HOW TO CONTROL RISKS IN ARTIFICIAL INTELLIGENCE

Prof. Hernan Huwyler MBA CPA
ie Executive Education

RISK PROCESS FOR AI

DISCOVERY

- Identify all AI assets and libraries including shadow AI implementations
- Establish a searchable database of AI assets
- Foster a user community to promote collaboration and the adoption of AI best practices

MANAGEMENT

- Perform an initial risk assessment and prioritize AI assets for further review
- Identify opportunities for targeted governance to reduce risks, optimize costs, and enhance the value derived from AI resources

COMPLIANCE

- Develop a policies and procedures with AI-specific controls
- Incorporate AI applications into the IT, risk and compliance policies
- Train compliance with AI-related controls
- Test the AI-specific controls

AI Model Risk Management

DESIGN

- *Embed into decision-making: innovation plan, product approvals, third-party sourcing, and end-user computing of AI-based software*
- *Improve risk criteria to cover the materiality, customer and societal impacts and complexity*

IMPLEMENT

- *Assess data, algorithmic, performance, computational feasibility, and vendor risks*
- *Enhance controls for change management, targeted model reviews, and ongoing monitoring*

COMMUNICATE

- *Keep senior management updated on AI model development, review, and usage*
- *Report the testing and outcomes analysis for AI models*
- *Hold developers and model owners accountable for safe deploying*

OBJECTIVES AT RISK

DATA

- Training data
- Use data
- Feedback data
- Model code

SECURITY

- Deployment systems
- Core systems
- Infrastructure
- Edge systems

PERFORMANCE

- Third-parties
- Dependencies
- User experience

COMPLIANCE

- AI and privacy regulations
- Contracts
- Insurance

EXPLANABILITY

- Fairness
- Infrastructure
- Incorrect feedback

COSTS

- Budgeted resources
- Time objectives

RISK AND COMPLIANCE

US EXECUTIVE ORDERS 13859 ON AI AND 14028 ON CYBER SEC

- *Demonstrate controls on privacy, human rights, data, model and cyber security risks*
- *Model threats, attacks and surface to assess software risks (NIST 800 218 PW.1.1)*
- *Analyze vulnerabilities to gather risk data and plan responses (NIST 800 218 RV.2.1)*
- *Have a qualified independent person to review that the design addresses the identified risks (NIST 800 218 RW.2.1)*
- *Analyze the risk of applicable technology stacks*

EU ARTIFICIAL INTELLIGENCE LAW

- *Implement risk management system to proactively mitigate liabilities*
- *Implement granular scenario analysis to address AI-specific threats*
- *Evaluate the potential human rights impacts when AI interacts with users as intended by the model (bias, privacy, explainability, robustness)*

New Skills for Risk Managers

Practices

Understand

- *AI scope reviews*
- *Bias testing*
- *Explainability reports*
- *Data features and testing*
- *Statistical data checks*
- *Model output reviews*
- *Bayesian hypothesis testing*

Scenarios

Understand

- *AI specific threat vectors*
- *Performance versus interpretability trade-off*
- *Degradation and flagging*

Structure

Understand

- *Roles of data modelers, and analytics*
- *Model and use documentation*
- *Escalation mechanisms and workflows*
- *AI use cases*

AI

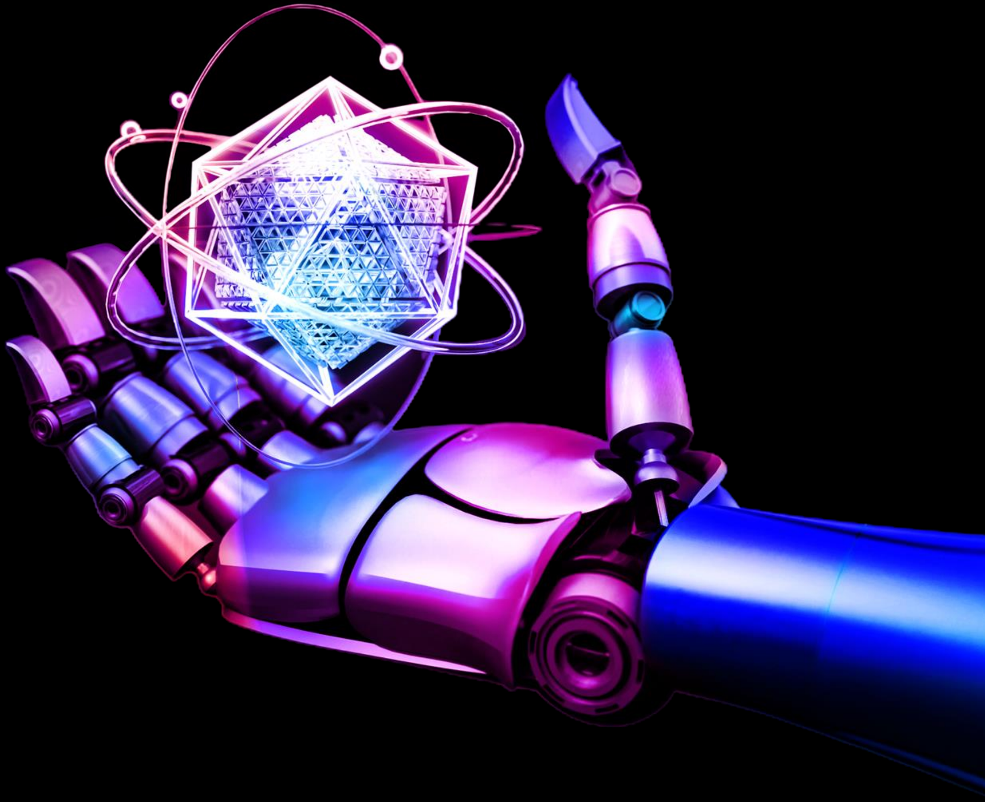
COMPUTER POWER + PROBABILITIES + DATA

Probabilistic model for predictions

Beyond machine learning

Curated
Classified
Complete
Clean
Consistent

DATA-RELATED RISKS



INADEQUATE DATA REPRESENTATIVENESS

RISK

Uneven training data from skewed sampling may lead to erroneous AI model predictions

FACTORS

- *Inadequate data preprocessing*
- *Biased data collection methods*
- *Unrepresentative training datasets*
- *Selection biases in sampling*
- *Data augmentation mishandling*
- *Lack of diversity in input sources*
- *Data drift over time*

CONTROLS

- *Verify data representativeness*
- *Conduct sensitivity analysis*
- *Implement advanced modeling techniques to mitigate selection bias*
- *Integrate fairness-aware AI algorithms to identify and rectify bias*

LOW DATA QUALITY

RISK

Incomplete, inconsistent, or incorrect data without detection processes may lead to inaccurate AI model predictions

FACTORS

- *Lack of data quality standards*
- *Absence of standardized data improvement processes*
- *Data privacy and ownership issues*
- *Weak data fitness for the use case*
- *High data customization*
- *Reuse of prior model-derived data*

CONTROLS

- *Establish data quality rules*
- *Enhance remediation processes*
- *Centralize data management*
- *Test the data formatting, metadata completeness, and indexing*
- *Address root causes of errors*

DATA SCARCITY

RISK

Inadequate selection, size and relevance of data sets may lead to inaccurate AI model predictions

FACTORS

- *Privacy, data source, and logistical constraints*
- *Limited available data due to the unique nature of the domain*
- *Lack of resources to process large volume of data*
- *Need for long-term historical data*
- *Use of third-party data*

CONTROLS

- *Define standard identifiers and consistent definitions*
- *Integrate physics-informed machine learning*
- *Review training, synthetic and augmented data for accuracy and relevance*

INSUFFICIENT EXTERNAL DATA QUALITY

RISK

A data vendor's inadequate lineage information, components, and processes may compromise data quality and model performance

CONTROLS

- *Enforce standards of model risk management*
- *Set service level agreements with provenance data*
- *Monitor aggregation*
- *Validate inputs and reliability*
- *Track response times*

FACTORS

- *Lack of vendor due diligence and requirements*
- *Insufficient quality control measures*
- *Data volume exceeds infrastructure capacity*
- *Inadequate response time optimization*

MODEL-RELATED RISKS

Misjudged AI Risk Ratings

Risk

Inadequate risk assessments on model pattern recognition, probability theory, and engineering may lead to vulnerabilities and implementation issues

Factors

- *Lack of traditional statistical foundation for AI modeling*
- *Complex AI model patterns*
- *Probability-based decision-making*
- *Model engineering nuances*

Controls

- *Integrate standards into modeling processes, approvals, third-party sourcing, and IT*
- *Maintain an AI model inventory with metadata*
- *Adapt design reviews to model-specific risks*

Missing Business Requirements

RISK

Inadequate review of use case, operationalization, and consumption the AI model development may overlook business requirements

FACTORS

- *Incomplete use case analysis*
- *Lack of operationalization clarity*
- *Insufficient consumption strategy*

CONTROLS

- *Assess AI's unique behavior potential beyond documented requirements*
- *Identify and address capability gaps across conceptualization, pilot, and operationalization phases before project initiation*

MODEL USAGE MISSUNDESTANDING

RISK

Inadequate understanding of intended uses of the model may hinder informed decision-making, leading to potential human right harm and compliance losses

FACTORS

- *Lack of context understanding*
- *Unrecognized impact on different groups*
- *Absence of continuous assessment plan*
- *Failure to consider the broader system impact*

CONTROLS

- *Perform impact assessments considering the context to address vulnerabilities and avoid unfair outputs*
- *Implement a continuous assessment plan throughout the model's lifecycle*
- *Consider the broader system impact during model integration*

UNIDENTIFIED MODEL LIMITS

RISK

Inadequate understanding of AI model boundary conditions may produce unintended and suboptimal outcomes

FACTORS

- *Neglecting system limitations awareness and boundary condition considerations*
- *Lack of ongoing monitoring*
- *Lack of interdisciplinary collaboration*
- *Ethical misalignment in using AI*

CONTROLS

- *Develop ongoing boundary testing processes during project framing*
- *Define monitoring procedures throughout all the lifecycle*
- *Form multi-disciplinary development teams*

Unidentified Model Limits

Challenges

- *Human decision-making may be shaped by a multitude of complex factors, including subjective experiences and individual interpretations*
- *Behaviors may defy predictability owing to the presence of free will, cognitive biases, and the impact of unique and unforeseen stimuli*
- *Behavioral data may engender concerns regarding ethical, compliance, and privacy considerations*
- *Behaviors may exhibit substantial cultural and contextual variations, rendering generalized predictions challenging*
- *Predictive models may struggle to extrapolate behavior patterns consistently across diverse individuals*
- *Behaviors may be dynamic and can undergo significant transformations over time*

Model Security Gaps

RISK

Inadequate security requirements may expose AI models to adversarial attacks and loss of data integrity and availability

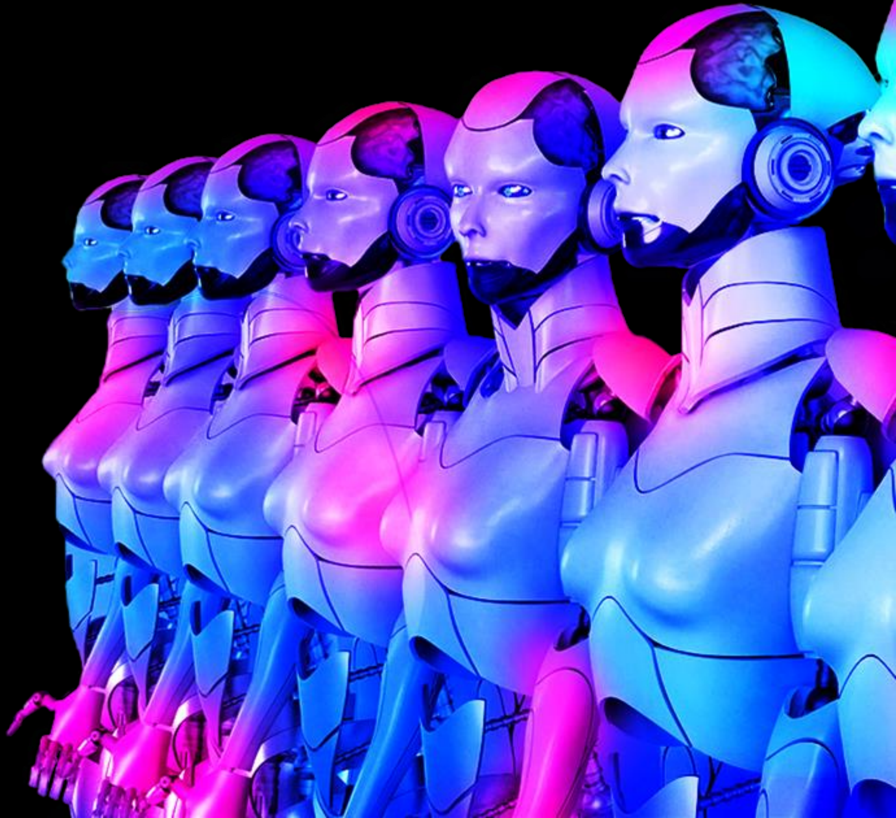
FACTORS

- *Incomplete security specifications*
- *Lack of adversarial robustness testing*
- *Neglected threat modeling*

CONTROLS

- *Enhance security program with AI considerations*
- *Address potential system compromise methods*
- *Assess dynamic AI-related risks*
- *Integrate information security into risk management policies*

Algorithm-Related Risks



ETHICS OVERSIGHT GAPS

RISK

Insufficient algorithm ethics checks lead to discrimination and unreliable models unfit for predictions

FACTORS

- *Lack of defined AI Ethics Governance to address ethical considerations in training data*
- *Inadequate controls in algorithm development*
- *Neglecting hidden biases in model outputs.*
- *Ethical misalignment with agents' preferences*
- *Absence of AI ethics council inspections*

