POSITIVE AI

Responsible IA Label
Technical Framework

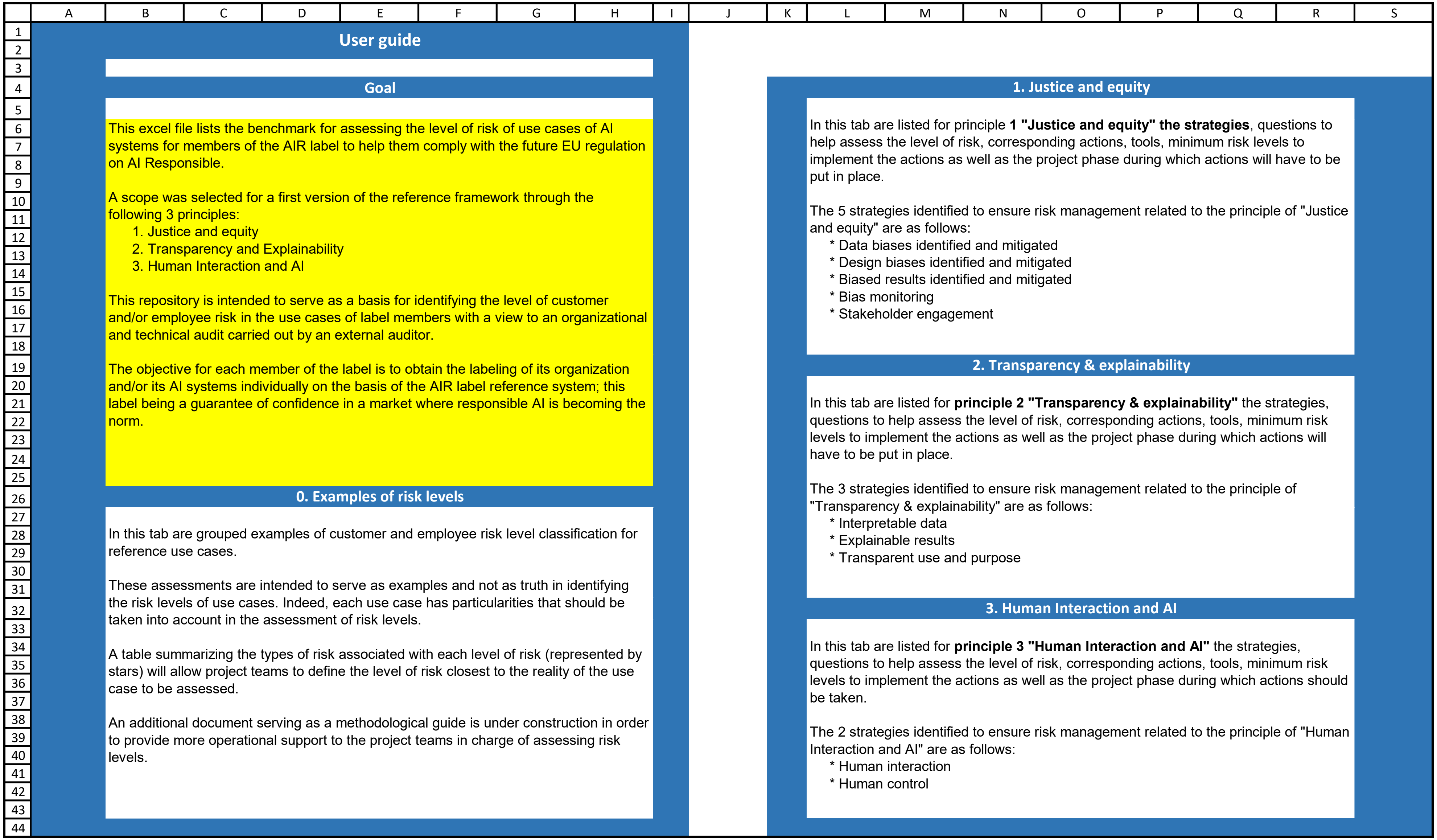
The version number and release date for the underlying report are :

Version 3.0

June 5, 2024

Version history

VERSION	DATE	EDITOR	MAIN CHANGES
1	Jan 31, 2022	Olivier Kahn	Initial version
1.1	March 31, 2022		Addition of the DVC tool in the Transparency & Explainability
1.2	April 12, 2022	Malo Grisard	Addition of a user guide tab + cleanliness of the repository
2.0	March 17, 2023	Amir Kroudir	Addition of Social and environmental impact,responsibility,data & private life, technical robustness and security
2.2	April 17, 2023	Amir Kroudir	Refinement of Social and environmental impact,responsibility,data & private life, technical robustness and security
3.0	June 5, 2024	Olivier Kahn/Niels Freier/Hugo Vallet/Corentin Delloye	Significant rework to extend the framework to GenAI systems and apps



User guide

Goal

This excel file lists the benchmark for assessing the level of risk of use cases of AI systems for members of the AIR label to help them comply with the future EU regulation on AI Responsible.

A scope was selected for a first version of the reference framework through the following 3 principles:

1. Justice and equity
2. Transparency and Explainability
3. Human Interaction and AI

This repository is intended to serve as a basis for identifying the level of customer and/or employee risk in the use cases of label members with a view to an organizational and technical audit carried out by an external auditor.

The objective for each member of the label is to obtain the labeling of its organization and/or its AI systems individually on the basis of the AIR label reference system; this label being a guarantee of confidence in a market where responsible AI is becoming the norm.

0. Examples of risk levels

In this tab are grouped examples of customer and employee risk level classification for reference use cases.

These assessments are intended to serve as examples and not as truth in identifying the risk levels of use cases. Indeed, each use case has particularities that should be taken into account in the assessment of risk levels.

A table summarizing the types of risk associated with each level of risk (represented by stars) will allow project teams to define the level of risk closest to the reality of the use case to be assessed.

An additional document serving as a methodological guide is under construction in order to provide more operational support to the project teams in charge of assessing risk levels.

1. Justice and equity

In this tab are listed for principle **1 "Justice and equity" the strategies**, questions to help assess the level of risk, corresponding actions, tools, minimum risk levels to implement the actions as well as the project phase during which actions will have to be put in place.

The 5 strategies identified to ensure risk management related to the principle of "Justice and equity" are as follows:

* Data biases identified and mitigated
* Design biases identified and mitigated
* Biased results identified and mitigated
* Bias monitoring
* Stakeholder engagement

2. Transparency & explainability

In this tab are listed for **principle 2 "Transparency & explainability"** the strategies, questions to help assess the level of risk, corresponding actions, tools, minimum risk levels to implement the actions as well as the project phase during which actions will have to be put in place.

The 3 strategies identified to ensure risk management related to the principle of "Transparency & explainability" are as follows:

