318,000

The difference between the actual and the predicted results can be measured with the residuals (Figure 12). In this case, they are acceptable.

Figure 12: Comparison between actual and predicted administrative fines

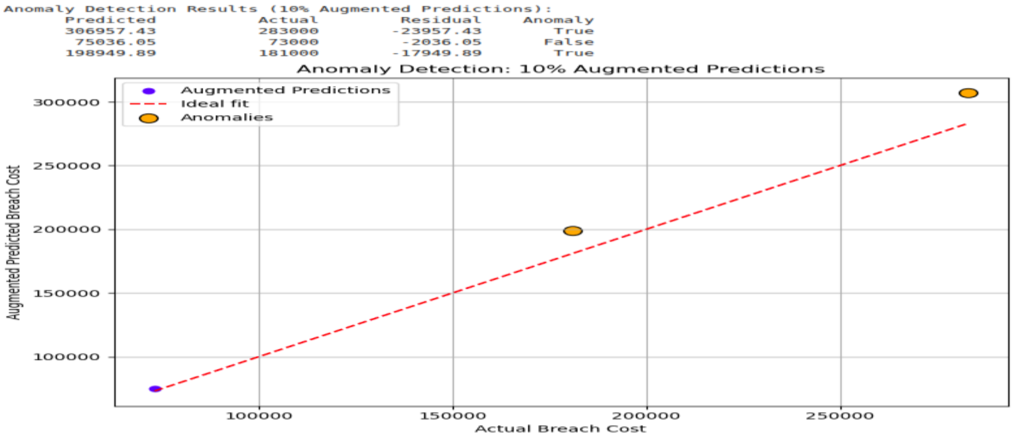


```
Predicted vs Actual Values for RMSE:
      Predicted            Actual
      279052.21            283000
       68214.59             73000
      180863.54            181000

Step-by-step RMSE Calculation:
# (283000 - 279052.21)^2 = 15585072.18
(Actual - Predicted)^2 = (283000 - 279052.21)^2 = 15585072.18
# (73000 - 68214.59)^2 = 22900178.23
(Actual - Predicted)^2 = (73000 - 68214.59)^2 = 22900178.23
# (181000 - 180863.54)^2 = 18622.00
(Actual - Predicted)^2 = (181000 - 180863.54)^2 = 18622.00

# RMSE = sqrt(12834624.14) = 3582.54
RMSE = sqrt(MSE) = sqrt(12834624.14) = 3582.54
```

Source: Author

However, in a data poisoning risk scenario, the predicted values can be manipulated as a result of calculating over tampered data. A convenient robustness risk control may be an anomaly detection system. Just like accuracy, robustness controls are also important when the risk scenarios are linked with consumer protection, civil law, or even criminal law. The following example shows the tampered results of a linear regression model after a data poisoning attack that arbitrarily modified 10% of the predicted values and how an anomaly detection system risk control with a threshold of 4,000 can detect the predicted values that have exceeded it.

Figure 13: Comparison between actual and predicted administrative fines



```
Anomaly Detection Results (10% Augmented Predictions):
      Predicted            Actual          Residual     Anomaly
      306957.43            283000         -23957.43        True
       75036.05             73000          -2036.05       False
      198949.89            181000         -17949.89        True
```

Source: Author

The risk model would be similar to the ones already presented in the personal data and fairness risk ontologies. Likewise, the ROSI can also be estimated as it was previously shown. Nonetheless, they could also be used only as resistance strength. The logic behind this assumption is logical and pragmatic; if the model has better accuracy and is robust enough, it becomes a very important resistance strength control.

5. Modelling Information Security: Integrating Personal Data Protection, Fairness, Accuracy and Robustness

Once we get the output of data protection, fairness, accuracy, and robustness assessments, it is very convenient to integrate them into a single operational information security risk model. The FAIR model is ideal for information security risk scenarios related to artificial intelligence, such as AI hallucinations (Maleki et al., 2024) and adversarial machine learning (McCarthy et al., 2023). The trick is integrating personal data protection, accuracy, robustness, and fairness in the right factors, while decomposing the risk scenario problem.