CONTROLS

- *Enhance algorithm development controls to address ethics in training data and mitigate hidden biases*
- *Inspect algorithms through an AI ethics council*
- *Combine game theory and machine learning*

Non compliance

RISK

Limited insight into decision-making processes may harbor hidden biases and threaten the model credibility

FACTORS

- *Lack of embedded compliance checks*
- *Insufficient model validation for regulatory requirements*
- *Lack of transparency in AI decision-making*
- *Incomplete understanding of decision processes*
- *Unexamined societal biases in data and models*

CONTROLS

- *Strengthen model validation, emphasizing transparent decision understanding*
- *Conduct extensive pre-deployment testing of models*
- *Enhance algorithm validation to meet ethical AI standardsand mitigate societal biases in data and models*

BLACK BOX

RISK

Inscrutable machine learning algorithms may prevent human understanding of the decision-making processes

FACTORS

- *Model bias and errors*
- *Lack of transparency and accountability of business and technical owners and stakeholders*
- *Ethical and regulatory concerns*

CONTROLS

- *Embed black box testing in the model lifecycle to explain, debug, and improve models for stakeholders*
- *Require explanations from model owners*
- *Operationalize tools for audits and monitoring model impact on humans*

Algorithm Aversion

Challenges

- *End users may exhibit skepticism when it comes to placing trust in AI algorithm-generated decisions, even when these decisions surpass human judgment*
- *End users may harbor doubts about algorithmic recommendations when they lack clarity about the decision-making process*
- *End users may view AI algorithms as dehumanizing, which can result in a sense of diminished control or the loss of a personal touch*
- *End users may not rely on recommendations when there isn't a readily identifiable entity to hold accountable for those recommendations*
- *End users may be concerned about the potential for algorithmic discrimination or harm*
- *End users may prefer for their own judgments, which can lead to cognitive dissonance when algorithms offer disparate recommendations*

MODEL INSTABILITY

RISK

Changing relationships between the input features, the data and the target variables (concept and data drifts) may lead to deteriorating model accuracy over time

CONTROLS

- *Embed model stability checks in the model lifecycle*
- *Check for data distribution changes and input-output relationship shifts*
- *Conduct sensitivity analysis and scenario testing to enhance model robustness, accuracy, and understanding*

FACTORS

- *Changes in data distribution*
- *Altered input-output relationships*
- *Multicollinearity-induced parameter instability*
- *High redundancy in the model structure focused on improving the performance*

Model Misselection

RISK

Inadequate model selection may result in suboptimal AI performance and inaccurate decisions

FACTORS

- *Poor candidate model generation and development*
- *Neglect of key parameters*
- *Insufficient data and considerations for analysis*
- *Inadequate system understanding*

CONTROLS

- *Tailor model selection to the specific context*
- *Form diverse review teams for model categorization*
- *Ensure models are fit-for-purpose, explainable, reproducible, and robust*

Missing Checks

Risk

Failure to validate and cross-validate the model features may compromise the integrity and lead to data-driven errors

Factors

- *Weak governance tools for model monitoring, explainability, and biases*
- *Omission of feature validation*
- *Lack of cross-validation*
- *Limited stakeholder involvement*

Controls

- *Integrate reasonability, accuracy checks and cross-validation checks into risk management*
- *Embed checks throughout the model lifecycle*
- *Engage business and technical stakeholders in issue mitigation*

OVERFITTING AND UNDERFITTING

RISK

Excessively complex or simplistic modelling unable to capture good patterns in the training data may limit the prediction accuracy

FACTORS

- *Inadequate model complexity control*
- *Limited data diversity*
- *Inadequate feature selection*
- *Data noise and anomalies*

CONTROLS

- *Define and train risk management procedures for fit risks*
- *Distinguish data issues from fit issues*
- *Continuously assess fit's impact on outputs and equity gaps*

Suboptimal Hyperparameter Configuration

Risk

Inadequate hyperparameter configurations may impair AI model performance, effectiveness and reliability

Factors

- *Complex hyperparameter tuning*
- *Poor calibration understanding*
- *Lack of documentation*
- *Neglecting continuous assessment*
- *Failure to capture decision evidence*

Controls

- *Incorporate hyperparameter specifications and calibration into risk and impact assessments*
- *Document software package choices and specific values*
- *Continuously evaluate and adapt hyperparameters*

SUBOPTIMAL DIMRED

RISK

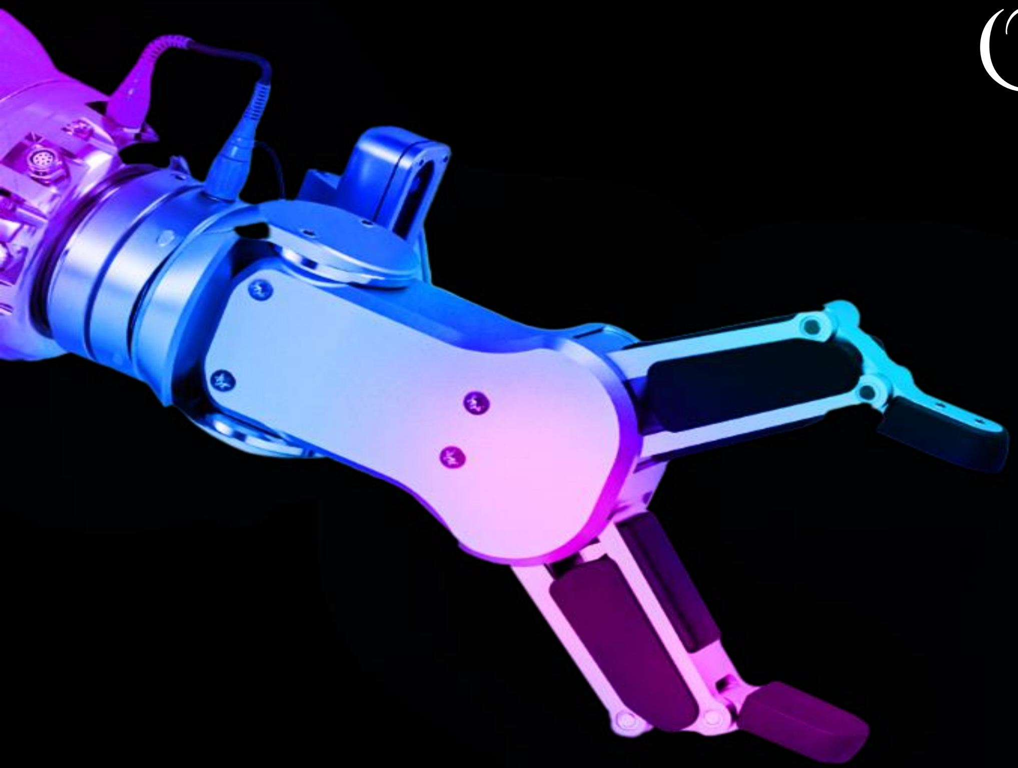
Dimensionality reduction, such as feature selection and extraction, may hinder interpretability

FACTORS

- *Inadequate dimensionality reduction techniques*
- *Incorrect feature selection*
- *Poorly chosen feature extraction methods*
- *Lack of interpretability in models*

CONTROLS

- *Use linear discriminant analysis, distributed stochastic neighbor embedding and auto encoders*
- *Optimize dimensionality reduction*
- *Justify and document the chosen dimensionality reduction approach*



OPERATION-RELATED RISKS

AI POLICY GAPS

RISK

Inadequate AI policies and procedures may lead to unmanaged risks in AI systems, hindering their benefits and exposing to unforeseen challenges

CONTROLS

- *Develop AI governance, infrastructure and use policies with articulated roles for model developers, users, and validators*
- *Include indirect AI influences in supporting technology policies*
- *Promote AI knowledge sharing in broader organizational functions*

FACTORS

- *Absence of specific AI governance policies*
- *Neglect of indirect AI technology influences in policies*
- *Lack of infrastructure support policies for AI at scale*
- *Insufficient consideration of AI knowledge in supporting non-AI functions*

THIRD-PARTY FAILURE

RISK

Inadequate third-party components and vendors may compromise software integrity, reliability, security, and performance

FACTORS

- *Lack of visibility into and security vulnerabilities of open source and commercial components*
- *Incomplete or outdated software inventory, source components and licenses*
- *Inadequate prioritization of third-party vulnerability mitigation*

CONTROLS

- *Create and maintain a detailed software Bill of Materials*
- *Monitor external threats and vulnerability disclosures*
- *Prioritize vulnerability mitigation based on risk assessments*
- *Perform due diligence audits in vendors*

UNATTAINABLE SCALABILITY

RISK

The utilization of actual data, users, and customers in deploying the AI solution may degrade the performance

CONTROLS

- *Monitor the performance to address slowdowns*
- *Ensure AI system's scalability with appropriate servers, computer power and communication*
- *Automate model updates when metrics deviate from performance objectives*

FACTORS

- *Inadequate infrastructure for scaling*
- *Increased computing power demands*
- *Model performance degradation*
- *Delayed model updates and approvals*

BIG O NOTATION IN SCALABILITY

Assesses the efficiency of searching, sorting, data manipulation, and other tasks within an AI algorithm. It quantifies efficiency by considering the relationship between processing time and data usage in proportion to the size of the input data.

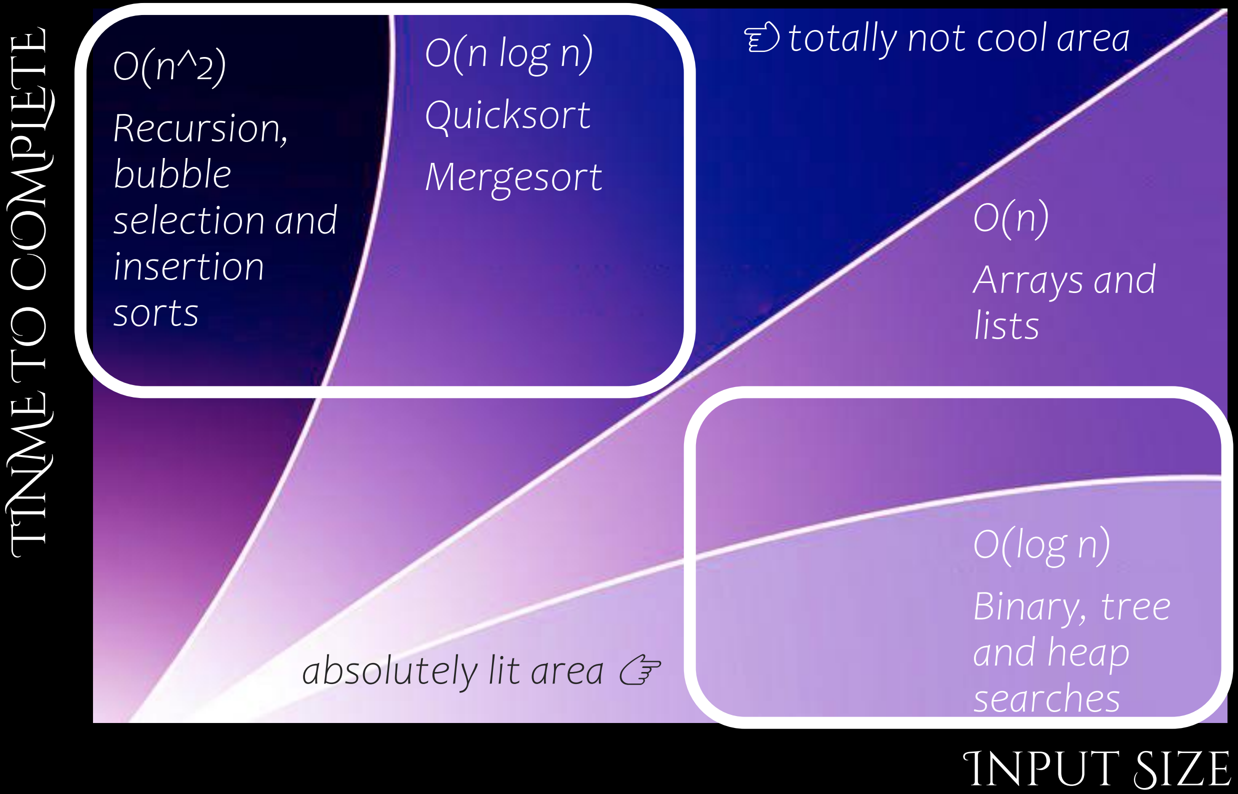
CONTROLS

- *Assess and compare the trade-off between efficiency and complexity in the given use case*
- *Enhance the AI algorithm for scalability and efficient memory, arrays and resource allocation*
- *Ensure that the computation time meets end-users' expectations*

PROCESSES

- *Performance impact assessments for different alternatives*
- *Benchmarking pre and post-changes*
- *Selection of data structures*
- *Quick, merge and heap sorting selection*
- *Linear, binary and tree searching selection*

BIG O NOTATION IN SCALABILITY



Infrastructure Malfunction

RISK

Infrastructure malfunctions stemming from diverse factors may result in software performance degradation or data loss

FACTORS

- *Inadequate IT support*
- *Outdated AI knowledge*
- *Undefined roles and responsibilities*

CONTROLS

- *Establish expert IT support for AI systems*
- *Ensure ongoing AI training for IT teams*
- *Define clear roles and responsibilities*
- *Monitor the operation of supporting infrastructure*

Insecure edge systems

RISK

Insecure edge hardware may be compromised affecting the data residing at the edge

FACTORS

- *Physical components in edge devices*
- *Data and model theft risks*
- *Interconnected systems vulnerability*
- *Overlapping edge uses*

CONTROLS

- *Implement remote disablement methods*
- *Apply data and model obfuscation techniques*
- *Isolate critical assets among systems*
- *Minimize redundancy in edge use cases*

FAILED ACCESS CONTROLS

RISK

Granting excessive permissions in the AI environments may expose to attack vectors leading to data breaches and security incidents