* Interpretable data
* Explainable results
* Transparent use and purpose

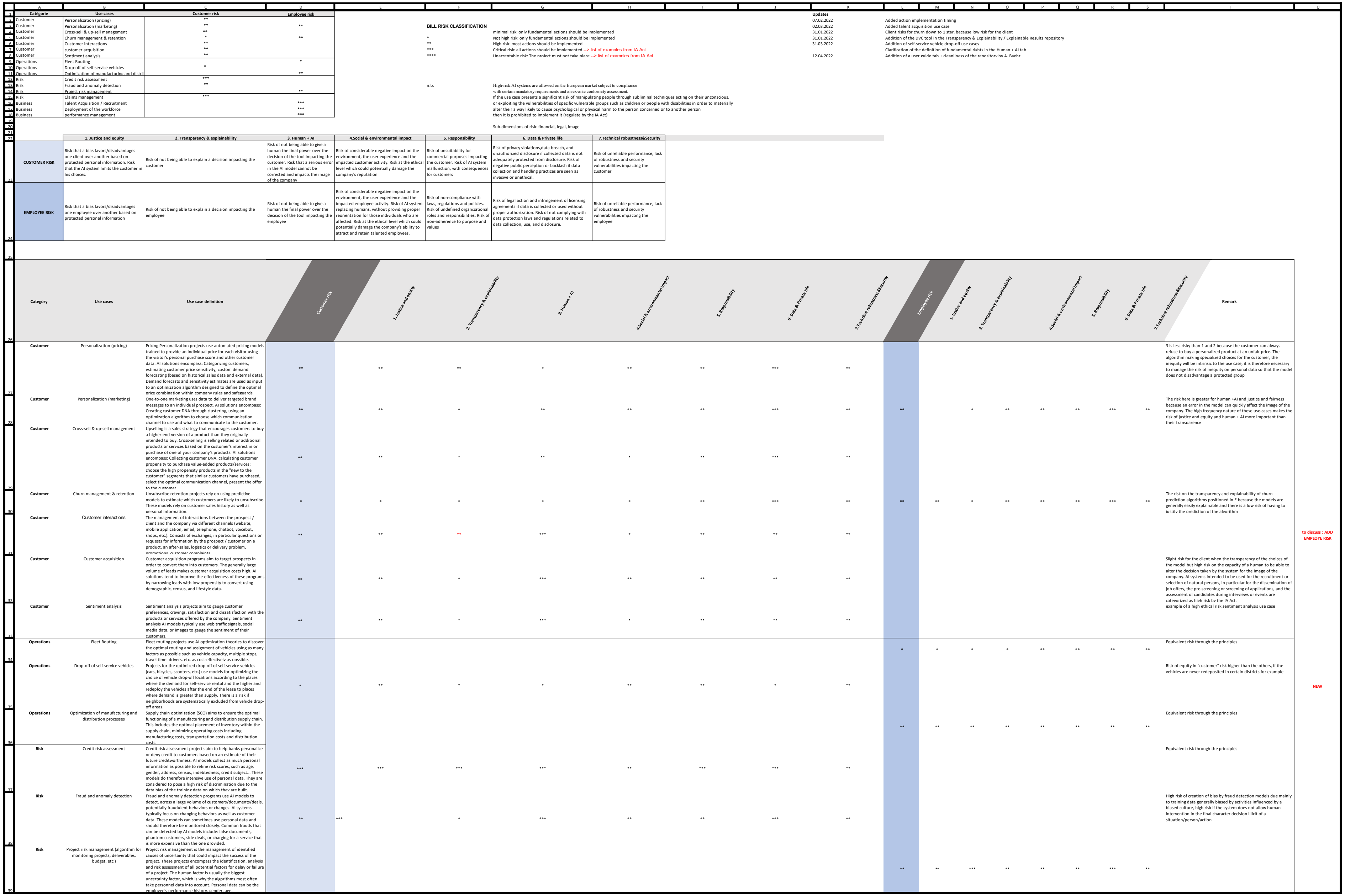
3. Human Interaction and AI

In this tab are listed for **principle 3 "Human Interaction and AI"** the strategies, questions to help assess the level of risk, corresponding actions, tools, minimum risk levels to implement the actions as well as the project phase during which actions should be taken.

The 2 strategies identified to ensure risk management related to the principle of "Human Interaction and AI" are as follows:

* Human interaction
* Human control

Technical Framework v3.0



Catégorie	Use cases	Customer risk	Employee risk
Customer	Personalization (pricing)	**	
Customer	Personalization (marketing)	**	**
Customer	Cross-sell & up-sell management	**	
Customer	Churn management & retention	*	**
Customer	Customer interactions	**	
Customer	customer acquisition	**	
Customer	Sentiment analysis	**	
Operations	Fleet Routing		*
Operations	Drop-off of self-service vehicles	*	
Operations	Optimization of manufacturing and distril		**
Risk	Credit risk assessment	***	
Risk	Fraud and anomaly detection	**	
Risk	Project risk management		**
Risk	Claims management	***	
Business	Talent Acquisition / Recruitment		***
Business	Deployment of the workforce		***
Business	performance management		***

BILL RISK CLASSIFICATION

	minimal risk: only fundamental actions should be implemented
*	Not high risk: only fundamental actions should be implemented
**	High risk: most actions should be implemented
***	Critical risk: all actions should be implemented --> list of examples from IA Act
****	Unacceptable risk: The project must not take place --> list of examples from IA Act

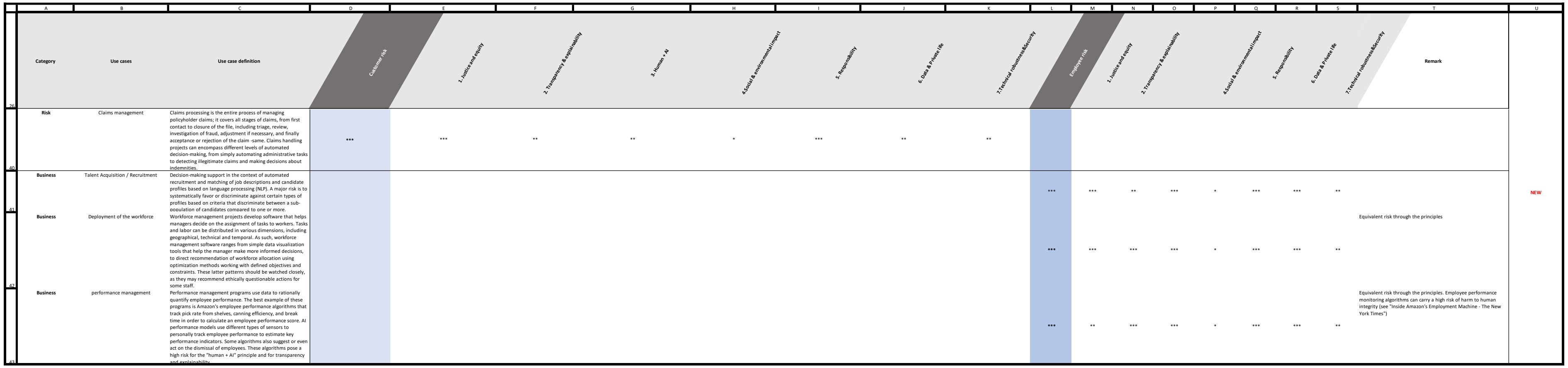
Updates	
07.02.2022	Added action implementation timing
02.03.2022	Added talent acquisition use case
31.01.2022	Client risks for churn down to 1 star. because low risk for the client
31.01.2022	Addition of the DVC tool in the Transparency & Explainability / Explainable Results repository
31.03.2022	Addition of self-service vehicle drop-off use cases
	Clarification of the definition of fundamental rights in the Human + AI tab
12.04.2022	Addition of a user guide tab + cleanliness of the repository by A. Baehr

n.b. High-risk AI systems are allowed on the European market subject to compliance
with certain mandatory requirements and an ex-ante conformity assessment.
If the use case presents a significant risk of manipulating people through subliminal techniques acting on their unconscious,
or exploiting the vulnerabilities of specific vulnerable groups such as children or people with disabilities in order to materially
alter their a way likely to cause psychological or physical harm to the person concerned or to another person
then it is prohibited to implement it (regulate by the IA Act)

Sub-dimensions of risk: financial, legal, image

	1. Justice and equity	2. Transparency & explainability	3. Human + AI	4.Social & environmental impact	5. Responsibility	6. Data & Private life	7.Technical robustness&Security
CUSTOMER RISK	Risk that a bias favors/disadvantages one client over another based on protected personal information. Risk that the AI system limits the customer in his choices.	Risk of not being able to explain a decision impacting the customer	Risk of not being able to give a human the final power over the decision of the tool impacting the customer. Risk that a serious error in the AI model cannot be corrected and impacts the image of the company	Risk of considerable negative impact on the environment, the user experience and the impacted customer activity. Risk at the ethical level which could potentially damage the company's reputation	Risk of unsuitability for commercial purposes impacting the customer. Risk of AI system malfunction, with consequences for customers	Risk of privacy violations,data breach, and unauthorized disclosure if collected data is not adequately protected from disclosure. Risk of negative public perception or backlash if data collection and handling practices are seen as invasive or unethical.	Risk of unreliable performance, lack of robustness and security vulnerabilities impacting the customer
EMPLOYEE RISK	Risk that a bias favors/disadvantages one employee over another based on protected personal information	Risk of not being able to explain a decision impacting the employee	Risk of not being able to give a human the final power over the decision of the tool impacting the employee	Risk of considerable negative impact on the environment, the user experience and the impacted employee activity. Risk of AI system replacing humans, without providing proper reorientation for those individuals who are affected. Risk at the ethical level which could potentially damage the company's ability to attract and retain talented employees.	Risk of non-compliance with laws, regulations and policies. Risk of undefined organizational roles and responsibilities. Risk of non-adherence to purpose and values	Risk of legal action and infringement of licensing agreements if data is collected or used without proper authorization. Risk of not complying with data protection laws and regulations related to data collection, use, and disclosure.	Risk of unreliable performance, lack of robustness and security vulnerabilities impacting the employee