Personal data

Its Personal data Value at Risk shall be calculated in its own risk ontology. The rationale of a personal data value at risk is based on an intelligence assessment about the sanctioning psychology of the data protection authority, and the maturity level of data protection compliance (such as GDPR compliance). The logic of the Jurimetrical Pd-VaR (Enríquez, 2024) relies on a historical analysis of the judges or administrative authorities in order to solve the difficulty of calibrating the material impact of a data protection risk on the rights and freedoms of the data subjects. Instead of blindly guessing such impact, a very useful alternative is understanding the sanctioning psychology of the Data Protection Authority by using qualitative and quantitative information and argument retrieval analytical methods and adding the results as a secondary loss within an information security risk scenario linked to AI. The jurimetrical Pd-VaR is then complemented with the calibrated Pd-VaR, which is about the internal risk-based compliance maturity of the controller.

Fairness

It can also be calculated in an independent model ontology, and the output of the loss can be integrated as a secondary loss within the FAIR model. In the near future, we may also get fairness-related data coming from existing administrative fines issued by National AI authorities. Quantitative information and argument retrieval analytical methods may also be applied in order to calibrate a Fairness Value at Risk (F-VaR). The logic behind it relies on the difficulty of calibrating the loss of AI users, following the logic behind a Personal Data Value at Risk previously explained.

Accuracy and robustness

Even though that could also become part of the secondary risks related to fines and judgments, in the following case they will be only part of the resistance strength factor. They are integrated as resilience strength because their purpose is the good function and redundancy of the system that helps mitigate the risk of harm against the fundamental rights of natural persons. These factors shall be weighed, combining them with the resilience level of the risk controls that shall provide protection from a threat community trying to compromise the system. In a nutshell, a bad accuracy and robustness assessment within the model would result in harm to the AI users, just like a black hat hacker threat community may exploit vulnerabilities of confidentiality, integrity, and availability controls. Therefore, in consumer protection legal risk scenarios, they may also become part of the *fines and judgments'* secondary losses factor.

Information security

It may be convenient in many risk scenarios to combine the data protection, fairness, accuracy, and robustness rationales into an information security model. For instance, let's consider data poisoning combined with a biased algorithm ranking risk scenario. The protected asset is confidential data in CSV format, stored in a confidential RAG environment. The threat community are cybercriminals. The purpose of the attack is changing the weights of a decision-making process in order to favour men over women. Despite that this risk scenario is about fairness, it has the four risk perspectives as are presented in following Table 6:

Table 6: A description of each AI risk regulatory dimension

Dimension	Description
Data protection risk:	Confidential data of natural persons will be accessed by the attacker without authorization (GDPR, articles 5(f), 32).
Fairness risk	The candidates' selection process will be biased (AI Act, article 10 (2) (f), 10 (2) (g)).
Accuracy and robustness risk	Even though the classification metrics are properly applied, the result of the selection will be wrong because data was contaminated (AI Act, article 15 (1) (2) (3)).
Information security risk	A security policy has been violated, as access controls failed. (AI Act, article 15 (4) (5)).