FACTORS

- *Inadequate permission governance tools, in particular for cloud-services*
- *Permissions mischaracterized as misconfigurations*
- *Anomalous activity monitoring gaps*
- *Unauthorized users with excessive permissions*
- *Overly broad permission grants*

CONTROLS

- *Detect and rectify excessive permissions, including mischaracterized misconfigurations*
- *Continuously monitor for anomalous activities*
- *Minimize the gap between granted and used permissions*

INADEQUATE FALLBACK SYSTEMS

RISK

Weak backup and restoration processes may affect the resilience and security of the AI application

FACTORS

- *Lack of backup systems, and contingency plans*
- *Incomplete risk-tiering*
- *Insufficient data encryption for backups*
- *Unassigned responsibilities*

CONTROLS

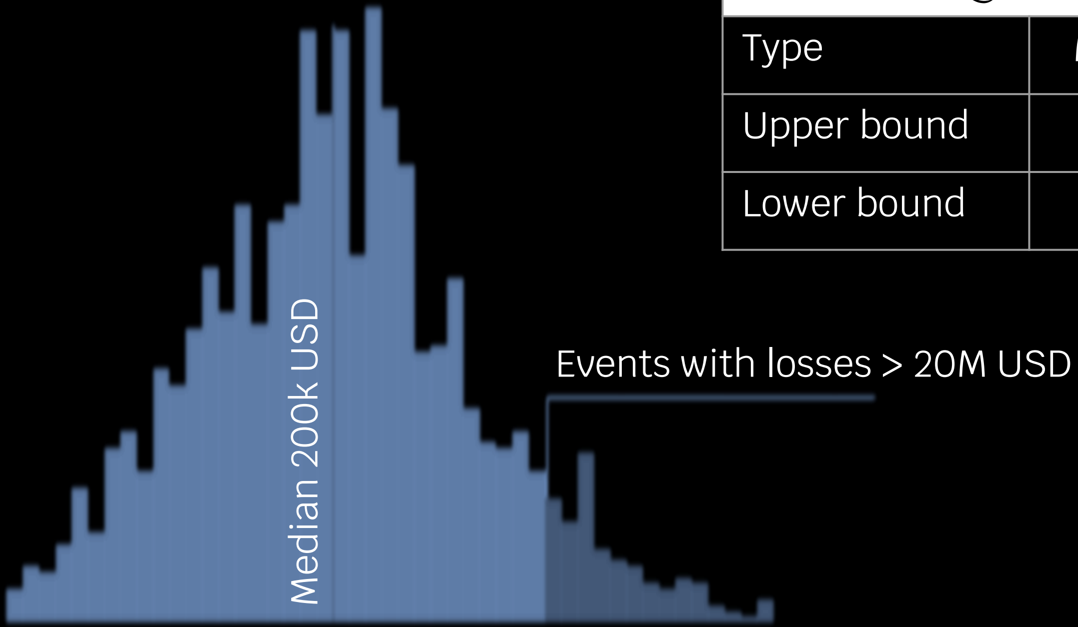
- *Establish a comprehensive fallback plan with responsibilities*
- *Define clear risk triggers and tiering and contingency plans*
- *Implement varied data backup schedules*

QUANTIFYING AI RISKS

WHAT IS THE PROBABILITY OF A?	P(A) PROBABILITY
HOW MUCH MORE LIKELY IS A THAN B?	PDF (A) / PDF (B) PROBABILITY DENSITY FUNCTIONS
HOW A IS PARAMETRIZED?	MLA (A) & MAP (A) MAXIMUM LIKELIHOOD ESTIMATION MAXIMUM A POSTERIORI
HOW A IS APROXIMATED?	RV (A) & CLT (A) RANDOM VARIABLE (MONTE CARLO SIM) CENTRAL LIMIT THEOREM
IS A SIGNIFICANT?	P VALUES (A) BOOTSTRAP RESAMPLING,

LOSSES FROM CYBER INCIDENTS

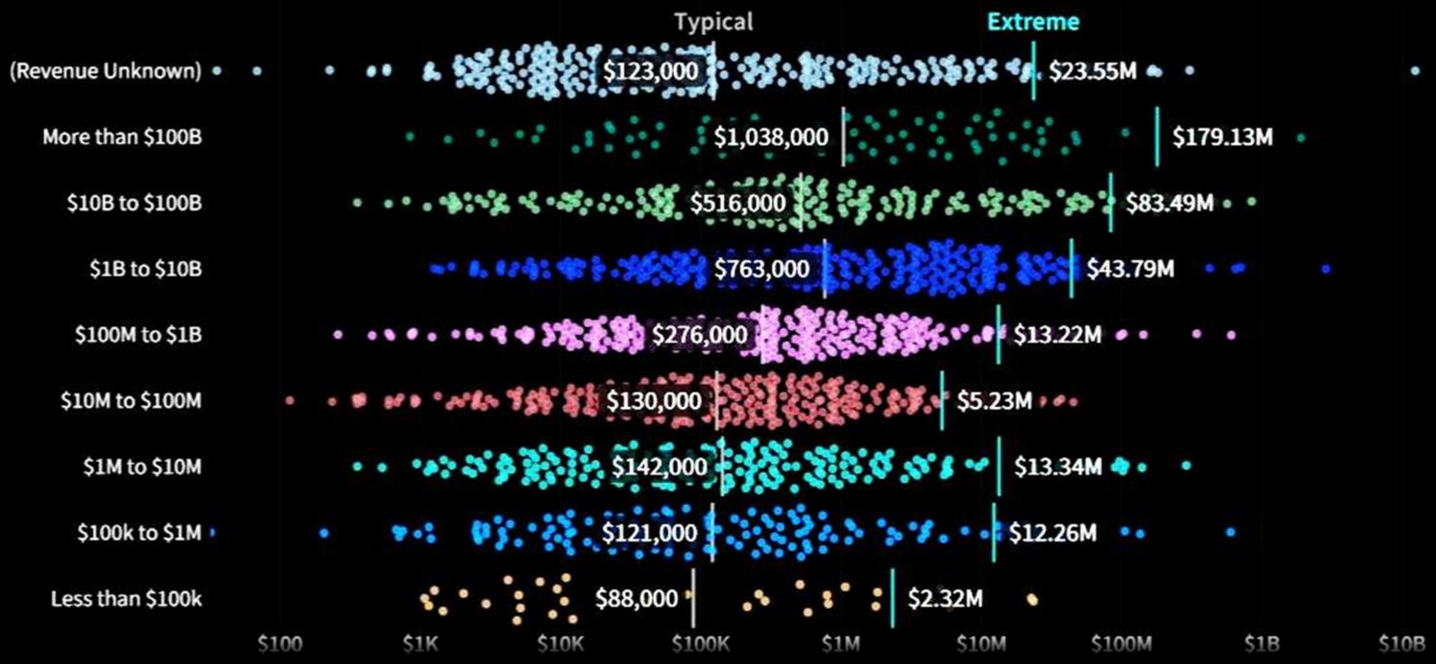
LOGNORMAL DISTRIBUTION FIT



Information Risk Insights Study 2017-2022

EVENT FREQUENCY MODEL PARAMETERS		
Type	Mean μ	St deviation σ
Upper bound	-2.28	.87
Lower bound	-6.39	1.78

LOSSES FROM CYBER INCIDENTS



IRIS 2022 Distribution of reported cyber event losses by company revenue

CONFIDENTIALITY	INTEGRITY	AVAILABILITY
Disclosure: employee or consultant may expose the model or training data through either negligence or intention **Oracle**: attacker may input the model to analyze its outputs, aiming to reverse engineer and extract either the model itself or the training data	**Poisoning**: attacker may manipulate the training data, aiming to distort outputs, introduce backdoors, or sabotage the model, causing it to produce inaccurate predictions or classifications **Evasion:** attacker may modify a minor portion of the training data to trigger significant changes in outputs, thereby influencing decisions made by the model **Bias:** developer may use incomplete training data leading to discriminatory outcomes	**Denial of service**: attacker may flood the system with numerous requests or data to overwhelm its capacity and slow down service consumption **Perturbation:** attacker may input data to exploit the model's fragility, intentionally inducing errors or unexpected behavior **Abuse:** user may misuse the model for fraudulent or unethical purposes **Third party:** vendor may fail to deliver expected supporting services

QUANTIFICATION

PERFORMANCE

Loss of revenue
Loss productivity hours
Cost of changes

FAIRNESS

Compensations
Costs of rebuilding
Costs of remediation

PRIVACY

Penalties for privacy laws
Compensations
Costs of notifications
Legal fees

ROBUSTNESS

Costs of response, restoration and remediation
Loss of competitiveness (IPs)
Loss of fraud

REGULATIONS

Penalties for AI laws
Compensations
Legal fees
Overcosts of customer acquisition

EXPLAINABILITY

Costs of wrong decision-making
Compensations
Costs of reprogramming

AGGREGATION

NORMAL

$X1, \ldots, Xn \sim \text{Normal}(\mu, \sigma^2)$
$X1 + X2 \sim \text{Normal}(2\mu, 2\sigma^2)$

Combined impact or exposure of a specific number of Normal-distributed events when both X1 and X2 are considered together

POISSON

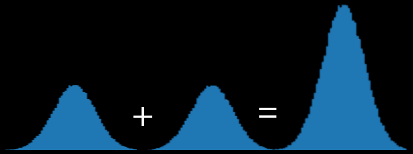
$X1, X2, \ldots, Xn \sim \text{Poisson}(\lambda)$
$X1 + X2 \sim \text{Poisson}(2\lambda)$

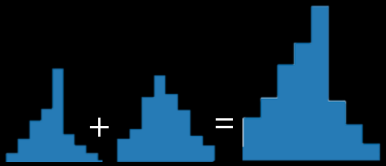
Likelihood of observing a specific number of Poisson-distributed events when both X1 and X2 are considered together

BINOMIAL

$X1 \sim \text{Binomial}(n_1, p)$
$X2 \sim \text{Binomial}(n_2, p)$

Likelihood of observing a specific number of independent events when both X1 and X2 are considered together





LAW OF TOTAL EXPECTATION

Calculate the expected loss average of a random variable by considering all possible loss values and their associated probabilities

$$E(X)=E(E(X|Y))=\sum E(X|Y=y)\cdot P(Y=y)$$

$E(X)$ represents the expected value of the random variable X
$E(X|Y)$ represents the expected value of X given a specific value of the random variable Y
$E(E(X|Y))$ means taking the expected value of $E(X|Y)$ over all possible values of Y

Decompose independent losses to be able to aggregate a total exposure

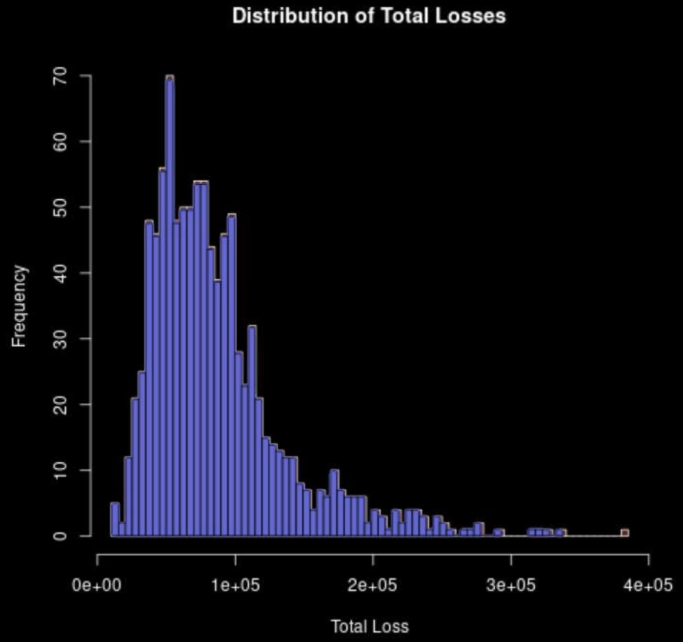
EXAMPLE IN R

```
P_B <- 0.05                            # 5% probability of a security incident on the IA model causing a data breach
mean_compensation <- 50000 # Mean compensation costs (customer compensations, legal and notification costs)
sd_compensation <- 2               # Standard deviation of compensation cost
mean_regeneration <- 20000   # Mean data regeneration costs (model rebuilding and data regeneration costs)
sd_regeneration <- 2                # Standard deviation of data regeneration cost

# Generate random samples from lognormal distributions for compensation and regeneration costs
compensation_samples <- rlnorm(1000, log(mean_compensation), log(sd_compensation))
regeneration_samples <- rlnorm(1000, log(mean_regeneration), log(sd_regeneration))

# Calculate the expected losses using the Law of Total Expectation
Expected_Cost_Compensation <- P_B * mean(compensation_samples)
Expected_Cost_Regeneration <- P_B * mean(regeneration_samples)
Expected_Loss <- Expected_Cost_Compensation + Expected_Cost_Regeneration

# Print the results
cat("The expected compensation cost of a data breach is $", Expected_Cost_Compensation, "\n")
cat("The expected data regeneration cost of a data breach is $", Expected_Cost_Regeneration, "\n")
cat("The total expected loss of a data breach is $", Expected_Loss, "\n")
```

```
The expected compensation cost of a data breach is $ 3195.646
The expected data regeneration cost of a data breach is $ 1291.59
The total expected loss of a data breach is $ 4487.237
```

EXAMPLE IN R

```
samples <- regeneration_samples + compensation_samples
hist(samples, breaks = 100, main = "Distribution of Total Losses",
     xlab = "Total Loss", ylab = "Frequency", col = "lightblue")
```

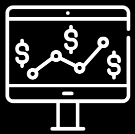


COUNTING

Experiment ⇨ *Outcome*

Risk scenario
- Event A Failed — *P(A)*
- Event B Succeeded — *P(B)*

Model attack
- Event A Failed — *P(A)= 95%*
- Event B Succeeded — *P(B)= 5%*
  - Compensations
  - Regeneration



How many options?