Category	Use cases	Use case definition	Customer risk	1. Justice and equity	2. Transparency & explainability	3. Human + AI	4.Social & environmental impact	5. Responsibility	6. Data & Private life	7.Technical robustness&Security	Employee risk	1. Justice and equity	2. Transparency & explainability		4.Social & environmental impact	5. Responsibility	6. Data & Private life	7.Technical robustness&Security	Remark	
Customer	Personalization (pricing)	Pricing Personalization projects use automated pricing models trained to provide an individual price for each visitor using the visitor's personal purchase score and other customer data. AI solutions encompass: Categorizing customers, estimating customer price sensitivity, custom demand forecasting (based on historical sales data and external data). Demand forecasts and sensitivity estimates are used as input to an optimization algorithm designed to define the optimal price combination within company rules and safeguards.	**	**	**	*	**	**	***	**									3 is less risky than 1 and 2 because the customer can always refuse to buy a personalized product at an unfair price. The algorithm making specialized choices for the customer, the inequity will be intrinsic to the use case, it is therefore necessary to manage the risk of inequity on personal data so that the model does not disadvantage a protected group	
Customer	Personalization (marketing)	One-to-one marketing uses data to deliver targeted brand messages to an individual prospect. AI solutions encompass: Creating customer DNA through clustering, using an optimization algorithm to choose which communication channel to use and what to communicate to the customer.	**	**	*	**	**	**	***	**	**		*	**	**	**	***	**	The risk here is greater for human +AI and justice and fairness because an error in the model can quickly affect the image of the company. The high frequency nature of these use-cases makes the risk of justice and equity and human + AI more important than their transparency	
Customer	Cross-sell & up-sell management	Upselling is a sales strategy that encourages customers to buy a higher-end version of a product than they originally intended to buy. Cross-selling is selling related or additional products or services based on the customer's interest in or purchase of one of your company's products. AI solutions encompass: Collecting customer DNA, calculating customer propensity to purchase value-added products/services; choose the high propensity products in the "new to the customer" segments that similar customers have purchased, select the optimal communication channel, present the offer to the customer.	**	**	*	**	*	**	***	**										
Customer	Churn management & retention	Unsubscribe retention projects rely on using predictive models to estimate which customers are likely to unsubscribe. These models rely on customer sales history as well as personal information.	*	*	*	*	*	**	***	**	**	**	*	**	**	**	***	**	The risk on the transparency and explainability of churn prediction algorithms positioned in * because the models are generally easily explainable and there is a low risk of having to justify the prediction of the algorithm	
Customer	Customer interactions	The management of interactions between the prospect / client and the company via different channels (website, mobile application, email, telephone, chatbot, voicebot, shops, etc.). Consists of exchanges, in particular questions or requests for information by the prospect / customer on a product, an after-sales, logistics or delivery problem, promotions, customer complaints.	**	**	**	***	*	**	**	**										**to discuss : ADD EMPLOYE RISK**
Customer	Customer acquisition	Customer acquisition programs aim to target prospects in order to convert them into customers. The generally large volume of leads makes customer acquisition costs high. AI solutions tend to improve the effectiveness of these programs by narrowing leads with low propensity to convert using demographic, census, and lifestyle data.	**	**	*	***	**	**	**	**									Slight risk for the client when the transparency of the choices of the model but high risk on the capacity of a human to be able to alter the decision taken by the system for the image of the company. AI systems intended to be used for the recruitment or selection of natural persons, in particular for the dissemination of job offers, the pre-screening or screening of applications, and the assessment of candidates during interviews or events are categorized as high risk by the IA Act.	
Customer	Sentiment analysis	Sentiment analysis projects aim to gauge customer preferences, cravings, satisfaction and dissatisfaction with the products or services offered by the company. Sentiment analysis AI models typically use web traffic signals, social media data, or images to gauge the sentiment of their customers.	**	**	*	***	*	**	**	**									example of a high ethical risk sentiment analysis use case	
Operations	Fleet Routing	Fleet routing projects use AI optimization theories to discover the optimal routing and assignment of vehicles using as many factors as possible such as vehicle capacity, multiple stops, travel time, drivers, etc. as cost-effectively as possible.									*	*	*	*	**	**	**	**	Equivalent risk through the principles	
Operations	Drop-off of self-service vehicles	Projects for the optimized drop-off of self-service vehicles (cars, bicycles, scooters, etc.) use models for optimizing the choice of vehicle drop-off locations according to the places where the demand for self-service rental and the higher and redeploy the vehicles after the end of the lease to places where demand is greater than supply. There is a risk if neighborhoods are systematically excluded from vehicle drop-off areas.	*	**	*	*	**	**	*	**									Risk of equity in "customer" risk higher than the others, if the vehicles are never redeposited in certain districts for example	**NEW**
Operations	Optimization of manufacturing and distribution processes	Supply chain optimization (SCO) aims to ensure the optimal functioning of a manufacturing and distribution supply chain. This includes the optimal placement of inventory within the supply chain, minimizing operating costs including manufacturing costs, transportation costs and distribution costs.									**	**	**	**	**	**	**	**	Equivalent risk through the principles	
Risk	Credit risk assessment	Credit risk assessment projects aim to help banks personalize or deny credit to customers based on an estimate of their future creditworthiness. AI models collect as much personal information as possible to refine risk scores, such as age, gender, address, census, indebtedness, credit subject... These models do therefore intensive use of personal data. They are considered to pose a high risk of discrimination due to the data bias of the training data on which they are built.	***	***	***	***	**	***	***	**									Equivalent risk through the principles	
Risk	Fraud and anomaly detection	Fraud and anomaly detection programs use AI models to detect, across a large volume of customers/documents/deals, potentially fraudulent behaviors or changes. AI systems typically focus on changing behaviors as well as customer data. These models can sometimes use personal data and should therefore be monitored closely. Common frauds that can be detected by AI models include: false documents, phantom customers, side deals, or charging for a service that is more expensive than the one provided.	**	***	*	***	**	**	***	**									High risk of creation of bias by fraud detection models due mainly to training data generally biased by activities influenced by a biased culture, high risk if the system does not allow human intervention in the final character decision illicit of a situation/person/action	
Risk	Project risk management (algorithm for monitoring projects, deliverables, budget, etc.)	Project risk management is the management of identified causes of uncertainty that could impact the success of the project. These projects encompass the identification, analysis and risk assessment of all potential factors for delay or failure of a project. The human factor is usually the biggest uncertainty factor, which is why the algorithms most often take personnel data into account. Personal data can be the employee's performance history, gender, age.									**	**	***	**	**	**	***	**		



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
26	**Category**	**Use cases**	**Use case definition**	Customer risk	1. Justice and equity	2. Transparency & explainability	3. Human + AI	4.Social & environmental impact	5. Responsibility	6. Data & Private life	7.Technical robustness&Security	Employee risk	1. Justice and equity	2. Transparency & explainability		4.Social & environmental impact	5. Responsibility	6. Data & Private life	7.Technical robustness&Security	**Remark**	
40	**Risk**	Claims management	Claims processing is the entire process of managing policyholder claims; it covers all stages of claims, from first contact to closure of the file, including triage, review, investigation of fraud, adjustment if necessary, and finally acceptance or rejection of the claim -same. Claims handling projects can encompass different levels of automated decision-making, from simply automating administrative tasks to detecting illegitimate claims and making decisions about indemnities.	***	***	**	**	*	***	**	**										
41	**Business**	Talent Acquisition / Recruitment	Decision-making support in the context of automated recruitment and matching of job descriptions and candidate profiles based on language processing (NLP). A major risk is to systematically favor or discriminate against certain types of profiles based on criteria that discriminate between a sub-population of candidates compared to one or more.									***	***	**	***	*	***	***	**		NEW
42	**Business**	Deployment of the workforce	Workforce management projects develop software that helps managers decide on the assignment of tasks to workers. Tasks and labor can be distributed in various dimensions, including geographical, technical and temporal. As such, workforce management software ranges from simple data visualization tools that help the manager make more informed decisions, to direct recommendation of workforce allocation using optimization methods working with defined objectives and constraints. These latter patterns should be watched closely, as they may recommend ethically questionable actions for some staff.									***	***	***	***	*	***	***	**	Equivalent risk through the principles	
43	**Business**	performance management	Performance management programs use data to rationally quantify employee performance. The best example of these programs is Amazon's employee performance algorithms that track pick rate from shelves, canning efficiency, and break time in order to calculate an employee performance score. AI performance models use different types of sensors to personally track employee performance to estimate key performance indicators. Some algorithms also suggest or even act on the dismissal of employees. These algorithms pose a high risk for the "human + AI" principle and for transparency and explainability									***	**	***	***	*	***	***	**	Equivalent risk through the principles. Employee performance monitoring algorithms can carry a high risk of harm to human integrity (see "Inside Amazon's Employment Machine - The New York Times")	