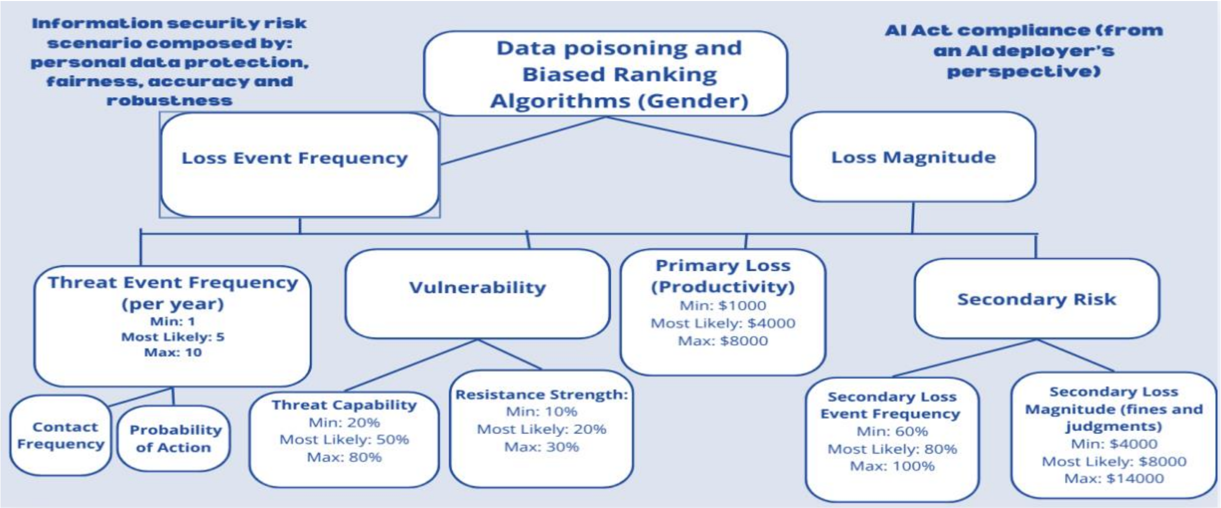
6. Artificial Intelligence Value at Risk (AI-VaR)

These AI risk dimensions can be included in a classic FAIR model, and the predicted interval will provide the input of an AI-VaR. To make it simple, it will only include *productivity* as a Primary Loss, and *fines and judgments* as a Secondary Loss, with the average of a Personal Data Value at Risk and a Fairness Value at Risk. The *resistance strength* factor will have two sources: information security controls that may fail because the access controls are not good enough, while the accuracy and robustness of the AI system may fail due to the lack of model performance checking controls. In this example, they will be included with the following input values presented in Table 7:

Table 7: Input values

Values	MIN	Most Likely	Max
Threat Event frequency:	Min = 1;	Most Likely = 5;	Max = 10.
Threat Capability:	Min = 20%;	Most Likely = 50%;	Max = 80%.
Resistance Strength	Min: 10%;	Most Likely = 20%;	Max = 30%.
Primary Loss (Productivity)	Min = $1000	Most Likely = $4000	Max = $8000.
Secondary Loss Frequency (Fines and Judgments)	Min= 60%;	Most Likely = 80%;	Max= 100%.
Secondary Loss Magnitude	Min = $4000	Most Likely = $8000	Max = $14000

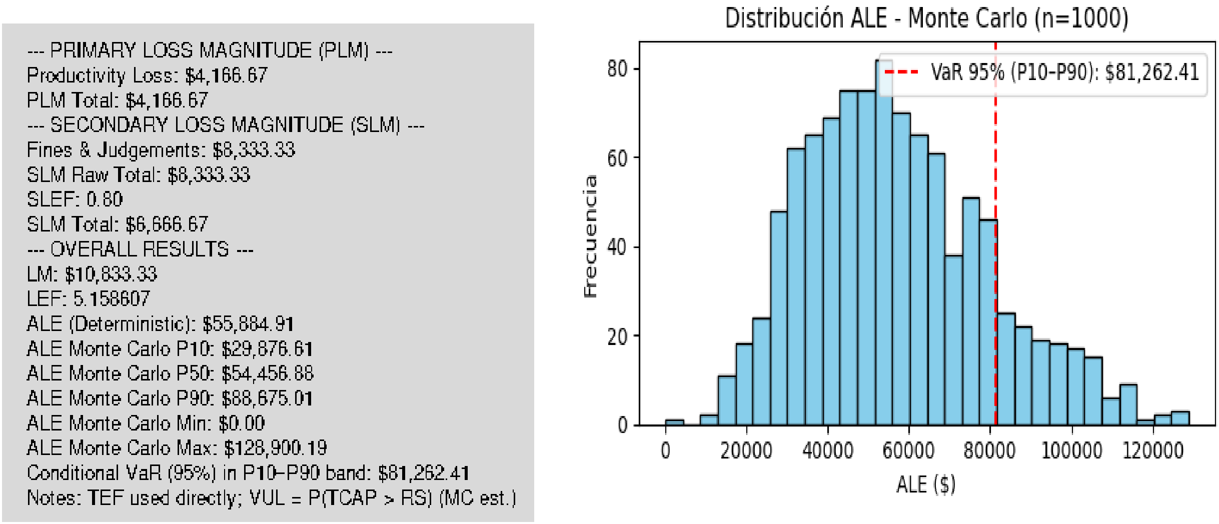
Figure 14: A risk scenario of data poisoning and biased ranking algorithms (classic FAIR model)