Product Rule of Counting

If an experiment has two independent parts, where the first part can result in one of |m| outcomes and the second part can result in one of |n| outcomes, then the total number of outcomes for the experiment is |m| * |n|.

|m| * |n| = 2 * 2 = 4

(Fail, 0,0)
(Succ, Comp, Rege)
(Succ, 0, Rege)
(Succ, Comp, 0)

PROBABILITY

$$P(E) = \lim_{n \to \infty} \frac{n(E)}{n}$$

n

# of total trials

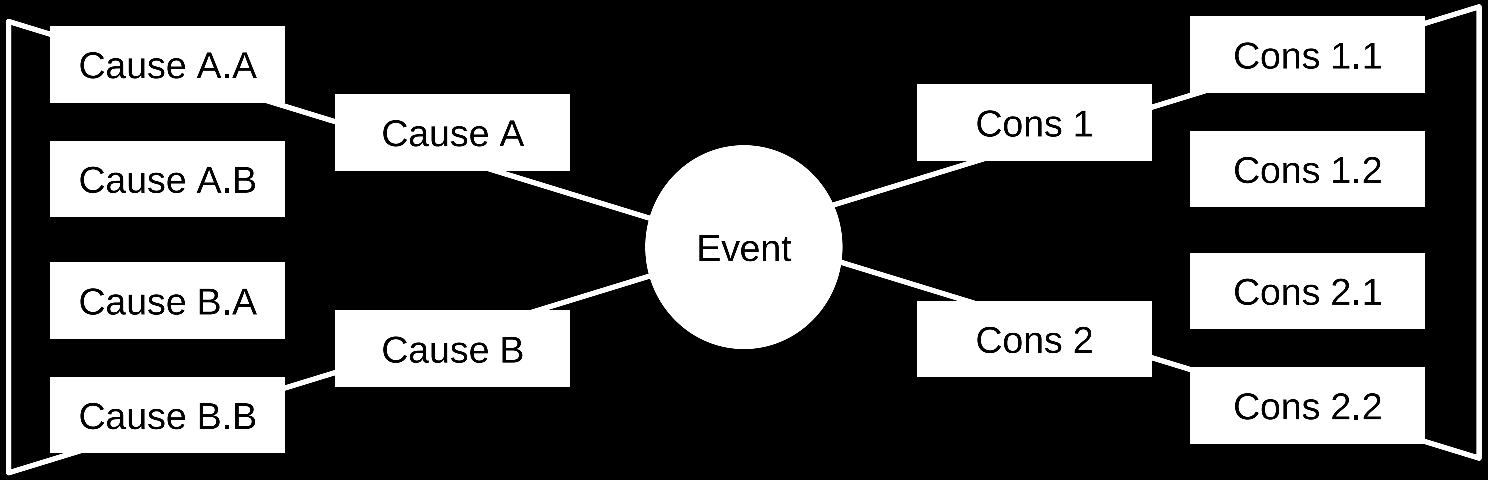
$n(E)$

trials where E occurs

Our belief that an event E occurs

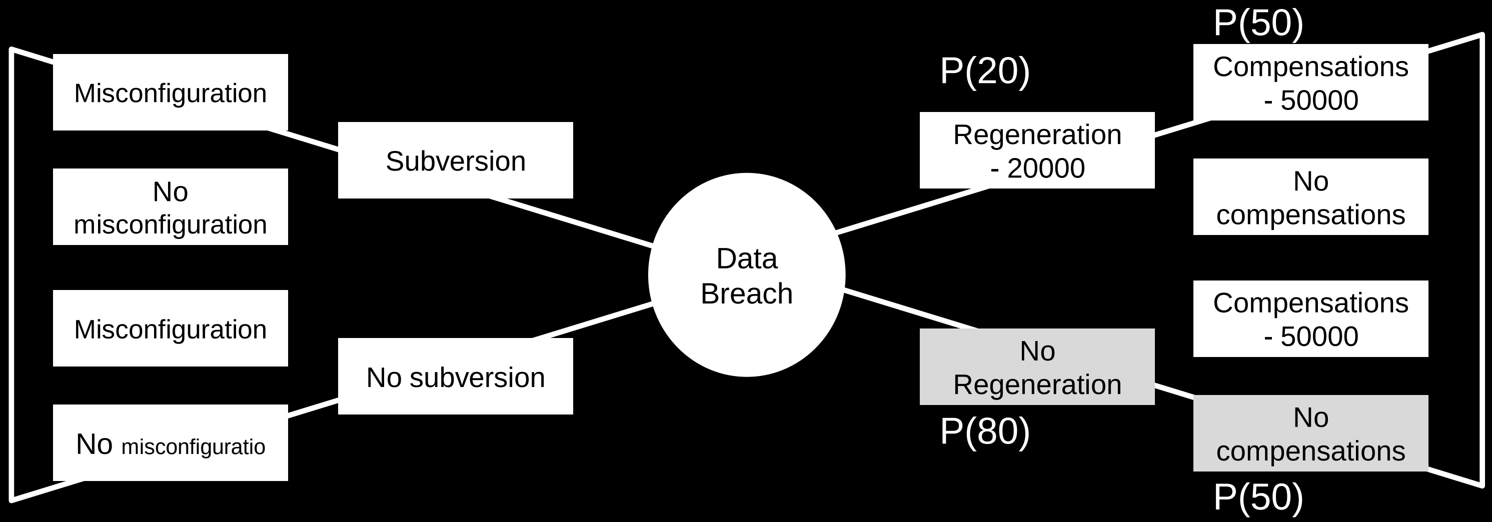
- *quantitative way to express our degree of belief or confidence in the occurrence of an event*
- *number between 0 and 1 to which we ascribe meaning*
- *represented through probability distributions, which describe how probabilities are distributed among different possible outcomes*

COMBINATORICS & BOW TIE



Tier 1 = 2 outcomes (direct or primary impact)
Tier 2 = 4 outcomes (indirect or secondary impact)
Tier 3 = 8 outcomes
$|A| = m$, $|B| = n$, $A \cap B = \emptyset$
Tier n = 2^n = **exponential growth!**

COMBINATORICS & BOW TIE

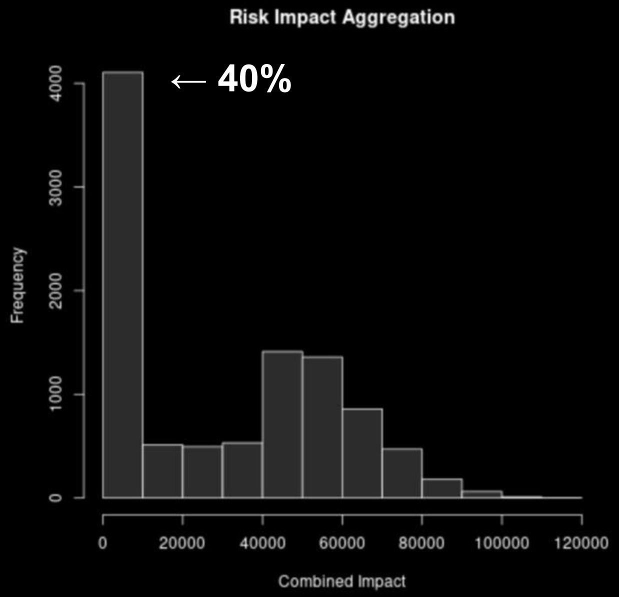


What is the probability of avoiding data regeneration and compensation costs?

80%*50%= 40%

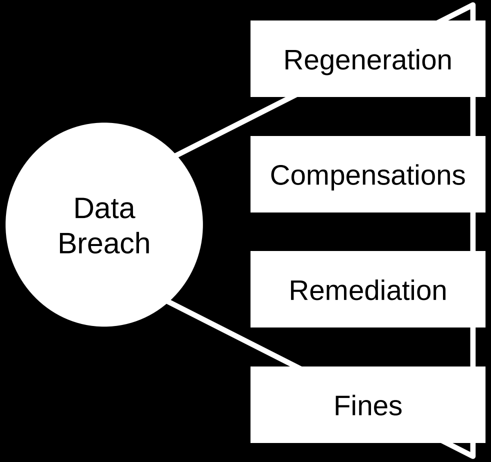
R CODE

```
# Set the parameters for risk assessment
Simulations <- 10000  # Number of random simulations
Loss1 <- 20000           # Estimated loss of the tier 1 impact
StDev1 <- 0.2              # Estimated standard deviation 1
Prob1 <- 0.2 # Probability of occurrence of the tier 1 impact
Loss2 <- 50000           # Estimated loss of the tier 2 impact
StDev2 <- 0.2              # Estimated standard deviation 2
Prob2 <- 0.5 # Probability of occurrence of the tier 2 impact

# Calculate and combine impacts
x1 <- rlnorm(Simulations, log(Loss1), StDev1) *
rbinom(Simulations, 1, Prob1)
x2 <- rlnorm(Simulations, log(Loss2), StDev2) *
rbinom(Simulations, 1, Prob2)
hist(x1 + x2, main="Risk Impact Aggregation",
xlab="Combined Impact")
```



Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	0	30777	29888	53441	126206

COMBINATORICS & IMPACTS



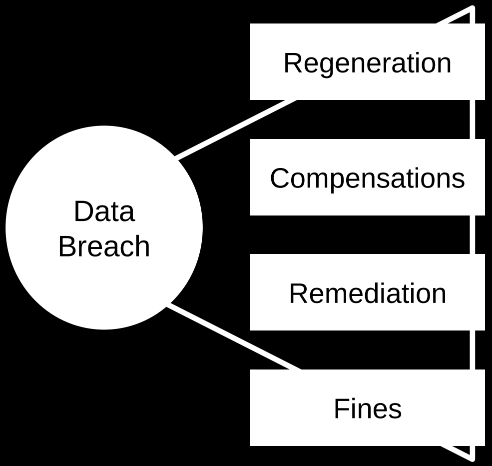
Product rule of counting

List of combinations of potential impacts without repetition

R code

```
impacts <- c("Regeneration", "Compensations",
"Remediation", "Fines")
all_combinations <- list()

for (n_impacts in 1:4) {
  combinations <- combn(impacts, n_impacts, simplify =
FALSE)
  all_combinations[[as.character(n_impacts)]] <- combinations
}
all_combinations
```

COMBINATORICS & IMPACTS




```
$`1`[[1]] "Regeneration"
$`1`[[2]] "Compensations"
$`1`[[3]] "Remediation"
$`1`[[4]] "Fines"
$`2`[[1]] "Regeneration“ "Compensations"
$`2`[[2]] "Regeneration" "Remediation"
$`2`[[3]] "Regeneration" "Fines"
$`2`[[4]] "Compensations" "Remediation"
$`2`[[5]] "Compensations" "Fines"
$`2`[[6]] "Remediation" "Fines"
$`3`[[1]] "Regeneration"  "Compensations" "Remediation"
$`3`[[2]] "Regeneration"  "Compensations" "Fines"
$`3`[[3]] "Regeneration" "Remediation"  "Fines"
$`3`[[4]] "Compensations" "Remediation"   "Fines"
$`4`[[1]] "Regeneration“ "Compensations" "Remediation“ "Fines"
```

FAIRNESS RISK INDICATORS

Indicators associated with the risk for discrimination

Commonly used for eLending, eRecruiting, healthcare and criminal justice

STATISTICAL PARITY DIFFERENCE

- *Difference in the probability of a favorable outcome for different groups, often based on sensitive attributes like gender, race, or age*
- *SPD = |P(Y = 1 | D = privileged) - P(Y = 1 | D = underprivileged)|*

EQUAL OF OPPORTUNITY DIFFERENCE

- *Difference in the probability of a favorable label, specifically assessing whether the fraction of actual positives correctly classified is similar across all groups*
- *EOD = |TPR(D = privileged) - TPR(D = underprivileged)|*

FAIRNESS RISK INDICATORS

EQUAL PREDICTIVE PERFORMANCE

- *Difference in the probability of errors in the accuracy, precision, recall, and F1 scores across all groups*
- *EPP = |P(Y = 1 | D = privileged) - P(Y = 1 | D = underprivileged)|*
- *Y= (Number of Correct Predictions) / (Total Number of Predictions)*

EQUAL OUTCOMES

- *Difference in the probability of getting equal outcomes for individuals or groups regardless race, gender, age and any demographics*
- *EO = actual outcomes achieved by different groups*

HELLINGER DISTANCE

- *Difference of the distribution of predicted outcomes for different groups*
- *H(Group 1, Group 2)=1 / sqrt(2) * sqrt(sum((sqrt(p_i) - sqrt(q_i))^2))*

ROBUSTNESS RISK INDICATORS

STABILITY	• *Difference in the predictions when the input data changes or when the model is subjected to noise and perturbations in real word*
TREE DISTANCE	• *Difference in the number of insertions, deletions, and modifications required to transform one hierarchical structure into another* • *Assess the stability of predictions are when dealing with variations in the hierarchical structure of input data*
OUT-OF-DISTRIBUTION	• *Difference between the input data and the training data* • *Patterns that the model has never encountered during training* • *OoD = uncertainty estimates and anomaly detection methods*

EXPLAINABILITY RISK INDICATORS

SURROGACY EFFICACY SCORE

- *Measures how well complex input-output relationships as a black box can be deducted and explained using decision trees and linear regression in an auxiliary surrogate model*
- *SEC = accuracy and R-squared for regression for model and surrogate*

α-FEATURE IMPORTANCE

- *Measures how well individual features can explain the predictions by introducing a parameter (α) to control the balance between individual feature importance and feature interactions*

USER SATISFACTION

- *Measures how well the explanations meet the end-user expectations and needs for transparency*
- *Interviews and surveys to end-users or stakeholders to gather their feedback on the quality and effectiveness of AI model explanation*

SECURITY RISK INDICATORS

FALSE REJECTION RATE

- *Rate at which legitimate users are incorrectly denied access when they should have been accepted in biometric systems*
- *FRR = Number of False Rejections / Total Legitimate Access Requests*

FALSE ACCEPTANCE RATE

- *Rate at which unauthorized users are incorrectly granted access when they should have been rejected in biometric systems*
- *FAR = Number of False Acceptance/Total Unauthorized Access Requests*

PRIVACY RISK INDICATORS

ANONYMITY SET SIZE

- *Amount of individuals that an AI model is unable to identify*
- *Quantify how many individuals are protected from identification and re-identification*

ENTROPY

- *Amount of uncertainty introduced in the data to protect individual privacy*
- *Measures the level of privacy protection by assessing the level of information dispersion in the data*

INFORMATION LEAKAGE RATE

- *Amount of sensitive or private information disclosed by the model's output in relation to other units of information*

All Risk Indicators

SAFE FRAMEWORK

- *Sustainable indicators measure the energy consumption, workload, and operational costs*
- *Accurate indicators measure the prediction quality*
- *Fair indicators measure the*
- *Explainable indicators measure the prediction quality for rankings and human-understandable insights.*

KEY ARTIFICIAL INTELLIGENCE RISK INDICATORS

Sustainability	How secure an AI model works for robustness and stability Tests to evaluate accuracy in handling extreme values and data manipulation and to optimize the model while preserving simplicity
Accuracy	How well an AI model predicts things compared to what actually happens Tests to ascertain differences in predictions between various models and to compare the consistency of predictions
Fairness	How an AI model don't favor one group over another Tests to compare the distribution of variables across different population groups and identify biases
Explainability	How an AI model can be explained to stakeholders Tests to evaluate how to interpret the outputs and behaviors by assigning contributions to each predictor.