Technical Framework v3.0

	A	B	C	D	E	F	G	H
1	**Principle**	**Strategy**	**Evaluation question**	**Corresponding action**	**Tools (illustrative, non-exhaustive)**	**Customer risk level minimum to implement the action**	**Employee risk level minimum to implement the action**	**Project phase for the implementation of the action**
2	The development, deployment and use of AI systems must be equitable. While we recognize that fairness can be interpreted in multiple ways, we consider fairness to be characterized by both a material and a procedural component. The material component involves a commitment to ensure an equal and fair distribution of benefits and costs, and to ensure that individuals and groups are not subject to unfair bias, discrimination and stigma. If unfair biases can be avoided, AI systems might even improve the fairness of society. Equal opportunity should also be encouraged in terms of access to education, goods, services and technology. Furthermore, the use of AI systems should never have the effect of misleading (end) users or limiting their freedom of choice. Fairness further implies that AI professionals should respect the principle of proportionality between ends and means, and carefully consider how to balance competing interests and goals. The procedural aspect of fairness involves the ability to challenge the decisions made by AI systems and by the human beings who use them, as well as the ability to introduce an effective remedy against these decisions.	**Data biases identified and mitigated**	• Do you have an appropriate definition of “fairness” that you apply in the design of AI systems? - Is your definition commonly used? Did you consider other definitions before choosing this one? - Have you planned a quantitative analysis or indicators to measure and test the applied definition of equity? - Do you have mechanisms in place to ensure fairness in your AI systems? Have you considered other potential mechanisms? • Do you have a strategy in place to avoid creating or reinforcing unfair biases in the use of input data? - Have you assessed and recognized any limitations arising from the composition of the datasets used? - Have you thought about the diversity and representativeness of users in the data? Did you test for specific populations or problematic use cases?	• Establish a list of potentially discriminating variables to be excluded by default, unless specific exemption justified by the use case • For each project, explicitly define a notion of fairness adapted to the AI system in question • For each project, conduct a bias analysis: - identify the issues that are inherent in the data collection process - identify the groups of individuals to be protected from possible bais - identify the proxy variables representing these groups - evaluate the biases of historical data using statistical tests and a bias indicator - if necessary, reduce identified data biases (e.g. re-weighting of data)	Package Aequitas AIF 360 package pandas.DataFrame.corr List of variables likely to bring a discriminatory bias)	*	*	- Definition of the use case - Prototyping and preparation of the first pilot
3		**Design biases identified and mitigated**	• Do you have a strategy in place to avoid creating or reinforcing unfair biases in the design of the algorithm? - Do you have processes in place to test and control for potential bias during the development, deployment and use phase of the system?	Conduct a bias analysis on the algorithm: Estimate the bias indicator of the model and compare it to the historical bias If necessary, reduce design bias: - delete sensitive variables if necessary - simulate bias reduction strategies and analyze their impact on model accuracy, - for example in: - using model recalibration methods (e.g Adversarial debiasing) - modifying the thresholds for decisions on protected groups	Package Aeaquitas AIF 360 package pandas.DataFrame.corr	**	**	- Prototyping and preparation of the first pilot
4		**Biased results identified and mitigated**	• Have you assessed whether, under identical conditions, a possible variability of decisions is possible? - If so, have you thought about probable causes? - Regarding variability, have you set up a mechanism to measure or assess the potential impact of this variability on fundamental rights?	• Analyze the variance of model predictions within groups of individuals to be protected	Pandas: pandas.core.groupby.GroupBy.var	***	***	- Prototyping and preparation of the first pilot
5		**Bias monitoring**	• Have you implemented bias monitoring?	• plan thresholds and configure tools • prepare actions in the event of bias detection (analysis of data format causes, and retraining decision)		**	**	preparation for deployment
6		**Stakeholder engagement**	• Have you thought about a mechanism to include the participation of different stakeholders in the development and use of the AI system? • Have you prepared the way for the introduction of the AI system within your organization by informing and mobilizing the affected workers and their representatives beforehand? • Depending on the use case, do you have a mechanism for others to report issues related to bias, discrimination, or poor performance of the AI system?	• Include teams from the system design phase • Introduce the AI system gradually through pilots and to obtain stakeholder feedback on the system • Organize training for stakeholders on issues of justice and equity. In particular, inform about possible problematic biases of the AI system.		**	*	- Definition of the use case - Deployment preparation
7								
8								
9								
10								
11								
12								