Information security risk scenario composed by: personal data protection, fairness, accuracy and robustness
Data poisoning and Biased Ranking Algorithms (Gender)
AI Act compliance (from an AI deployer's perspective)
Loss Event Frequency
Loss Magnitude
Threat Event Frequency (per year)
Min: 1
Most Likely: 5
Max: 10
Vulnerability
Primary Loss (Productivity)
Min: $1000
Most Likely: $4000
Max: $8000
Secondary Risk
Contact Frequency
Probability of Action
Threat Capability
Min: 20%
Most Likely: 50%
Max: 80%
Resistance Strength:
Min: 10%
Most Likely: 20%
Max: 30%
Secondary Loss Event Frequency
Min: 60%
Most Likely: 80%
Max: 100%
Secondary Loss Magnitude (fines and judgments)
Min: $4000
Most Likely: $8000
Max: $14000

Source: Author

From this model, we can implement the PERT formula and a Monte Carlo analysis with 1,000 simulations, both with the aim of getting an accurate annual loss exposure (ALE). It is possible to ping the distribution at certain percentiles with the aim of setting limits. For obtaining an AI-VaR, it is necessary to set a prediction interval. Since the previous raw distribution is unstable around the low and high limits, the most convenient way to implement the VaR is an adjusted version of a Value at Risk with truncated quantiles. It may be called a Conditional value at Risk, but the CvaR approach has been designed only for the worst losses (Rockefellar et al. 2022). In the following example, the lowest limit has been set at the 10th percentile and the highest limit at the 90th percentile of the raw distribution. The result shows the worst annual loss of $81 262, at the 95th confidence level within a chosen interval between the p10th and the p90th.

Figure 15: A Monte Carlo analysis and raw distribution



```
--- PRIMARY LOSS MAGNITUDE (PLM) ---
Productivity Loss: $4,166.67
PLM Total: $4,166.67
--- SECONDARY LOSS MAGNITUDE (SLM) ---
Fines & Judgements: $8,333.33
SLM Raw Total: $8,333.33
SLEF: 0.80
SLM Total: $6,666.67
--- OVERALL RESULTS ---
LM: $10,833.33
LEF: 5.158607
ALE (Deterministic): $55,884.91
ALE Monte Carlo P10: $29,876.61
ALE Monte Carlo P50: $54,456.88
ALE Monte Carlo P90: $88,675.01
ALE Monte Carlo Min: $0.00
ALE Monte Carlo Max: $128,900.19
Conditional VaR (95%) in P10–P90 band: $81,262.41
Notes: TEF used directly; VUL = P(TCAP > RS) (MC est.)
```

Distribución ALE - Monte Carlo (n=1000)
VaR 95% (P10-P90): $81,262.41
Frecuencia
0
20
40
60
80
0
20000
40000
60000
80000
100000
120000
ALE ($)

Source: Author

Conclusion

This paper has explored the creation of meaningful metrics and adequate models for AI risk scenarios from several risk perspectives: personal data protection, fairness, accuracy, robustness, and information security. Personal data protection metrics are fundamental to comply with frameworks such as the GDPR in the field of AI systems, focusing the analysis in two controversial spots: personal data as an input and personal data as an output. Fairness metrics are essential to reduce bias conditions within the datasets, where the vulnerability factor in the model ontology helps to compensate the disadvantaged AI users. Accuracy and robustness metrics are essential for the right performance of an AI system, as a bad AI system performance may also threaten the fundamental rights of natural persons. Finally, information security scenarios have been presented as the holy grail for AI risk integration, as without information security, AI systems will certainly fail.

