1. Metrics

The TAI-RPM introduces a set of metrics to correlate KPIs to specific cases and products. The indicators obtained from the metrics will accredit users to manage and act on the system based on risk-related information.

Furthermore, the proposed metrics are associated to system states to provide understanding on the level of improvement, risk state, and performance blended with ethical considerations. Specifically, four categories of metrics are proposed for the manufacturing sector on the processes associated to TAI-RMP.

Table 1 includes the reference nomenclature used for the overall metrics description included in next Subsections.

Table 1: Metrics Nomenclature

Symbol	Name	Comments
β	Failure effect probability	Conditional probability related to the severity of a failure effect.
α	Failure mode ratio	Ratio of the failure mode with respect of the overall failure modes that can occur on a selected AI artifact.
γ	Failure rate	Failing frequency for an AI artifact, expressed in units of time or operational cycles
C_m	Failure mode Criticality Number	Metric to classify as combination of probabilistic and temporal risk effects, estimated as the multiplication of $\gamma\beta\alpha t$
C_r	Criticality number	Total probabilistic and temporal risk effects of the AI artifact. Calculated as the sum of the C_m over all he failure modes of the same component
t	Time / number of activities	Time or cycles in which the AI artifact has been used. It must be in same units as γ metric. Different sub-indexes are used to define the referencing time (e.g. $t_{maintenance}$)

1.1. Metrics for FMEA and FMECA

When the user will perform any of the analyses using FMEA or FMECA then these metrics can be used to trace the risk tendency. The metrics should be grouped by specific risks on each trustworthy requirement.

When the information for failure rates associated to each of these groups (γ) is available, the criticality number (C_r) estimations must be used for this propose. From these estimations, a criticality matrix should be constructed to facilitate its tracking and visualization during subsequent stages of development. When there is no information available for γ, the risk priority number – RPN – can be used as a replacement to create the KPIs associated to this metric.

For example, and considering γ is available, the Criticality Number is used as accumulated critical value over the same AI artifact and same e-risk (Human Agency and Oversight or Accountability). The same scale must be used if several possible sources of critical numbers are used for comparison of intrinsic risk level – j in the next formula: $C_r = \sum_{n=1}^{j} c_{m,i}$

The Table 2 describes the metrics associated to FMEA and FMECA:

Table 2: FMEA and FMECA metrics

Name	Definition
Ethical Critical Number (ECN)	Criticality number for an specific trustworthy requirement and AI artifact. $ECN = \sum_{n=1}^{k} c_{r,i}$ where k identifies the requirement.
Ethical Relative Criticality Number (ERCN)	Ratio of item criticality number an specific trustworthy requirement and AI artifact over the total critical numbers produced by the system: $ERCN = \sum_{n=1} kc_{r,i} / \sum_{n=1} jc_{r,i}$ Where k is previously defined and j is the scale.

1.2. *Framework general and ethical based metrics*

These metrics defined in this Subsection support the general evaluation of the Risk Management Process. They are based on the general findings of the TAI-PRM. These are complementary to the decision-making processes of the risk status on an AI artifact under evaluation. Most of the KPIs should be associated to these metrics that represent the ratio values based on the risk limits – the concrete risk appetite allocations within the heatmap and the 4T's.

Table 3: High level metrics for management of risk ratio at overall system level

Name	Definition
$\%_{LU}$ - Unacceptable likelihood risk ratio	The likelihood of risks to materialise with an intrinsic value higher than the chosen by the management. It is calculated as: $\%_{LU} = N_{i>LU}/N_{risk}$ where $N_{i>LU}$ is the number of risk with likelihood over the stated limits and N_{risk} is the failure modes identified.
$\%_{LA}$ - Acceptance likelihood risk ratio	$1 - \%_{LU}$
$\%_{SU}$ - Unacceptable severity risk ratio	Number of risks with severity over limit chosen by the management Number. It is calculated as: $\%_{SU} = N_{i>SU}/N_{risk}$ where $N_{i>SU}$ is the number of risk with severity over the stated limits and N_{risk} is previously defined.
$\%_{SA}$ - Acceptable severity risk ratio	1 - $\%_{SU}$
$\overline{DC}$ - Detection Capacity	Average detection capacity of failure modes with risk levels over those in the stated limits. This number describes the average on how the system fail. It is calculated as: $\overline{DC} = \sum_{i=1}^{k} RPN_{i>DU}/N_{risk}$ where $RPN_{i>DU}$ is the risk priority number with detection over the stated limits, k are all the risk over the stated limit and N_{risk} is previously defined.