SAFE TESTS FOR AI RISK MEASURE

Sustainability	Likelihood > F test for regression and X2 test for classification
	Accuracy > Diebold test for regression and DeLong test for classification
Accuracy	Root Mean Square Error > Diebold-Mariano test
	Area Under the Receiver Operating Characteristic > DeLong test
Fairness	Gini on estimated parameters > KS Kolmogorov-Smirnov test
	Gini on Shapey Values > KS Kolmogorov-Smirnov test
Explainability	Estimated Parameters > T test
	Shapley Values > T test

AI RISK IDENTIFICATION

SCOPE THE AI LIFE CYCLE

DESIGN

- *Use case for automatization*
- *Marketing plan*
- *Financial model*
- *Supply chain*
- *Feature intelligence*
- *Capability assessment*

BUY

- *Data gathering*
- *Third-party components*
- *Third-party software*
- *Vendor management*

DEVELOP

- *Environment setting*
- *Warehouse Configuration*
- *Modeling*
- *Training*
- *Refinement*
- *Testing*
- *Optimization*

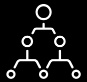
DEPLOY

- *Handoff*
- *Feedback*
- *Maintenance*
- *Post market evaluation*

RISK OWNERS

SPONSOR

- Set requirements
- Approve budget
- Monitor objectives
- Ensure testing and documentation



ARCHITECT

- Design and optimize high level features
- Decide the technology stack, tools, frameworks, and patterns
- Integrate solutions

ENGINEER

- Design and implement specific components
- Code, test and debug the model
- Prepare data
- Train the model

EXPERT

- Support in a domain area such as data ethics and compliance
- Facilitate solving issues
  Contribute to testing

HALF OF THE PROJECT TIMELINE IS TYPICALLY DEDICATED TO DATA PREPARATION

PREPARATION

- *Define the volume and quality of the data internally available*
- *Procure external data*
- *Clean and normalize*
- *Tools: Google Cloud Data Preparation, Hadoop, Alpine Miner*

MODELING

- *Determine the structure of the data and the analytic techniques*
- *Identify data from various sources.*

BULDING

- *Transform and label datasets*
- *Develop datasets for training, testing, and production*
- *Develop models on training data and test-on-test data*

TIPS TO IDENTIFY SCENARIOS

- *Assess the validity and accuracy of the model on the test data to ensure its generalization capabilities*
- *Assess whether the model's output and behavior align with the expectations and insights of domain experts*
- *Assess the reasonableness of parameter values within the context of the specific domain*
- *Assess whether the model sufficiently and accurately achieves the defined goals and objectives*
- *Assess the mechanisms in place to avoid intolerable mistakes or inaccuracies in its predictions*
- *Assess if additional data or inputs are required to enhance the model's performance and robustness*
- *Assess that the chosen model type is compatible with the expected run-time environment, considering factors such as speed and resource requirements*
- *Assess whether a different model may be necessary to effectively address the specific business problem or if adjustments to the existing model are required for optimal performance*

DATA QUALITY RISKS

Controls to identify data quality issues

- *Bivariate analysis: assess the relation between two variables*
- *Chi-Square test: assess the difference between the expected frequencies and the observed frequencies in one or more categories*
- *Z or T test: assess averages the difference between of two groups*
- *Analysis of variance: assess the difference between the average values of three or more independent groups*
- *Assess missing, null values, spaces in data sets*
- *Assess inconsistencies and outliners in data sets*

Controls to correct quality issues

- *Substitute values*
- *Segment data sets*
- *MinMaxScaler: transform the data to a standard range to ensure that all features have a consistent scale (often between 0 and 1)*
- *StandardScaler: transforms the data to have a mean of 0 and a standard deviation of 1 to ensure normal distribution without outliners*
- *Normalizer: transform the data to have a vector length of 1*

AI IMPACT ASSESSMENT

GOALS

Ethics by design principle

Identify early warnings of possible ethical and compliance issues

Assess appropriate controls on potential sources of bias, privacy , manipulation, dignity and security risks

Before the start of a life cycle stage when changes can be easily done

CONTEXT

Understand and describe

- *The benefit of the AI features for the endusers and stakeholders*
- *The technology, complexity, social and values context*
- *The planned new features or updates in a new release*
- *The integration of the AI system with other systems or products*
- *The intended users, sectors, and geographies*

AI IMPACT ASSESSMENT

LIFE CYCLE SCOPE

- *Planning stage: assess the feasibility risks for the goals, scope, and project specifications of the AI project*
- *Design stage: assess the architectural and system design based on the project specifications*
- *Development stage: assess the development and integration risks in the AI models, algorithms, and software components*
- *Testing stage: assess risks involved in testing the functionality, performance, and reliability of the AI system*
- *Deployment stage: assess risks related to the integration of the AI system with existing productive systems*
- *Maintenance stage: assess the risks associated with updates, improvements, and optimizations*
- *Retirements stage: assess the risks related to obsolescence and replacement, considering changes in business requirements*

CONTROLS ON AI
ISO 42001

AI GOVERNANCE CONTROLS

- *AI policy: The organization shall document a policy for the development or use of AI systems*
- *Alignment with other organizational policies: The organization shall determine where other policies can be affected by or apply to the organization's objectives with respect to AI systems*
- *Review of the AI policy: The AI policy shall be reviewed at planned intervals or additionally as needed to ensure its continuing suitability, adequacy, and effectiveness*
- *AI roles and responsibilities: Roles and responsibilities for AI shall be defined and allocated according to the needs of the organization.*
- *Reporting of concerns: The organization shall define and put in place a process to report concerns about the organization's role with respect to an AI system throughout its life cycle.*

IMPLEMENTATION

General recommendation

- *Regularly review and update the AI policy to ensure its continuing suitability, adequacy, and effectiveness in aligning with organizational objectives and managing AI-related risks effectively*

Practical guidance

- *Conduct a thorough analysis to determine intersections between AI policies and other organizational policies, updating them as necessary to ensure coherence and alignment.*
- *Designate a role approved by management responsible for the development, review, and evaluation of the AI policy, incorporating feedback from management reviews*
- *Define roles and responsibilities for AI within the organization, considering AI policies, objectives, and identified risks to ensure accountability throughout risk management, asset management, security, safety, privacy, development, performance, human oversight, supplier relationships, and legal requirements fulfillment*

RESOURCE ALLOCATION

- *Resource documentation: The organization shall identify and document relevant resources required for the activities at given AI system life cycle stages and other AI-related activities relevant for the organization.*
- *Data resources: As part of resource identification, the organization shall document information about the data resources utilized for the AI system.*
- *Tooling resources: As part of resource identification, the organization shall document information about the tooling resources utilized for the AI system.*
- *ystem and computing resources: As part of resource identification, the organization shall document information about the system and computing resources utilized for the AI system.*
- *Human resources: As part of resource identification, the organization shall document information about the human resources and their competences utilized for the development, deployment, operation, change management, maintenance, transfer and decommissioning, as well as verification and integration of the AI system*

IMPLEMENTATION

General recommendation

- *Assess resources required for AI system activities and other AI-related tasks to comprehensively understand and address risks and impacts*

Practical guidance

- *Ensure documentation of resources includes data resources, tooling resources, system and computing resources, and human resources, encompassing roles and competences necessary for the development, deployment, operation, maintenance, and integration of the AI system*
- *Consider diverse expertise and roles necessary for the system, including demographic groups related to data sets, to ensure inclusivity and effectiveness in system design and operation*
- *Recognize that different resources may be required at various stages of the AI system life cycle, and continually assess and adapt resource needs to support ongoing improvement and optimization of AI systems*

Impact Assessments

- *AI system impact assessment: The organization shall establish a process to assess the potential consequences for individuals or groups of individuals, or both, and societies that can result from the AI system throughout its life cycle*
- *Documentation of AI system impact assessments: The organization shall document the results of AI system impact assessments and retain results for a defined period*
- *Assessing AI system impact on individuals or groups of individuals: The organization shall assess and document the potential impacts of AI systems to individuals or groups of individuals throughout the system's life cycle*
- *Assessing societal impacts of AI systems: The organization shall assess and document the potential societal impacts of their AI systems throughout their life cycle*

IMPLEMENTATION

General recommendation

- *Establish a comprehensive process to assess the potential impacts of AI systems on individuals, groups, and societies throughout the system's life cycle*

Practical guidance

- *Consider the intended purpose and use of AI systems when assessing potential impacts on individuals, groups, and societies affected by the system*
- *Incorporate elements such as identification, analysis, evaluation, treatment, and documentation into the AI system impact assessment process*
- *Define circumstances under which an impact assessment should be performed, considering factors like criticality of purpose, complexity of technology, and sensitivity of data types*
- *Involve relevant stakeholders, experts, and users in the impact assessment process to obtain a comprehensive understanding of potential impacts*

IMPLEMENTATION

- *Document the results of AI system impact assessments and retain them for a defined period, considering legal requirements and organization retention schedules.*
- *Assess impacts on various aspects including fairness, accountability, transparency, security, privacy, safety, health, financial consequences, accessibility, and human rights*
- *Evaluate societal impacts considering environmental sustainability, economic factors, government processes, health and safety, cultural norms, traditions, and values*
- *Analyze potential misuse of AI systems and develop strategies to mitigate societal harms and reinforce positive impacts*
- *Consider both positive and negative outcomes when assessing impacts, particularly in scenarios involving health, safety, and societal well-being*
- *Continually update and refine impact assessments throughout the AI system's life cycle to address evolving risks and challenges*

LIFE CYCLE CONTROLS

Management guidance for AI system development

- *Objectives for responsible development of AI systems: The organization shall identify and document objectives to guide the responsible development of AI systems, and take those objectives into account and integrate measures to achieve them in the development life cycle.*
- *Processes for responsible AI system design and development: The organization shall define and document the specific processes for the responsible design and development of the AI system.*

AI system life cycle

- *AI system requirements and specification: The organization shall specify and document requirements for new AI systems or material enhancements to existing systems*
- *Documentation of AI system design and development: The organization shall document the AI system design and development based on organizational objectives, documented requirements, and specification criteria*

LIFE CYCLE CONTROLS

- *AI system verification and validation: The organization shall define and document verification and validation measures for the AI system and specify criteria for their use.*
- *AI system deployment: The organization shall document a deployment plan and ensure that appropriate requirements are met prior to deployment*
- *AI system operation and monitoring: The organization shall define and document the necessary elements for the ongoing operation of the AI system, including system and performance monitoring, repairs, updates, and support*
- *AI system technical documentation: The organization shall determine what AI system technical documentation is needed for each relevant category of interested parties and provide the technical documentation to them in the appropriate form*
- *AI system recording of event logs: The organization shall determine at which phases of the AI system life cycle record-keeping of event logs should be enabled, but at the minimum when the AI system is in use*

IMPLEMENTATION

General recommendation

- *Ensure that the organization identifies and documents clear objectives for responsible development of AI systems, integrating these objectives into the development life cycle*

Practical guidance

- *Identify objectives that impact AI system design and development processes, incorporating them into various stages such as requirements specification, data acquisition, and model training*
- *Provide requirements and guidelines to ensure that measures for achieving objectives are integrated into the development process, such as specific testing tools or methods to address fairness or bias*
- *Consider utilizing AI techniques to augment security measures, reinforcing protection for both AI systems and conventional software systems against security attack*

AI DATA CONTROLS

- *Data for development and enhancement of AI system: The organization shall define, document, and implement data management processes related to the development of AI systems.*
- *Acquisition of data: The organization shall determine and document details about the acquisition and selection of the data used in AI systems.*
- *Quality of data for AI systems: The organization shall define and document requirements for data quality and ensure that data used to develop and operate the AI system meet those requirements.*
- *Data provenance: The organization shall define and document a process for recording the provenance of data used in its AI systems over the life cycles of the data and the AI system.*
- *Data preparation: The organization shall define and document its criteria for selecting data preparations and the data preparation methods to be used.*

IMPLEMENTATION

General recommendation

- *Ensure that the organization defines and implements data management processes for the development of AI systems, encompassing aspects such as privacy, security, transparency, and data quality*