The integration of these four kinds of AI metrics in an information security risk scenario becomes a very challenging but feasible task. Conformal prediction is a reliable method to obtain rationales when reliable historical data is available, and it has been shown how it can be used for the Personal Data Value at Risk. The goal of this approach is to first solve the data protection and fairness values at risk, as secondary losses, by using legal predictive analytics in order to understand the sanctioning psychology of the AI National Competent Authorities. When historical data is not available, other methods can be used such as the PERT formula and the Monte Carlo analysis. Then, it was shown how to assess the accuracy and robustness assessments in order to combine them in a holistic operational risk scenario, where the loss due to the bad performance of the AI system shall be integrated as resistance strength and sometimes as a potential secondary loss. Yet, in several cases, it would be necessary to model data protection, accuracy, robustness, and fairness in their own risk ontologies in order to export their output into an information security risk scenario. As an experimental work, the aim of this paper has been to show how AI metrics and models can be implemented in order to comply with new best practice standards, AI governance, and AI legal frameworks.

Credit Authorship Contribution Statement

The author did all the research contained in this paper. It includes, text, metrics, models, figures and tables.

Acknowledgments/Funding

The author did not receive any kind of funding.

Conflict of Interest Statement

The author declares that this research was conducted in the absence of any potential conflict of interest.

Data Availability Statement

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

References

Albina, O. (2021). Cyber risk quantification: Investigating the role of cyber value at risk. *Risks*, 9(10), 184. https://doi.org/10.3390/risks9100184

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? Arguments against avoiding RMSE in the literature. *Geoscientific Model Development,* 7, 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014

Chan, C.-P., Tsai, C.-H., Tang, F.-K. & Yang, J.-H. (2025) A SHAP-Based Comparative Analysis of Machine Learning Model Interpretability in Financial Classification Tasks. *Journal of Applied Economic Sciences*, Volume XX, Fall, 3(89), 385-400. https://doi.org/10.57017/jaes.v20.3(89).03

Cox, L. A. (2008). What's wrong with risk matrices? *Risk Analysis,* 28(2), 497–512. https://doi.org/10.1111/j.1539-6924.2008.01030.x.

Cronk, R. J., & Shapiro, S. S. (2021). *Quantitative privacy risk analysis*. In 2021 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW) Vienna, Austria, 2021, pp. 340-350. https://doi.org/10.1109/EuroSPW54576.2021.00043

Dewolf, N., De Baets, B., et al. (2020). Valid prediction intervals for regression problems (Version 4). arXiv. https://arxiv.org/abs/2107.00363.

Enríquez, L. (2024). A personal data value at risk approach. *Journal of Research, Innovation and Technologies*, 141–158. https://doi.org/10.57017/jorit.v3.2(6).05

Enríquez, L. (2024). Personal data breaches: Towards a deep integration between information security risks and GDPR compliance risks (Ph.D. thesis). Université de Lille, France. https://theses.hal.science/tel-04723327

Enríquez, L. (2024). Using the FAIR model as Swiss army knife of privacy uncertainty quantification for GDPR. FAIR Institute. https://www.fairinstitute.org/blog/fair-model-privacy-uncertainty-quantification-gdpr.

Floridi, L., Holweg, M., et al. (2022). capAI: A procedure for conducting conformity assessment of AI systems in line with the EU Artificial Intelligence Act (Version 1.0). https://d110erj175o600.cloudfront.net/wp-content/uploads/2022/03/24144824/report.pdf

Freund, J., & Jones, J. (2015). *Measuring and Managing Information Risk: A FAIR Approach*. 2nd Edition. Elsevier. Paperback ISBN: 978-0443134845

Fuerriegel, S., Hartmann, J., et al. (2023). Generative AI. arXiv. https://arxiv.org/abs/2309.07930

Hubbard, D. W. & Seiersen, R. (2016). *How to Measure Anything in Cybersecurity Risk*. 2nd Edition, John Wiley & Sons, 368 pages. ISBN: 978-1-119-89230-4