1.3. Independent to TAI-PRM

These metrics are proposed to to track the AI artifacts status to to the likelihood of failure events. These metrics are proposed to to track the AI artifacts status with KPIs related to Computer Science specifics. The ones suggested in this subsection are the most representative ones: availability, capacity, performance, and the accuracy. However, these can be extended depending on the AI artifacts – AI related or not.

Table 4: Metrics for AI components for software engineering non ethical based

Name	Definition
$AIEC$ AI Effective Capacity	This metric represents the AI artifact up-time from deployment (not training if applicable). If the use of the AI artifact is discrete it refers to the number of uses. When it a DSS it must be considered the outcomes accepted by users with respect to the times it run. It is calculated as: t_{used}/t_{total} for discrete AI utilisation; $t_{up-time}$ for continuous time.
$AIPPM$ AI Planned Maintenance	This metric represents the ratio of time or cycles used for scheduled downtime operations. It includes training /parametrization and AI maintenance. This is calculated as: $t_{scheduled}/t_{total}$
$AIDR$ AI Downtime Rate	This metric represents the ratio of unscheduled downtime. It must consider the unexpected events when the AI was idle or offline. This metric provides insights on AI stability. This is calculated as: $t_{unscheduled}/t_{total}$
$AICU$ AI Capacity Utilization	This metric provides the amount of time when the AI should be utilised with respect to the total available. The metric estimation is similar to $AIEC$ with the difference that it considers only the functional time. It is calculated as: $t_{used}/(t_{total} - t_{scheduled})$ for discrete AI utilisation; $t_{up-time}/(t_{total} - t_{scheduled})$ for continuous time.
$AICI$ AI correction indicator	This metric provides insights on the efficiency of the AI artifact with respect of a KPI (KPI) to perform contingency actions before they occur, compared to previous information available on the system. It is calculated as: $(KPI_{initial} - KPI_{new})/(KPI_{initial} - KPI_{old})$
TP True Positives	A classification of data performed by the AI artifact that provides a correct outcome. This metric represents the accumulated true positives over a period of time.
TN True Negatives	A classification of data performed by the AI artifact that provides an outcome that is not incorrect. This metric represents the accumulated true negatives over a time period.
FP False Positive	A classification of data performed by the AI artifact that provides an outcome that is not correct discarding the item from the right cluster. This metric is the accumulated false positives over a period of time.
FN False Negative	A classification of data that provides an outcome not correct adding the item into a category that is not adequate cluster. It represents the accumulated false negatives over a period of time.
AI_A Accuracy rate	Correct estimation as result of all TP and TN from the accumulated classification operations of the AI artifact. It is calculated as: $AI_A = (TP + TN)/(TP + TN + FP + FN)$
AI_M Error rate	Incorrect estimation as result of all FP and FN from the accumulated classification operations of the AI artifact. It is calculated as: $AI_M = (FP + FN)/(FP + FN + FP + FN)$
AI_P Precision rate	This metric provide the frequency of positive estimations. It is calculated as: $AI_P = TP/(TP + FP)$
$F1 - score$	This metric provides a mechanism to measure the AI classification performance for predictors. It is calculated replacing the true negative by true positive to avoid the negative factors. The equation is: $AI_{F1} = 2TP/(2TP + FP + FN)$

1.4. Environmental, social, and governance metrics

Environmental, Social, and Governance (ESG) is a broad field with many different investment approaches addressing various investment objectives that cover three areas. The first is the ESG integration, which improves the risk-return characteristics of investment. The second is the values-based investing, in which the investor seeks to align his investment with his norms and beliefs. Finally, impact investing seeks to trigger changes on the social or environmental scope.

A Morgan Stanley Capital International ESG Rating is designed to measure a company's resilience to long-term, industry material, environmental, social and governance (ESG) risks [1]. They propose a rules-based methodology to identify industry leaders and laggards according to their exposure to ESG risks and the efficiency on the management of those risks by peers.

They also rate equity and fixed income securities, loans, mutual funds, ETFs and countries.

Although ESG ratings are not directly linked to AI, they are suitable for manufacturing companies to reference their environmental social and governance status. However, the use of the ESG metrics is interesting for governance, but not usable to track internal risk management processes associated to the AI artifacts.

[1] https://www.msci.com/our-solutions/esg-investing/esg-ratings