Practical guidance

- *Consider privacy and security implications when using sensitive data, implementing measures to mitigate associated risks*
- *Ensure transparency and explainability by documenting data provenance and providing explanations of how data are used in determining AI system outputs*
- *Assess the representativeness, accuracy, and integrity of training data compared to the operational domain of use, addressing biases and ensuring suitability for the intended purpose*

IMPLEMENTATION

- *Define criteria for data acquisition, including categories, quantity, sources, and characteristics, and document details about data acquisition and use using established frameworks such as ISO/IEC 19944-1*
- *Define and document requirements for data quality, considering the impact of bias on system performance and fairness, and make necessary adjustments to improve performance and fairness*
- *Establish a process for recording data provenance throughout the data and AI system life cycles, considering factors such as data source, content, and context of use*
- *Define criteria for selecting data preparation methods and transforms, ensuring that data are properly prepared to increase quality and avoid errors in AI system outputs*

REPORT CONTROLS

- *System documentation and information for users: The organization shall determine and provide the necessary information to users of the AI system*
- *External reporting: The organization shall provide capabilities for interested parties to report adverse impacts of the AI system*
- *Communication of incidents: The organization shall determine and document a plan for communicating incidents to users of the AI system*
- *Information for interested parties: The organization shall determine and document its obligations to reporting information about the AI system to interested parties*

IMPLEMENTATION

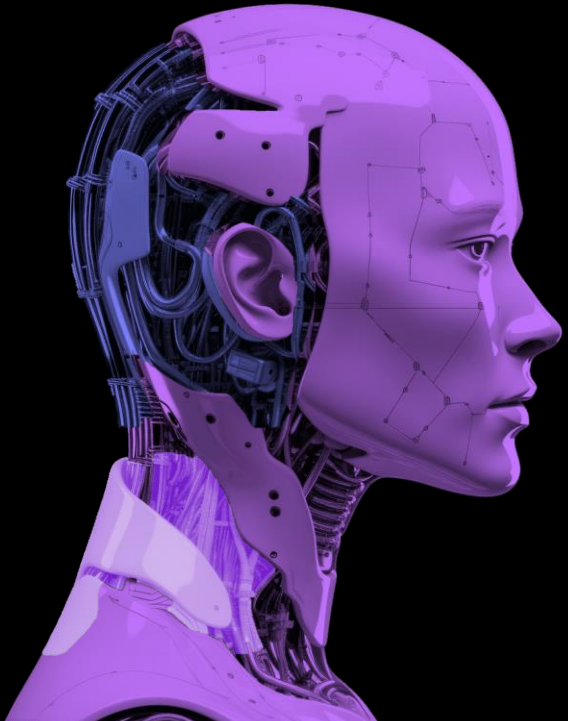
General recommendation

- *Ensure that relevant interested parties have access to comprehensive information about the AI system, including its purpose, operation, potential impacts (both positive and negative), and avenues for reporting adverse impacts or incidents*

Practical guidance

- *Provide users with clear and understandable information about the AI system, including technical details, instructions for interaction, and notifications about AI-generated outputs*
- *Tailor system documentation to the needs of different user groups, considering their technical expertise and specific requirements*
- *Make information accessible and easy to find, considering users' accessibility needs*
- *Document criteria for determining what information to provide, considering the intended use and potential impacts of the AI system*

CONTROLS ON AI USES



- *Processes for responsible use of AI systems: The organization shall define and document the processes for the responsible use of AI systems*
- *Objectives for responsible use of AI systems: The organization shall identify and document objectives to guide the responsible use of AI systems*
- *Intended use of the AI system: The organization shall ensure that the AI system is used according to the intended uses of the AI system and its accompanying documentation*

IMPLEMENTATION

General recommendation

- *Establish and document processes for the responsible use of AI systems to ensure alignment with organizational policies and objectives*

Practical guidance

- *Define and document processes for determining the suitability of using a particular AI system, considering factors such as required approvals, costs, legal requirements, and sourcing criteria*
- *Identify and document objectives to guide the responsible use of AI systems, considering factors such as fairness, accountability, transparency, reliability, safety, privacy, security, and accessibility*
- *Implement mechanisms to achieve these objectives, which may include incorporating human oversight at relevant stages of the AI system life cycle*
- *Ensure that human oversight activities are informed by AI system impact assessments and that personnel involved are adequately trained and informed of their duties*

Implementation

- *Deploy AI systems according to their intended uses and accompanying documentation, ensuring that resources, including human oversight, are provided as required*
- *Monitor the operation of AI systems and communicate concerns regarding their impact or compliance with legal requirements to relevant personnel and third-party suppliers*
- *Maintain event logs or other documentation related to the deployment and operation of AI systems to demonstrate adherence to intended use and facilitate communication of concerns*
- *Determine the retention period for event logs and documentation based on the intended use of the AI system, organizational data retention policies, and relevant legal requirements*

OBJECTIVES

Responsible AI

- *Fairness*
- *Accountability*
- *Transparency*
- *Explainability*
- *Reliability*
- *Safety*
- *Robustness and redundancy*
- *Privacy*
- *Security*
- *Accessibility*

THIRD-PARTY CONTROLS

- *Allocating responsibilities: The organization shall ensure that responsibilities within their AI system life cycle are allocated between the organization, its partners, suppliers, customers, and third parties.*
- *Suppliers: The organization shall establish a process to ensure that its usage of services, products, or materials provided by suppliers aligns with the organization's approach to the responsible development and use of AI systems.*
- *Customers: The organization shall ensure that its responsible approach to the development and use of AI systems considers their customer expectations and needs.*

IMPLEMENTATION

General recommendation

- *Ensure that the organization understands its responsibilities, remains accountable, and appropriately manages risks when third parties are involved at any stage of the AI system life cycle*

Practical guidance

- *Responsibilities within the AI system life cycle should be clearly allocated between the organization, its partners, suppliers, customers, and third parties*
- *Establish a process to ensure that the organization's usage of services, products, or materials provided by suppliers aligns with its approach to responsible development and use of AI system*
- *Understand customer expectations and needs when supplying products or services related to an AI system.*

FAIRNESS IN AI

EQUITY AND ETHICS RISKS

FAIRNESS

IMPORTANCE

- *AI systems should treat individuals and groups in a just, equitable, and unbiased manner*
- *Algorithmic systems can perpetuate unfair stereotypes and negative associations, leading to real harm and unequal access to opportunities.*

COMPLEXITY

- *Definitions of fairness can vary and conflict with one another, making it crucial to make a conscious choice*
- *Algorithmic fairness is highly context-dependent*

IMPLICATIONS

- *Algorithmic decision-making can increase consistency and reduce bias compared to individual judgments*
- *Ensures equitable treatment for all individuals*
- *Prevents negative consequences for marginalized groups*

DEFINITIONS

INDIVIDUAL FAIRNESS

- *Treating similar individuals similarly, based on relevant attributes*

GROUP FAIRNESS

- *Treating different groups to receive similar treatment or outcomes on average*

EQUALIZED ODDS

- *Achieving similar true positive and false positive rates across different groups*

DEMOGRAPHIC PARITY

- *Ensuring that the proportion of each group receiving a positive outcome is the same*

GROUP UNAWARE

- *Making decisions without access to sensitive attributes*

COUNTER FACTUAL

- *Considering hypothetical outcomes for different groups, else being equal*

RECOMMENDATIONS

- *Be aware that unfair biases and stereotypes can become embedded in AI systems, whether deliberately or accidentally*
- *When building AI decision-making tools, carefully consider upfront which specific definition and approach to fairness to adopt. Different technical approaches optimize for different notions of fairness*
- *Recognize that boosting one type of fairness often comes at the expense of other priorities like overall accuracy or efficiency*
- *Look to governments and civil society to help provide frameworks and best practices for navigating fairness tradeoffs, especially in high-stakes public sector applications*
- *Engage in discussions to clarify how factors like gender, race, age, and socioeconomic status should or should not be considered by algorithms in different contexts.*
- *Seek to establish some shared directional principles*
- *Address fairness requires ongoing ethical reasoning and a willingness to grapple with complex tradeoffs*

Non-Discrimination

- *Fairness is based on the belief that all human beings have equal moral status and deserve equal respect, concern, protection, and regard before the law*
- *Wrongful discrimination occurs when decisions, actions, institutional dynamics, or social structures do not respect the equal moral standing of individual persons*
- *Fairness involves the moral duty to treat others as moral equals and to secure the membership of all in a moral community where every person has equal value*
- *Discriminatory harassment > Unwanted or abusive behavior linked to a protected characteristic that violates someone's dignity, degrades their identity, or creates an offensive environment.*
- *Direct discrimination > Treating individuals adversely based on their membership in a protected class, also known as disparate treatment*
- *Indirect discrimination > Existing provisions, criteria, policies, arrangements, or practices that disparately harm or unfairly disadvantage members of a protected class, also known as disparate impact*

BIAS TYPES

- Historical Bias > *Be aware of differences between the current world and the values and objectives of your AI model. Ensure your model reflects the world you want to create, not just the one that exists*
- Representation Bias > *Make sure your model represents all groups in the population. Avoid under-representation or failure to generalize for certain groups.*
- Measurement Bias > *Choose features and labels that accurately reflect real-world quantities. Avoid using noisy proxies that can lead to biased results.*
- Aggregation Bias > *Be cautious when combining distinct groups into a single model. Ensure you're not masking important differences between groups. Develop separate models for distinct groups when necessary to ensure fair treatment and accurate representation of each group's unique characteristics.*
- Evaluation Bias > *Use performance metrics and testing benchmarks that accurately represent the entire population. Avoid using metrics that favor one group over another.*
- Deployment Bias > *Provide clear guidelines and training for the appropriate use and interpretation of AI models in real-world environments.*

PROTECTED CLASSES

PERSONAL

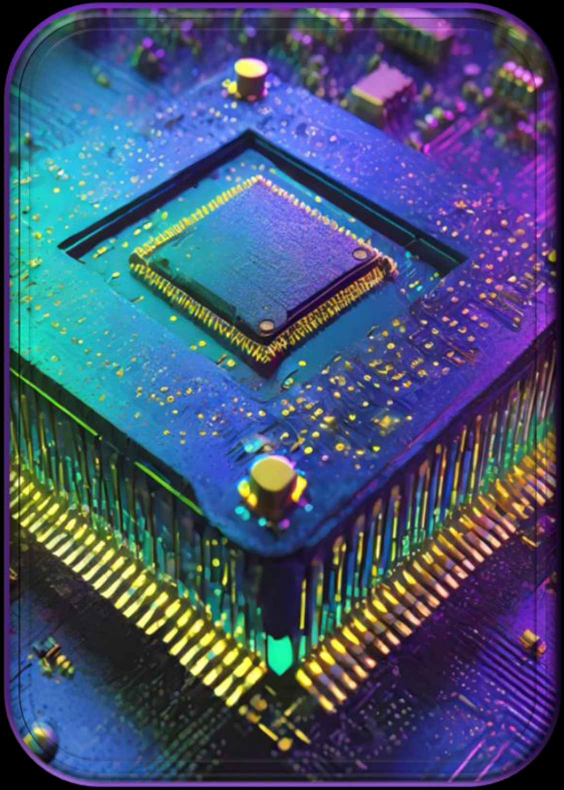
- *Age*
- *Gender*
- *Marital status*
- *Pregnancy or maternity leave*
- *Disability*
- *Sex*
- *Sexual orientation*
- *Medical conditions*
- *Criminal records*

IDENTIFY

- *Race, color, nationality. ethnic, and national origin*
- *Religion and beliefs*
- *Language*
- *Political and other opinions*
- *National and social origins*
- *Association with a national minority*

CONDITIONS

- *Socioeconomic status*
- *Property ownership*
- *Place of birth*
- *Crime victim status*