ISO/IEC. (2022). *ISO/IEC 22989:2022 Information technology - Artificial intelligence - Artificial intelligence concepts and terminology*. *International Organization for Standardization*. https://www.iso.org/standard/74296.html

ISO/IEC. (2023). ISO/IEC 23894:2023 Information technology - Artificial intelligence - Risk management. *International Organization for Standardization.* https://www.iso.org/standard/77304.html

ISO/IEC. (2022). ISO/IEC 27005:2022 Information security, cybersecurity and privacy protection - Information security risk management. *International Organization for Standardization*. https://www.iso.org/standard/80585.html

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). Noise: A flaw in human judgment. HarperCollins, 464 pp., ISBN: 978-0316451406

Krishnan, N. (2025). AI agents: Evolution, architecture, and real-world applications (Version 1). arXiv. https://arxiv.org/abs/2503.12687

Lawlor, R. (1963). What computers can do: Analysis and prediction of judicial decisions. *American Bar Association Journal*, 49(4), 337.

Maleki, N., Padmanabhan, B., et al. (2024). AI hallucinations: A misnomer worth clarifying. arXiv. https://arxiv.org/abs/2401.06796.

Manokhin, V. (2023). *Practical Guide to Applied Conformal Prediction in Python*. Packt Publishing. https://www.oreilly.com/library/view/practical-guide-to/9781805122760/

McCarthy, J. (2007). What is artificial intelligence? Stanford University. http://jmc.stanford.edu/articles/whatisai/whatisai.pdf

McCarthy, A., Ghadafi, E., Andriotis, P., & Legg, P. (2023). Defending against adversarial machine learning attacks using hierarchical learning: A case study on network traffic attack classification. *Journal of Information Security and Applications, 72*, Article 103398. https://doi.org/10.1016/j.jisa.2022.103398.

Minsky, M. (Ed.). (1968). *Semantic Information Processing*. MIT Press. http://geca.area.ge.cnr.it/files/6570.pdf

Mirković, M. S. (2020). Triangular distribution and PERT method vs. payoff matrix for decision-making support in risk analysis of construction bidding: A case study. *Facta Universitatis, Series: Architecture and Civil Engineering*, 18(3), 287–307. https://doi.org/10.2298/FUACE201117020M

Ouyang, L., Wu, J., et al. (2022). Training language models to follow instructions with human feedback. arXiv. https://arxiv.org/abs/2203.02155

Regulation (EU) 2016/679. (2016). General Data Protection Regulation. *Official Journal of the European Union*, L 119. https://eur-lex.europa.eu/eli/reg/2016/679/oj

Regulation (EU) 2024/1689. (2024). Artificial Intelligence Act. *Official Journal of the European Union*. https://eur-lex.europa.eu/legal-content/ro/ALL/?uri=oj:L_202401689

Rockafellar, R. T., & Uryasev, S. (2002). Conditional value-at-risk for general loss distributions. Journal of Banking and Finance, 26(7), 1443–1471. https://doi.org/10.1016/S0378-4266(02)00271-6

Samarati, P., & Sweeney, L. (1998). Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression (Technical Report). SRI International. https://epic.org/wp-content/uploads/privacy/reidentification/Samarati_Sweeney_paper.pdf

Shapiro, S. (2022). Time to modernize privacy risk assessment. Issues in Science and Technology, 38(1), 20–22. https://issues.org/wp-content/uploads/2021/10/20-22-Shapiro-Time-to-Modernize-Privacy-Risk-Assessment-Fall-2021.pdf

Sidorenko, A. (2017). Risk management used to be a science, then became an art, and now it's just bullsh@t. Risk Academy Blog. https://riskacademy.blog/first-blog-post

Society of Actuaries. (2025). Fundamentals of actuarial practice. https://www.soa.org/49347f/globalassets/assets/files/edu/edu-2012-c2-1.pdf

Vaswani, A., Shazeer, N., et al. (2017). Attention is all you need. arXiv. https://arxiv.org/abs/1706.03762