Technical Framework v3.0

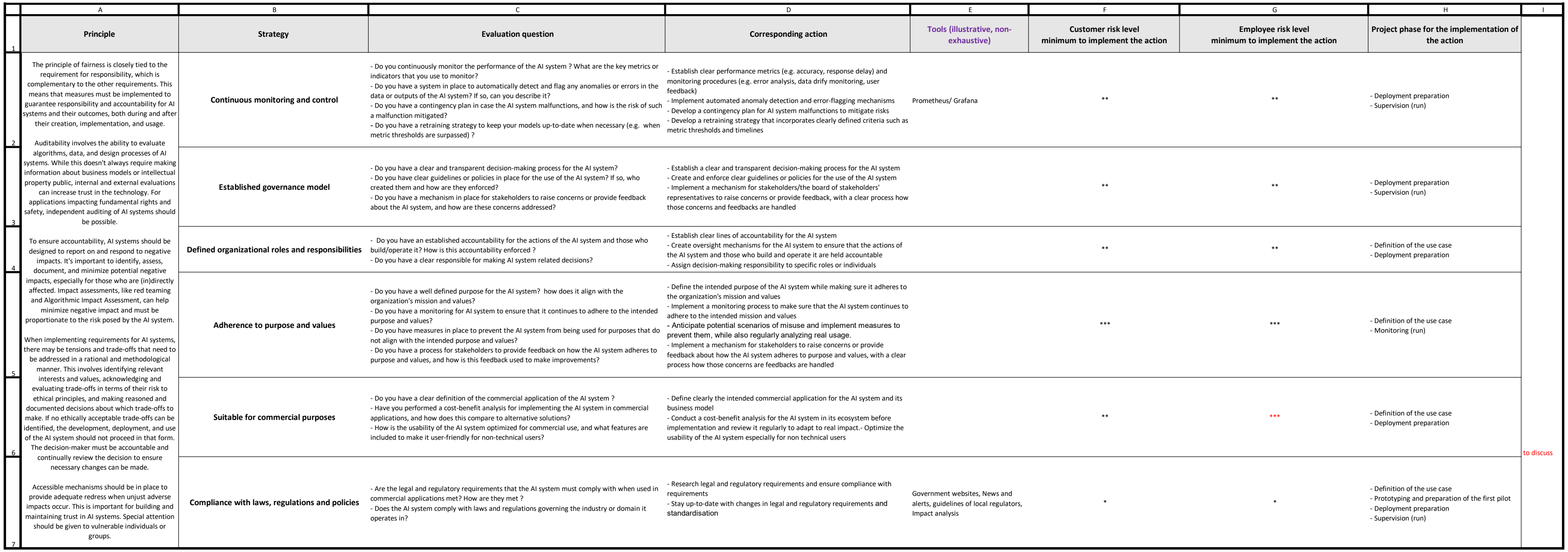
	A	B	C	D	E	F	G	H
1	**Principle**	**Strategy**	**Evaluation question**	**Corresponding action**	**Tools (illustrative, non-exhaustive)**	**Customer risk level minimum to implement the action**	**Employee risk level minimum to implement the action**	**Project phase for the implementation of the action**
2	Explainability is key to building and maintaining user trust in AI systems. This means that processes must be transparent, the capabilities and purpose of AI systems must be openly communicated, and decisions – where possible – must be explainable to those directly and indirectly affected. Without this information, a decision cannot be properly challenged. It is not always possible to explain why a model generated a particular outcome or decision (and what combination of input factors contributed to it). These are known as “black box” effect algorithms. These should be given special attention. In such circumstances, other measures of explainability (e.g. traceability, auditability and transparent communication about the capabilities of the system) might be required, provided that the system as a whole respects fundamental rights. The extent to which explainability is needed depends heavily on the context and the severity of the consequences if that result is wrong or otherwise inaccurate.	**Interpretable data**	Have you assessed whether you are able to analyze the data you used for training and testing purposes? Can this be changed and updated over time?	• Establish data traceability: provenance, lineage and history - Nomenclature of variables, check the human interpretability of databases • Conduct an analysis and document the quality of the data in order to verify that they are of sufficient quality to allow the interpretability of the results: - missing data, uniqueness, distribution of variables, correlations, etc GenAI Specific : • Investigate the position of pre-trained LLM providers with regard to the interpretability of training data - Verify wether the documentation of the pre-trained LLM specifies the data sources used for pre-training - Assess the credibility and reliability of the organizations providing the pre-trained LLMs - Investigate the measures taken by the pre-trained LLM providers to mitigate risks of data unfairness • If necessary, implement additional bias mitigation and fairness strategies . (for eg, fine-tuning the pre-trained models with supplementary, unbiased, and diverse data sets)	Pandas Profiling package: df.profile_report() DVC Datasets : https://github.com/mlabonne/llm-datasets	*	*	- Prototyping and preparation of the first pilot
3		**Explainable results**	Have you thought to use the simplest and easiest to interpret model for the application in question?	• Conduct a rigorous model selection analysis by: - choosing an appropriate model validation metric (RMSE, AUC, MAPE, etc.) - evaluating whether the gains in precision obtained compensate for the loss of interpretability induced by the type of model used GenAI Specific : • Determine the trade-offs between using a full-scale LLM versus simpler alternatives like decision trees or rule-based systems, NLP models, in terms of accuracy, response time, and understandability	Gamma facet SHAP value analysis packages LIME package (or ELI5) for classifiers interpretability Scikit learn: sklearn.model_selection	**	**	- Prototyping and preparation of the first pilot
4			Have you assessed the extent to which the decisions made, and therefore the results achieved, by the AI system can be understood?	Conduct an explainability analysis of the models used within the AI system: • identify the types of algorithms used: system of rules, interpretable model, or "black box", • map the dependencies between these models • document rule systems • for models that are difficult to interpret, use a SHAP importance variable plot to understand the influence of each variable on the results of a model GenAI specific : • Enhance explainability of LLMs : - Document how different components of the LLM interact, especially when multiple models or layers contribute to the final decision - Implement advanced visualization tools like attention maps or feature influence charts for educational purposes - Develop prompting protocols to understand the influence of prompts on LLM decision-making	Gamma Facet LIME package Shap values analysis packages BertViz Develop protocol using method as CoT prompting...	*	*	- Supervision (run)
5			Have you ensured that an explanation of why a system has made a certain choice leading to a certain outcome can be made understandable to all users who may wish to obtain an explanation? Have you assessed whether any solutions are available to you as a result of training and fine-tuning the model to examine the interpretation or whether you have access to the model's sequence of operations?	• Establish a traceability of the actions of the AI system in order to be able to replay an individual result and link it to the explainability analysis • Use an individual SHAP value plot to assess the impact of the variables that led to the individual result GenAI specific : • Establish a detailed logging system that records step-by-step computations and decisions made by the LLM, allowing for the replay of specific decisions • Develop user-friendly interfaces that allow users to visualize how input affect output via tools adapted for LLMs, such as example-based explanations or simplified narrative explanations	MLFlow Shap value analysis package LIME package LIT by Google Giskard IBM WatsonX.ai API (LLM "management" tool - model tracking, explainability…)	**	**	- Prototyping and preparation of the first pilot
6			Did you survey end users to verify their understanding of the underlying explanatory elements that led to the model's results or recommendations?	• Conduct a usage study (survey) with end users to gauge understanding • Organize structured surveys and interactive sessions where users can engage with the LLM outputs and provide feedback on their understanding of the explanations provided • Adjust the model's output presentation based on user feedback (fine tuning model with RLHF)		***	***	- Supervision (run)
7		**Transparent use and purpose**	Have you assessed why this particular system was deployed in this specific area?	• Clearly identify and document the benefits expected from the AI system that justify its deployment		*	*	- Definition of the use case -Run
8			Did you design the AI system with interpretation in mind from the start?	• Organize training for stakeholders around the concept of “correlation is not causation”		*	*	
9			Have you assessed the extent to which the system decision influences the organization's decision-making processes?	• *Map the business units and associated decision-making processes impacted by the AI system* • *Simulate the impact of adopting the AI system by performing backtests to estimate the impact on well-chosen KPIs*	Scikit learn	**	*	

	A	B	C	D	E	F	G	H
1	**Principle**	**Strategy**	**Evaluation question**	**Corresponding action**	**Tools (illustrative, non-exhaustive)**	**Customer risk level minimum to implement the action**	**Employee risk level minimum to implement the action**	**Project phase for the implementation of the action**
2	The fundamental rights on which the EU is founded are intended to guarantee respect for the freedom and autonomy of human beings. Humans interacting with AI systems must be able to maintain full and effective self-determination and participate in the democratic process. In the absence of justification, AI systems should not subordinate, coerce, deceive, manipulate, condition or dictate human beings. Rather, AI systems should be designed to augment, complement and foster cognitive, social and cultural skills. The division of labor between humans and AI systems should follow human-centered design principles and give humans real opportunity to make choices. In other words, human supervision and control over the working processes of AI systems should be ensured. AI systems could also fundamentally change the sphere of work. These systems should support human beings in the work environment, and aim to create meaningful jobs.	**human interaction**	In use cases that may lead to adverse effects on fundamental rights, have you carried out a fundamental rights impact assessment? Have you identified and documented the potential use of trade-offs between different principles and rights? Does the AI system interact with human end-user decision-making (for example, recommending actions or decisions to take, or presenting possible choices)? o In such cases, is there a risk that the AI system will affect human autonomy by unintentionally interfering with end-user decision-making? o Do you believe that an AI system should communicate to users that a decision, content, advice or result stems from an algorithmic decision? o When the AI system includes a robot or chat system, are human users aware that they are interacting with a virtual agent?	Conduct a risk analysis on fundamental rights: - Have you identified risks of loss of capacity for action or influence for the people who use or interact with the system? (loss of autonomy, addiction, attachment, abusive trust, confinement in a personal bubble, loss of free will, limitation of individual or collective freedoms, surveillance, influence on behavior or vulnerable people) - Does the system concern vulnerable people (children, the elderly, the disabled, etc.) For cases presenting risks: - make explicit to the user the virtual and automatic nature of the system - make available documentation explaining clearly and in simple terms how the AI system produces its recommendation / decision - as far as possible, allow the user to challenge the recommendation / decision GenAi Specific : - Implement clear mechanisms to inform users when interacting with LLMs and when outputs are algorithmically generated - Provide transparency about the LLM's capabilities and limitations - Regularly monitor the impact of LLMs on human users, specifically focusing on any unintended interference with decision-making	PIA CNIL tool: https://www.cnil.fr/fr/outil-pia-telechargez-et-installez-le-logiciel-de-la-cnil	**	**	- Definition of the use case - Prototyping and preparation of the first pilot - Supervision (run)
3								
4		**Human control**	Have you considered the appropriate level of human control for the AI system and use case in question? o Can you describe the level of human control or involvement, if any? Who is the "human in charge" and when is there human intervention, or with what tools? o Do you have mechanisms and measures in place to ensure potential human control or oversight of this nature, or to ensure that decisions are made under the overall responsibility of human beings? o Have you taken steps to enable audits and resolve issues related to the governance of AI autonomy?	Put in place measures to assess the governance of the autonomy of the system: - set up a traceability of the automated decisions of the AI system impacting human autonomy - regularly carry out audits on these traces in order to assess whether the governan		**	**	- use case definition - Supervision (run)
5				Put in place measures to assess the governance of the autonomy of the system: - set up a traceability of the automated decisions of the AI system impacting human autonomy - regularly carry out audits on these traces in order to assess whether the governance in place is appropriate to the extent of the risks - ensure that public authorities are able to exercise control in accordance with their mandate - when necessary, supplement this traceability with automated risk detection mechanisms - the degree of governance control must be adapted to the potential risks. In particular, it should be strengthened for self-learning systems		***	***	- Definition of the use case
6								