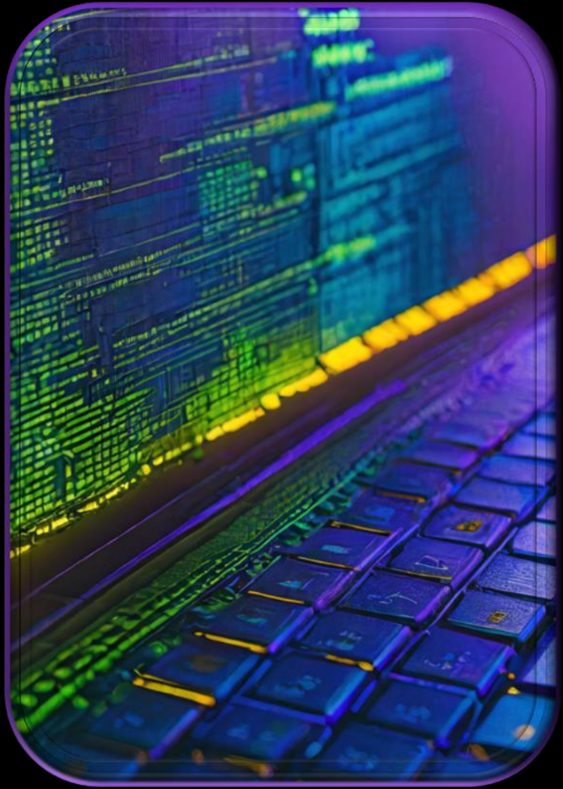


CLASSES

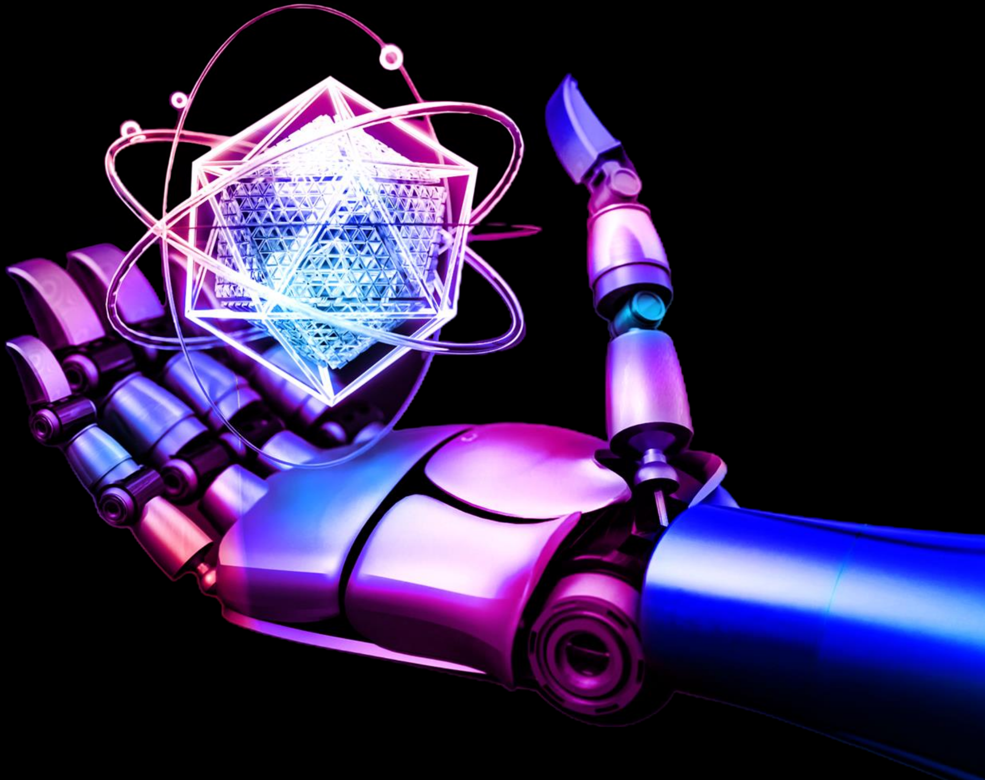
- *Data Fairness > Ensure datasets are properly representative, fit-for-purpose, relevant, accurately measured, and generalizable. Use data that is free from bias and accurately reflects the population being served.*
- *Application Fairness > Ensure policy objectives and agenda-setting priorities do not create or exacerbate inequity, structural discrimination, or systemic injustices. Ensure AI systems align with the aims, expectations, and sense of justice of impacted people.*
- *Model Design and Development Fairness > design models that do not include discriminatory variables, features, processes, or analytical structures. Ensure models do not encode social and historical patterns of discrimination.*

CLASSES

- Metric-Based Fairness > Establish lawful, clearly defined, and justifiable formal metrics of fairness. Make metrics transparently accessible to relevant stakeholders and impacted people.
- System Implementation Fairness > Ensure users are sufficiently trained to implement AI systems in a bias-aware manner. Ensure users understand the limitations and strengths of AI systems and deploy them with due regard to individual circumstances.
- Ecosystem Fairness > Ensure the wider economic, legal, cultural, and political structures or institutions do not entrench or amplify discriminatory power dynamics. Ensure policies, norms, and procedures promote equitable outcomes for protected, marginalized, vulnerable, or disadvantaged social groups.



TOOLS FAIR TESTTING

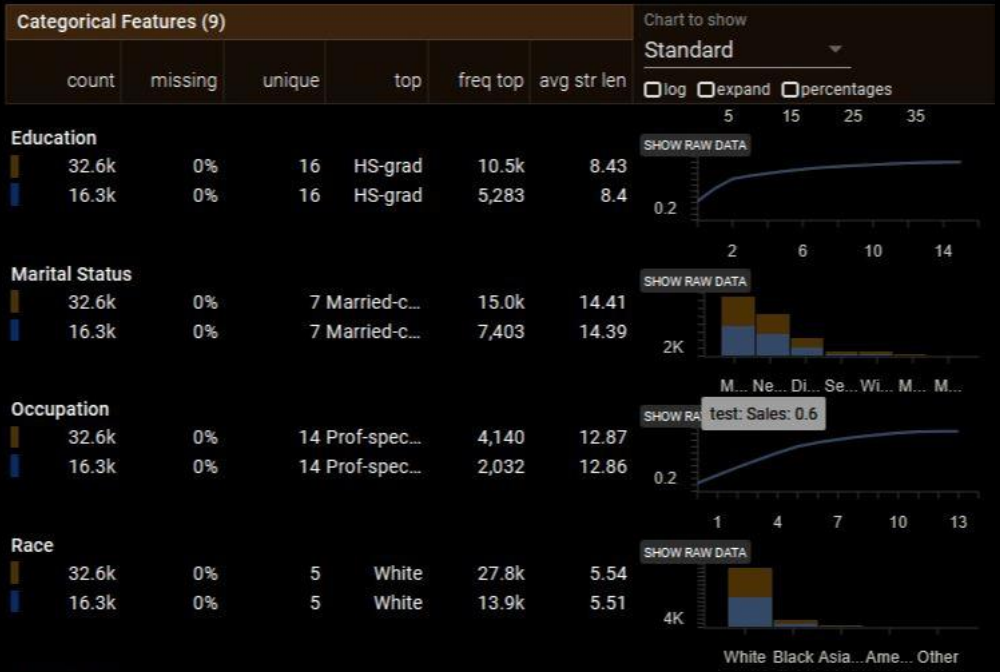


TOOLS

- *Facets Overview and Facets Dive to explore their datasets and identify potential sources of bias*
- TensorFlow What-If Tool *as a way to probe and understand models, and to take into account constraints such as fairness criteria.*
- *Model and Data Cards to document the models and datasets*
- *TensorFlow algorithms to train AI systems that satisfy fairness goals, and encourage students to experiment with these tools in their projects.*

Encourage ongoing monitoring and evaluation of models to ensure they continue to meet fairness goals as they are deployed in real-world contexts.

FACETS OVERVIEW

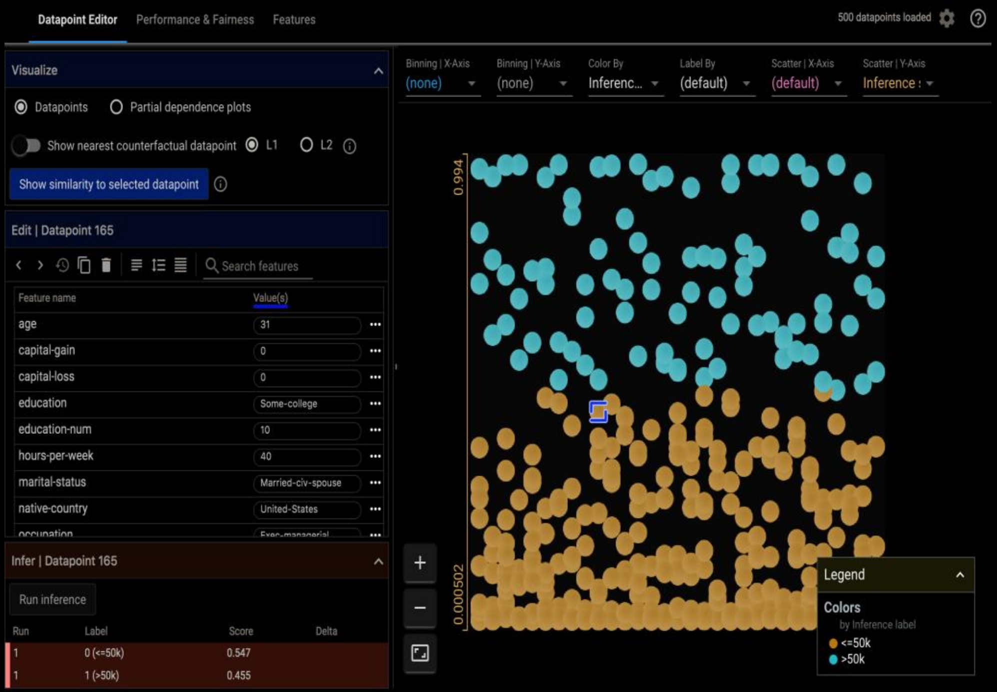


Visualization tool to analyze input feature data from multiple datasets and understand the distribution of values across features

It uncovers unexpected feature values, missing feature values, and skew between training, serving, and validation sets

It allows users to explore individual observations and get a deeper understanding of their dataset to mitigate the risk of bias

TENSORFLOW WHAT-IF



Visualization tool to probe and analyze machine learning models to assess their fairness across different subgroups and hypothetical scenarios to surfacing potential fairness issues

It connects to a model server and dataset in TensorBoard for exploration

DATA FAIRNESS

- *Ensure your training data is representative of the population the AI system will impact. Underrepresentation or overrepresentation of disadvantaged groups can lead to discriminatory outcomes*
- *Collect sufficient data to capture the diversity of attributes in the population being modeled. Insufficient data may not equitably reflect qualities that should rationally factor into the AI's decision*
- *Scrutinize data sources and measurement instruments for potential biases. Basing an AI system on data reflecting biased human decisions will replicate those biases in the AI's outputs*

DATA FAIRNESS

- *Use current, up-to-date data that reflects the present distribution of characteristics in the population. Outdated data may introduce bias as social relationships and group dynamics change over time*
- *Incorporate domain expertise to select the most relevant and appropriate data features as model inputs. This helps optimize the AI system's accuracy and robustness*
- *Maintain a comprehensive data factsheet throughout the AI development lifecycle. Systematically document key information on data provenance, preprocessing, potential bias issues identified, and remediation steps taken*
- *Foster close collaboration between domain experts and the technical team to inform responsible data practices. Diverse perspectives help proactively identify and mitigate risks of bias.*

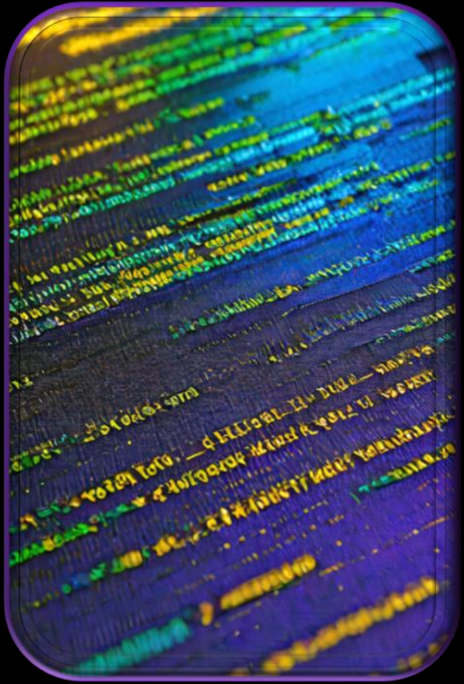
RECOMMENDATIONS

- *Policymakers and AI experts should collaborate to identify and address inadvertent harms that may arise from existing or proposed rules around fairness in AI*
- *Inferring sensitive attributes like race or gender can be essential for assessing the fairness of AI systems*
- *AI has the potential to surface and mitigate existing human and societal biases. By analyzing connections between input data and output predictions, AI can help identify embedded biases in current decision-making processes.*
- *If AI reveals that certain biases are unmerited, organizations should take steps to adjust their practices and limit the effect of these biases*
- *Continuous monitoring and evaluation of AI systems in production is essential to detect and mitigate any fairness issues that may emerge over time as the AI is exposed to new data and scenarios*

Metrics

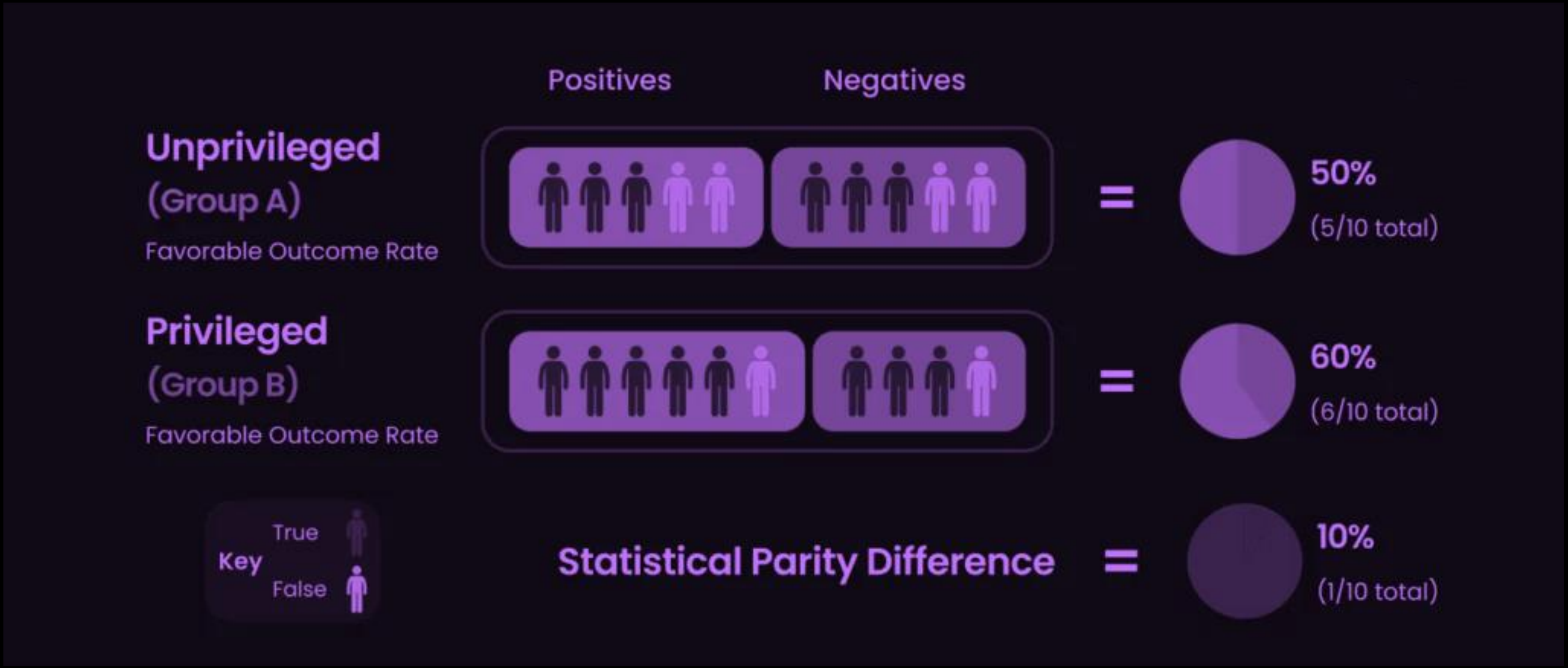
- *Individual Fairness > This approach judges fairness at an individual level rather than group level. It requires that similar individuals (based on a defined similarity metric) receive similar algorithmic outcomes. A challenge is agreeing on an appropriate similarity metric.*
- *Demographic Parity > An AI system satisfies this fairness criterion if each demographic group receives a positive outcome at equal rates. The goal is to prevent disparate impact, where certain groups are disproportionately harmed.*
- *Equalized Odds > Under this definition, an AI system is considered fair if both the true positive rates and false positive rates are equal across demographic groups. This aims to ensure parity in the AI's accuracy for each group.*
- *Counterfactual Fairness > An outcome is deemed fair if it would have been the same in a counterfactual scenario where the individual belonged to a different demographic group. This causal approach highlights factors that influence the AI's decision for a given individual.*