Technical Framework v3.0

	A	B	C	D	E	F	G	H
1	**Principle**	**Strategy**	**Evaluation question**	**Corresponding action**	**Tools (illustrative, non-exhaustive)**	**Customer risk level minimum to implement the action**	**Employee risk level minimum to implement the action**	**Project phase for the implementation of the action**
2	The AI system's life cycle should prioritize fairness and the prevention of harm by considering the broader society, other sentient beings, and the environment as stakeholders. It is important to encourage sustainability and ecological responsibility in AI systems and promote research into AI solutions that address global concerns, such as the Sustainable Development Goals. The ideal use of AI systems should benefit all human beings, including future generations. The development and use of AI systems have the potential to address significant societal concerns, but it is essential to ensure that they are sustainable and environmentally friendly. The entire life cycle of the AI system, including its development, deployment, and supply chain, should be evaluated to minimize its environmental impact. This can involve critically examining the resources used and energy consumption during training and choosing less harmful options. Promoting measures that secure the environmental friendliness of the entire supply chain of AI systems is crucial. The widespread use of AI systems in various aspects of our lives, such as education, work, care, or entertainment, can have a significant impact on our social agency and relationships. While AI systems can enhance social skills, they can also negatively affect them, potentially leading to a decline in physical and mental well-being. As a result, it is crucial to monitor and carefully consider the effects of these systems on individuals. In addition to evaluating the impact of AI systems on individuals during development, deployment, and use, it is also essential to assess their impact on institutions, democracy, and society as a whole. Special consideration should be given to the use of AI systems in situations related to the democratic process, such as political decision-making and electoral contexts.	**Sustainably developed by design**	- Have you evaluated the AI system to ensure its sustainability and environmental friendliness throughout its life cycle, including development, deployment, and monitoring? -Do you use a measuring tool to estimate/monitor the resources utilization of the AI system? do you have actions in place to minimize these resources consumption, thus limiting environmental impact	- Implement a measuring system for resources utilization during the AI system development (training) and use in production (inference): • Hardware Efficiency: what is the amount of embodied carbon, non-renewable raw materials utilization, water utilization and pollution impact? • Energy Efficiency: what is the amount of energy consumed? • Carbon Awareness: do you do more when the electricity is cleaner and do less when the electricity is dirtier? Put in place actions to minimize resources utilization, thus limiting environmental impact. e.g. of mechanisms include (categorized by level of difficulty to put in place): Level 1 - quick wins: - Review frequency of training, and the size of training dataset (subsampling) - Use carbon efficient standards for putting models into production (e.g ONNX) - Plan the heavy energy consuming tasks in datacenter location where energy is clean and in periods where energy is cleaner - **GenAI specific**: If model inferences exposed to large user base, setup mechanisms to limit number of inferences: API calls limit per user, inference batch pre-compute at fixed time intervals, exclusion of certain "low value" users from inference base, etc.	**AI systems in general** - Green Software Foundation standard for measurement - CodeCarbon - MLCO2 Impact (https://mlco2.github.io) - "energyusage" python package - FinOPS (reducing cloud expenses can lead to reducing carbon footprint) - Green Software Foundation Green patterns - ghgprotocol (standards to measure and manage emissions) - THE 17 GOALS - Sustainable Development Goals : https://sdgs.un.org/goals	***	***	- Definition of the use case - Prototyping and preparation of the first pilot - Supervision (run)
3				Level 2 - model complexity VS performance: - Review the complexity of AI models VS performance metric to find a middle ground (e.g. accuracy) - Review the feature engineering, perform feature selection, review number of serialized intermediate datasets, and results - Use a code monitoring tool for carbon emissions (e.g. Code Carbon) - GenAI specific: Review the use of externally trained models and AI components (e.g. LLMs), their complexity, the declared carbon footprint declared by provider (e.g. carbon footprint per API call) if available - GenAI specific: When using LLMs / GenAI trade-off between small/large models (e.g. Mistral 7B) Level 3 - optimisation of code and infrastructure: - Review the technical infrastrure to minimize idle time impact in training and inference - Optimise code focusing on limiting compute (CPU/GPU/TPU) utilisation, as it is by far the main carbon footprint contributor - Optimise your hardware setup (e.g. going from CPU architecture to GPU)	**GenAI specific** LLM footprint still subject of research debate as of June 2024, but current view is that footprint can be decomposed between the footprint of running the Fundation Model (FM), possibly on heavy infra, to do inference only and the footprint of fine-tuning the FM / doing RAG, which can be tracked using regular methods (see tools above). Bellow an interesting research paper synthetisizing known ways to estimate footprint for FMs - https://arxiv.org/html/2309.14393v2/#S4.E3			
4		**Promoting positive outcomes**	- Have you conducted an impact analysis of the design and/or testing of the AI system on the end-user experience and activities over the initial state? - Have you conducted an impact analysis of the design and/or testing of the AI system on groups with higher vulnerabilities? - Have you conducted an analysis of the impact of the design and/or testing of the AI system on the environmental impact of the activity in question? In the case where there is a negative impact on the end-user experience or/and environmental impact of the activity, have you built an action plan to mitigate and reduce the negative impact ? - Does the AI system address global concerns such as the Sustainable Development Goals? - Does the AI system benefit all human beings, including future generations?	- Conduct an impact analysis on end-user experience of the activity (productivity, quality of the output, stress-levels, growth and developement) before and after using the AI system - Conduct an impact analysis on society as a collective segmented by groups, with a focus on vulnerability groups - Conduct an analysis on the environmental impact of the activity before and after using the AI system In case of a negative impact on the end-user experience or/and the environmental impact of the activity : - Build an action plan to mitigate and reduce those negative impacts - Conduct an assessment of the long term implications of the AI System (impact on future generations) - Conduct ex-post evaluation and usage review for the AI system - Build a roadmap on how to sustain a long term positive implications of the AI system - **GenAI specific:** Publish disclaimers / trainings materials to explain the limitations and the known (accepted) risks of your system (e.g. LLM-based limitations disclaimer for public facing chatbot)	- Clarity AI (UN SDGs Impact, ESG Impact) - sesamm - giskard.ai (auto-test suite)	***	**	- Definition of the use case
5				- Perform an analysis of the required change management and work reorganization in the event of job cuts, and prepare plans for the reorientation of affected employees			*	
6		**Avoidance of societal harms**	- Have you assessed the potential negative social and ethical implications of the AI system's use and deployment ? How are those potential harms adressed ? - Have you developed policies to prevent the use and deployment of AI systems from having a negative impact on society? - Have you assessed the AI system impact on social skills and the potential negative effects it can have on individuals' physical and mental well-being? - Have you assessed the impact of AI systems on institutions, democracy, and society as a whole? - Do you have a strategy in place to avoid the use of AI systems in situations related to the democratic process, such as political decision-making and electoral contexts ?	-Develop policies and procedures,workshops to mitigate potential risks and negative impacts on society - Access AI system impact on physical, mental health and well being of individuals, and build an action plan to mitigate those impacts - Access AI system impact on democracy, political decision making and elections and build an action plan to mitigate those impacts - Create a board of stakeholders' representatives or establish a dialogue with them to regularly evaluate the impact and collaboratively generate proposals for mitigating it	- Surveys / Documentation - for builders (e.g. data scientists) - **GenAI specific:** Disclaimers / training - for users (e.g. chatbot users) - The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems	***	***	- Definition of the use case

Technical Framework v3.0



	A	B	C	D	E	F	G	H	I
1	**Principle**	**Strategy**	**Evaluation question**	**Corresponding action**	**Tools (illustrative, non-exhaustive)**	**Customer risk level minimum to implement the action**	**Employee risk level minimum to implement the action**	**Project phase for the implementation of the action**	
2	The principle of fairness is closely tied to the requirement for responsibility, which is complementary to the other requirements. This means that measures must be implemented to guarantee responsibility and accountability for AI systems and their outcomes, both during and after their creation, implementation, and usage. Auditability involves the ability to evaluate algorithms, data, and design processes of AI systems. While this doesn't always require making information about business models or intellectual property public, internal and external evaluations can increase trust in the technology. For applications impacting fundamental rights and safety, independent auditing of AI systems should be possible. To ensure accountability, AI systems should be designed to report on and respond to negative impacts. It's important to identify, assess, document, and minimize potential negative impacts, especially for those who are (in)directly affected. Impact assessments, like red teaming and Algorithmic Impact Assessment, can help minimize negative impact and must be proportionate to the risk posed by the AI system. When implementing requirements for AI systems, there may be tensions and trade-offs that need to be addressed in a rational and methodological manner. This involves identifying relevant interests and values, acknowledging and evaluating trade-offs in terms of their risk to ethical principles, and making reasoned and documented decisions about which trade-offs to make. If no ethically acceptable trade-offs can be identified, the development, deployment, and use of the AI system should not proceed in that form. The decision-maker must be accountable and continually review the decision to ensure necessary changes can be made. Accessible mechanisms should be in place to provide adequate redress when unjust adverse impacts occur. This is important for building and maintaining trust in AI systems. Special attention should be given to vulnerable individuals or groups.	**Continuous monitoring and control**	- Do you continuously monitor the performance of the AI system ? What are the key metrics or indicators that you use to monitor? - Do you have a system in place to automatically detect and flag any anomalies or errors in the data or outputs of the AI system? If so, can you describe it? - Do you have a contingency plan in case the AI system malfunctions, and how is the risk of such a malfunction mitigated? - Do you have a retraining strategy to keep your models up-to-date when necessary (e.g. when metric thresholds are surpassed) ?	- Establish clear performance metrics (e.g. accuracy, response delay) and monitoring procedures (e.g. error analysis, data drify monitoring, user feedback) - Implement automated anomaly detection and error-flagging mechanisms - Develop a contingency plan for AI system malfunctions to mitigate risks - Develop a retraining strategy that incorporates clearly defined criteria such as metric thresholds and timelines	Prometheus/ Grafana	**	**	- Deployment preparation - Supervision (run)	
3		**Established governance model**	- Do you have a clear and transparent decision-making process for the AI system? - Do you have clear guidelines or policies in place for the use of the AI system? If so, who created them and how are they enforced? - Do you have a mechanism in place for stakeholders to raise concerns or provide feedback about the AI system, and how are these concerns addressed?	- Establish a clear and transparent decision-making process for the AI system - Create and enforce clear guidelines or policies for the use of the AI system - Implement a mechanism for stakeholders/the board of stakeholders' representatives to raise concerns or provide feedback, with a clear process how those concerns and feedbacks are handled		**	**	- Deployment preparation - Supervision (run)	
4		**Defined organizational roles and responsibilities**	- Do you have an established accountability for the actions of the AI system and those who build/operate it? How is this accountability enforced ? - Do you have a clear responsible for making AI system related decisions?	- Establish clear lines of accountability for the AI system - Create oversight mechanisms for the AI system to ensure that the actions of the AI system and those who build and operate it are held accountable - Assign decision-making responsibility to specific roles or individuals		**	**	- Definition of the use case - Deployment preparation	
5		**Adherence to purpose and values**	- Do you have a well defined purpose for the AI system? how does it align with the organization's mission and values? - Do you have a monitoring for AI system to ensure that it continues to adhere to the intended purpose and values? - Do you have measures in place to prevent the AI system from being used for purposes that do not align with the intended purpose and values? - Do you have a process for stakeholders to provide feedback on how the AI system adheres to purpose and values, and how is this feedback used to make improvements?	- Define the intended purpose of the AI system while making sure it adheres to the organization's mission and values - Implement a monitoring process to make sure that the AI system continues to adhere to the intended mission and values - Anticipate potential scenarios of misuse and implement measures to prevent them, while also regularly analyzing real usage. - Implement a mechanism for stakeholders to raise concerns or provide feedback about how the AI system adheres to purpose and values, with a clear process how those concerns are feedbacks are handled		***	***	- Definition of the use case - Monitoring (run)	
6		**Suitable for commercial purposes**	- Do you have a clear definition of the commercial application of the AI system ? - Have you performed a cost-benefit analysis for implementing the AI system in commercial applications, and how does this compare to alternative solutions? - How is the usability of the AI system optimized for commercial use, and what features are included to make it user-friendly for non-technical users?	- Define clearly the intended commercial application for the AI system and its business model - Conduct a cost-benefit analysis for the AI system in its ecosystem before implementation and review it regularly to adapt to real impact.- Optimize the usability of the AI system especially for non technical users		**	***	- Definition of the use case - Deployment preparation	to discuss
7		**Compliance with laws, regulations and policies**	- Are the legal and regulatory requirements that the AI system must comply with when used in commercial applications met? How are they met ? - Does the AI system comply with laws and regulations governing the industry or domain it operates in?	- Research legal and regulatory requirements and ensure compliance with requirements - Stay up-to-date with changes in legal and regulatory requirements and standardisation	Government websites, News and alerts, guidelines of local regulators, Impact analysis	*	*	- Definition of the use case - Prototyping and preparation of the first pilot - Deployment preparation - Supervision (run)	

Technical Framework v3.0

	A	B	C	D	E	F	G	H
1	**Principle**	**Strategy**	**Evaluation question**	**Corresponding action**	**Tools (illustrative, non-exhaustive)**	**Customer risk level minimum to implement the action**	**Employee risk level minimum to implement the action**	**Project phase for the implementation of the action**
2	Privacy is a fundamental right that is closely linked to the principle of preventing harm in AI systems. Adequate data governance is necessary to prevent harm to privacy, which includes ensuring the quality and integrity of data, its relevance to the domain, access protocols, and the capability to process data in a way that protects privacy. AI systems must ensure privacy and data protection throughout the entire lifecycle of the system. This includes user-provided information and information generated about the user over time. AI systems can generate outputs and recommendations that may allow them to infer personal characteristics such as sexual orientation, age, gender, religion, or political views. To build trust with users, it must be ensured that the data collected will not be used to discriminate against them unlawfully or unfairly. The quality and integrity of data used are critical to the performance of AI systems. Socially constructed biases, inaccuracies, errors, and mistakes in data must be addressed before training. The integrity of the data must be ensured to prevent malicious data from changing the behavior of AI systems. Each data set used in a step in the process of building an AI system, from planning to deployment, must be tested and documented, including for externally acquired AI systems. Organizations that handle individuals' data must establish data access protocols that outline who can access data and under what circumstances. Only qualified personnel with the necessary competence and need should be allowed to access individuals' data.	**Collection of data traceable to requirements**	Do you have a mapping between the collected datasets and the corresponding AI system requirements ?	- Conduct a comprehensive audit of the collected datasets and map the collected datasets to the AI system requirements - Maintain an updated data documentation GenAI specific : (Very) low transparency on data used for core-model pre-training Only possiblie action : Assess the credibility and "reputation" of the organizations providing the pre-trained LLMs	- Excel - Collibra - Apache Atlas - Model cards documentation framework	***	***	- Definition of the use case
3		**License of data**	-Do you have the proper licences for the used datasets ? Does the data set licensing allow for the intended use of the data in the AI system? - Do you comply to the requirements for attribution or citation that need to be met when using the data set in the AI system? how do you comply to the requirements ? - If applicable, do you have the specific data security and privacy measures required by the licensing implemented when using the data set in the AI system? - Are there any limitations on the distribution or sharing of the data set that would affect the AI system's development or deployment? how are these limitations met ?	- Conduct a thorough review of the data sets licensing, ensure proper attribution and citation - Implement appropriate data security and privacy measures and consider limitations in term of distribution and sharing of datasets GenAI specific : (Very) low transparency on data used for core-model pre-training Only possiblie action : Assess the credibility and "reputation" of the organizations providing the pre-trained LLMs		**	**	- Definition of the use case
4		**Protected from disclosure**	- Do you have measures in place to protect the data from unauthorized access, disclosure, or theft? what are these measures ? - Do you have policies and procedures in place to ensure the proper access levels,handling and disposal of the data? what are these policies and procedures ? - Are there any specific regulatory or legal requirements that must be followed to protect the data from disclosure? how are these requirements met ?	- Implement robust security measures to protect data from unauthorized access, disclosure, or theft - Develop policies and procedures for access management, incident response plans, risk assessement - Research specific legal and regulatory requirements for data protection and disclosure, and ensure compliance with requirements GenAI specific : - Add GenAI specific strategies to detect vulnerabilities such as prompt injections, backdoor attacks, and poisoning, which could lead to unauthorized disclosure - Establish guardrails to prevent LLMs from generating outputs that could lead to unauthorized disclosure of sensitive information (content scrubbing techniques...) - Prompt Eng : Design specific prompts that explicitly instruct the model to avoid generating sensitive information	- Cloud Security e.g AWS Security - ISO 27001 - giskard.ai (red teaming) - Presidio - OpenAI's API and Hugging Face's Transformers - custom prompt design	*	*	- Prototyping and preparation of the first pilot - Deployment preparation - Supervision (run)
5		**Approaches to privacy preservation**	- Do you have specific measures in place to ensure that individuals' privacy is protected throughout the AI system's lifecycle? what are these measures ? - Do you have any machine un-learning techniques or process in place to protect individuals' identities and personal information? (e.g de-identification, data pertubration, encryption) - Do you have a mechanism for obtaining individuals' consent before their data is used in the AI system if it is required? - Do you have a measure in place to implement the right to be forgotten?	- Conduct privacy impact assessments (PIAs) - Implement privacy by design and use de-identification techniques inorder to protect individuals' identities and personal information - Establish data retention policies - Examine whether an AI model can expose or be classified as personal data and determine the methods for safeguarding the model in such cases. - Implement a measure to put in place the right to be forgotten GenAI specific : If the model is trained with personnal chat with users : - Use tools to detect and anonymize PII in user inputs before they reach the model - Establish guardrails to prevent LLMs from generating outputs that could lead to unauthorized disclosure of sensitive user informations	- GDPR - Privacy enhancig technologies (differential privacy, homomorphic encryption, federated learning) - Data protection laws - DPA guidlines - Presidio, spaCy	*	*	- Definition of the use case - Prototyping and preparation of the first pilot - Deployment preparation - Supervision (run)

Technical Framework v3.0

	A	B	C	D	E	F	G	H
1	**Principle**	**Strategy**	**Evaluation question**	**Corresponding action**	**Tools (illustrative, non-exhaustive)**	**Customer risk level minimum to implement the action**	**Employee risk level minimum to implement the action**	**Project phase for the implementation of the action**
2	Technical robustness is important in achieving Trustworthy (Gen)AI, it enables developing (Gen)AI systems that prevent harm, behave reliably, minimize unintentional,unexpected and unacceptable harm. This should also consider potential changes in the system's operating environment and interactions with other agents, including humans and artificial agents, while ensuring the physical and mental integrity of humans. To ensure Trustworthy (Gen)AI, resilience to attack and security must be considered. (Gen)AI systems should be protected against vulnerabilities and attacks that can result in data poisoning, model leakage, prompt leaking or system behavior changes, which can lead to incorrect decisions or system shutdown. Insufficient security processes can result in erroneous decisions and physical harm. It's essential to consider unintended applications of the (Gen)AI system and potential abuse by malicious actors, and take steps to prevent and mitigate these risks. (Gen)AI systems must have fallback plans to enable safe operation in case of problems, such as switching from a statistical to a rule-based procedure or requiring human intervention. The system must be designed to prevent harm to living beings and the environment while minimizing unintended consequences and errors. Processes should be established to assess and clarify potential risks associated with (Gen)AI systems across various application areas. The level of safety measures required depends on the system's capabilities and the magnitude of the risk posed. If high risks are anticipated, safety measures should be developed and tested proactively. Accuracy in AI refers to the system's ability to make correct judgements, such as classifying information into the correct categories or making accurate predictions and decisions based on data or models. A well-formed development and evaluation process can help mitigate risks associated with inaccurate predictions, and it is important for the system to indicate how likely errors are when they occur. High accuracy is crucial in situations where AI systems impact human lives. In AI, reliability and reproducibility are important factors. A reliable system functions properly with various inputs and in various situations, which helps to prevent unintended harms. Reproducibility refers to the ability of an AI experiment to exhibit the same behavior when repeated under the same conditions, which enables accurate description of AI systems' behavior. Replication files can aid in testing and reproducing behaviors.	**Reliable performance**	- Do you have accuracy metrics for the (Gen)AI model to make sure it makes the right judgements and predictions? - Do you have an appropriate back-testing methodology closest to the model production environement? - Do the (Gen)AI model handle outliers and anomalies properly? - Are all model errors the same cost? (False positive VS false negative). If not, are you choosing the right metric? - Have you performed a benchmark to compare how does the (Gen)AI system perform compared to similar systems? - Have you monitoring of 3rd parties API in place to control their performance - Have you established the domain of model validity and tracked both optimal and unfavorable conditions? - LLMs: have you estimated the cost of an hallucination? (e.g. reputational risk, risk to business, etc)	- Define appropriate metrics for evaluating the accuracy of the (Gen)AI model (such as precision, recall, F1 score, or accuracy) - Build a proper backtesting methodology closest to the model production environement - Identify and define what constitutes an outlier or anomaly in the data and ensure that the AI model can detect and handle them appropriately, and develop a testing strategy to evaluate the model's performance on outlier and anomaly data. - Identify appropriate benchmarks for comparison, such as similar (Gen)AI systems or industry standards and develop a testing methodology to compare the (Gen)AI system's performance to these benchmarks - Consider and monitor any external dependencies to to 3rd parties model as a potential realiability issue and protect your own system from it - Define and document model validity domain and trace it's optimal and unfavorable conditions	- scikit-learn for evaluation metrics - matplotlib and seaborn for visualizing metrics - PyOD or Scikit-learn to detect and handle outliers - Imbalance-learn to handle class imbalanceness	**	**	- Prototyping and preparation of the first pilot - Deployment preparation - Supervision (run)
3		**Robustness by design**	- Is the (Gen)AI system reliable and reproducible in various situations, and what measures are taken to prevent unintended harms? - Do you ensure that most of the possible failure scenarios have been considered in your (Gen)AI system design? - Have you developed a contingency plan in the event of a failure or degraded mode? - Do you design your (Gen)AI system to handle unexpected or novel inputs, and train it on a diverse and representative dataset to avoid biases? - Do you test your (Gen)AI system on a range of inputs and conditions,and of environments, including those that are different from its training environment to ensure its robustness?	- Ensure that the (Gen)AI system is reliable and reproducible across various situations and contexts by testing it in different scenarios - Conduct a thorough analysis of potential failure scenarios and adapt the (Gen)AI system design to prevent unintended harms - Develop a contingency plan to handle events of failure - Use techniques such as data augmentation and adversarial training to improve the system's ability to handle novel inputs - Use techniques such as cross-validation and stress testing to evaluate the system's performance under different conditions and identify potential weaknesses	- scikit-learn - Failure mode and effects analysis FMEA - Chaos Monkey (by Netflix) - latticeflow	***	***	- Deployment preparation - Supervision (run)
4		**Identification and protection of security vulnerabilities**	- Do you (Gen)AI system design handle adversarial attacks or attempts to exploit vulnerabilities?	- Review the state of the art of the most common vulnerabilities and attacks against (Gen)AI systems (MITRE Atlas, OWASP TOP 10 (Including LLM TOP 10) - Conduct a thorough analysis of the (Gen)AI system to identify potential vulnerabilities and any associated IT resources - Implement security measures such as access controls (and data lookup control), encryption, and authentication to prevent unauthorized access to the (Gen)AI system - Implement monitoring of the complete solution to surface any performance and security risks - Use adversarial training techniques to improve the (Gen)AI system's ability to handle attacks and exploits.	- Burp Suite, Nessus or Kali Linux for identifying vunerabilities - RBAC : Role based access control - MFA: Multi factor authentification - HTTPS, SSL/TLS for encryption - IDS : intrusion detection systems - SIEM : security information and event management - IBM Adversarial Toolbox - ETSI Securing Artificial Intelligence Mitigation Strategy Report	*	*	- Definition of the use case - Prototyping and preparation of the first pilot - Deployment preparation - Supervision (run)