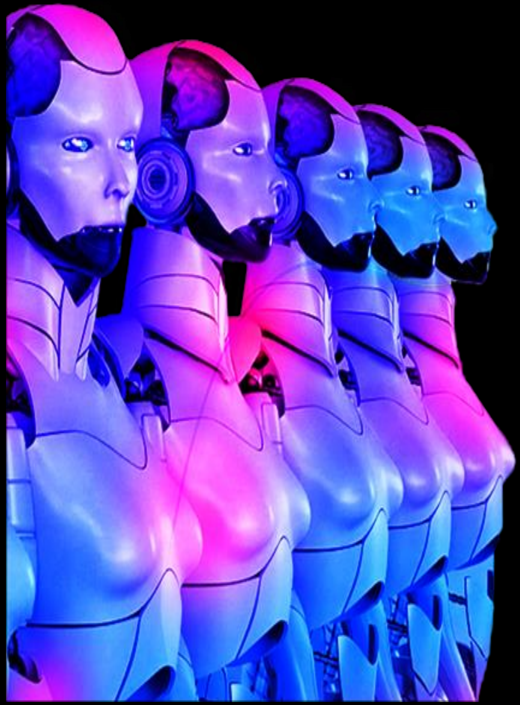
METRICS



- *Equal Opportunity True Positive Rate Parity > This metric defines fairness as equal true positive rates across groups - all qualified individuals should have equal probability of receiving a positive outcome regardless of group membership.*
- *Predictive Parity Positive Predictive Value Parity > According to this criterion, fairness means equal positive predictive value across groups - the probability of individuals predicted to be positive actually being positive should be the same for each group. It focuses on parity of precision.*

METRICS


Positives
Negatives
Unprivileged
(Group A)
Favorable Outcome Rate
=
50%
(5/10 total)
Privileged
(Group B)
Favorable Outcome Rate
=
60%
(6/10 total)
Key
True
False
Statistical Parity Difference
=
10%
(1/10 total)



FAIRNESS IN AI IS AN ONGOING PROCESS, NOT A ONE-TIME ACHIEVEMENT

REGULARLY REVIEW AND UPDATE AI SYSTEMS TO ALIGN WITH EVOLVING PRACTICES; REQUIREMENTS AND SOCIETAL EXPECTATIONS

AI IMPACT ASSESSMENT

ISO 42005 DRAFT

Dual Assessment

	AI Risk Assessment	AI Impact Assessment
Scope	Prevent deviations from the objectives in the AI software lifecycle	Prevent adverse impacts in human rights, security, and the environment
Focus on	internal losses for organizations involved in the development or use of AI systems	External losses for individuals, groups and societies caused by AI systems
Taxonomy	Cost overruns, fines and compensations, downtime, data corruption, profitability losses, IP losses	Discrimination, job displacement, fraud, extortion, humiliation, manipulation, disinformation, cyber-attacks, energy over consumption
ISO	ISO 23894 on AI risks, ISO 27005 IT Risks, ISO 31000 general risks	ISO 42005 D on AI impact, ISO 29134 on privacy

DUAL USE

	Intended use	Unintended use
Scope	Reasonable foreseeable purposes for which an AI system is designed, trained and tested by accepted users and information systems	Use or application of an AI system in a way not intended by the AI developer or provider which may cause beneficial, negative or neutral impacts
Restricted by	Laws, organizational policies, contractual agreements, training	Security controls, usage monitoring and audits, due diligence
Features	Predictive analytics, Automatic decision-making, content generation, optimization, object detection	Crime exploitation, misinformation, people tracking, device hijack
Access	Business and consumer users, vulnerable users	Cybercriminals, authoritarian regimes, curious users

COMPLETING IMPACT ASSESSMENTS

EVIDENCE

Document and date the process and artifacts used during assessments

INTEGRATION

Explain how the assessment is integrated with other organizational processes in risk, audit procurement, and security functions

TIMING

Consider legal requirements, risk levels, and stakeholder expectations to define the moment and level to initiate the assessment

COMPLETING IMPACT ASSESSMENTS

TRIGGERS

Assess when there's a change in AI system use, risk severity or impact type covering the modification of data, parameters, new users and uses, new features and changes in the compliance obligations

FREQUENCY

Define the stages of the AI system life cycle for completing and updating assessments

SCOPE

Consider the level of aggregation to include the entire AI system or specific components

COMPLETING IMPACT ASSESSMENTS

ROLES

Allocate responsibilities for assessment tasks in research, development, product and project management, data ethics, risk management, compliance and legal

LIMITS

Define thresholds for sensitive and restricted uses and intended users considering accountability, ethical frameworks, and suitable labels

EVALUATION

For intended uses and possible misuse, consider the beneficial and harmful impacts to individuals, groups, and societies

COMPLETING IMPACT ASSESSMENTS

SCENARIOS

Analyze results for responsible use and development of AI systems from the technical and management perspectives in analysis.

REPORT

Address internal and external reporting needs considering objectives, legal obligations

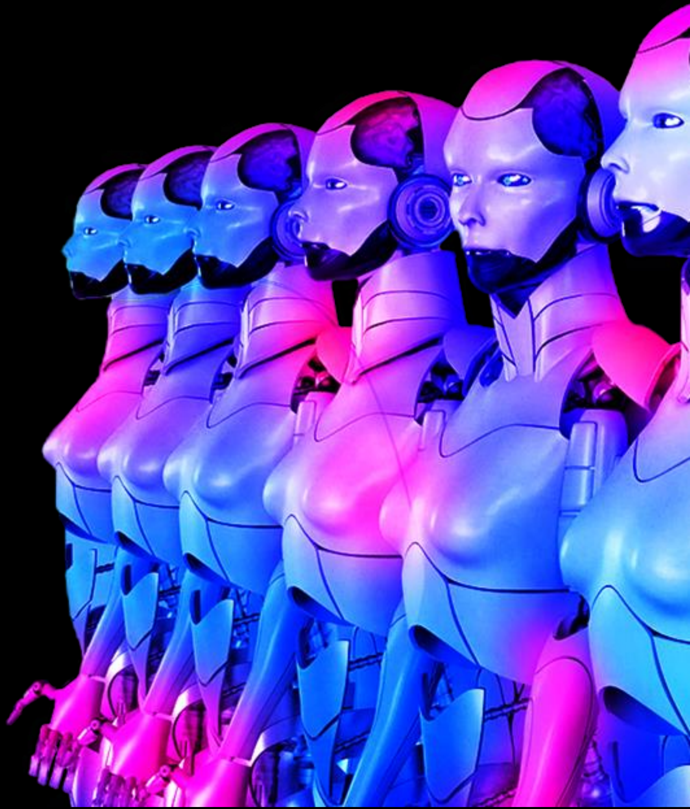
APROVE

Get internal and external approvals on the assessment and exceeding thresholds

INTERESTED GROUPS

- **End-users** Individuals or businesses that directly interact with the AI system as the the primary beneficiaries
- **Data subjects** Individuals whose personal data is processed by AI systems
- **Vulnerable groups** Segments of society that may be disproportionately affected by the implementation and impact of AI systems
- **Societies** People within a community, nation, or global context that may be influenced by the societal, economic, and cultural implications of AI systems
- **System developers and designers** Experts or teams responsible for conceptualizing, designing, and implementing the AI system or software
- **System owners and operators** Organizations that own, manage, and maintain the AI system or software responsible the operational integrity, security, and compliance
- **Regulatory authorities** Government agencies and legislative bodies
- **Academia** Entities involved in the theoretical study of AI technologies
- **Advocacy groups** Non-governmental organizations for ethical development
- **Media**: Entities shaping public perception surrounding AI technologies

AI DOCUMENTATION MANAGEMENENT



MODEL CARD

- *Document describing the intended context and use of an AI model*
- *It includes performance evaluation procedures and metrics to allow AI developers to compare results with other models for similar purposes*

Define the model's purpose clearly and explain its intended goal

Document the training data, including sources, size, and acquisition methods

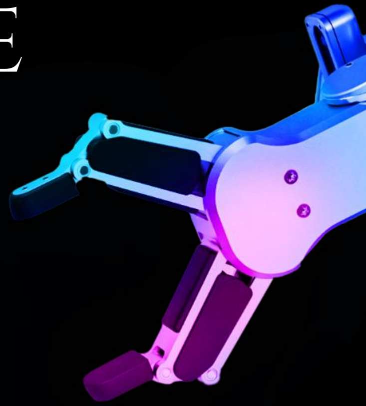
Describe ethical considerations and potential biases in the training data

Specify intended use cases and where the model might not perform well

MODEL CARD

- Include performance metrics like accuracy and generalizability
- Detail evaluation methodologies used to assess model performance
- Explain model decision-making processes and identify potential biases
- Outline techniques for mitigating biases and ensuring fair outcomes
- List known limitations, such as susceptibility to specific prompts or errors
- Optionally, estimate the environmental impact of training the model

TRANSPARENCY PRINCIPLE

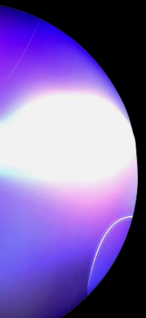


CLARITY

Produce detailed information about AI systems accessible and understandable to stakeholders, ensuring responsible and ethical use while complying with legal requirements

ASSESS

Model cards enhance transparency by providing clear explanations of machine learning models, including their construction, purpose, and performance, helping stakeholders understand the AI system's inner workings and potential biases



OBJECTIVES

INSIGHT

Model cards help stakeholders understand the model's design, data, and performance, highlighting its strengths and weaknesses

RISK FACTORS

Model cards reveal issues like bias, copyright violations, and factual errors, helping assess and manage risks

REPRODUCIBILITY

Model cards document the development process, enabling independent assessment and replication

DATA SHEETS

- *Document describing the in-depth technical description of AU models, detailing construction parameters and operational characteristics*
- *Explain the model development process, outlining the steps to build and train the model and any relevant algorithms used*

Include a transparent view of the model's internal logic to foster trust and enable more informed risk assessments

Break down the training data characteristics, including sources, size, distribution, and data quality checks performed

Document the training process, including optimization algorithms, success objectives, and convergence criteria

Provide performance metrics to evaluate the model's effectiveness using training and validation datasets

DATA SHEETS

Clearly define the model's output format, including data types and interpretation of results

Enable AI risk managers to identify potential failure points and develop mitigation strategies

Demonstrate regulatory compliance with relevant AI model development and deployment guidelines

Disclose assumptions and limitations made during development and inherent to the model architecture or data

Ensure data sheets facilitate rigorous validation of the model's effectiveness and generalizability

Promote reproducibility by ensuring data sheets enable independent parties to recreate and validate the model

RISK CARDS

- *Quick references describing common scenarios to standardize and facilitate AI risk assessments used by architects and data scientists*
- *Facilitate stress testing by using risk cards to brainstorm potential risks and analyze model behavior under different conditions without any concrete analysis*

Ensure risk cards remain dynamic, evolving with new risks and changes in context

Use risk cards to identify and mitigate biases in AI models

Incorporate fairness constraints in training processes to reduce biased outputs

Conceptualize risk cards as open-source assets, allowing anyone to add or edit risks

RISK CARD FIELDS

EXAMPLE

- Title > Brief, concrete, descriptive title of the risk
- Description > Description of the risk, affected AI tools and models, and impacted groups, Taxonomy > Main type and subcategory based on a chosen taxonomy
- Potential harms > List of possible negative impacts and affected stakeholders
- Stakeholders > Specific individuals, groups, or organizations affected by or responsible for managing the risk
- Evidence > References to laws, publications, or real-world examples demonstrating the risk

  Factors > Conditions or actions that may materialize the risk, including required access or resources and other risks that may be connected to or influenced
- Common controls > Potential mitigation strategies that can be implemented to reduce or eliminate the risk
- Monitoring > Metrics used to track the risk over time
- Example > Sample AI output showcasing the risk, with relevant model details

EXAMPLE

AI RISK TAXONOMY

- Responsible AI > *Privacy issues, generation of harmful content, and promotion of bias*
- Reputation > *Negative press due to inappropriate model usage*
- Cyberecurity > *Data breaches, manipulation attempts, and other security vulnerabilities*
- Societal > *Job displacement and misuse of AI for propaganda*
- Operational > *Challenges with limited training data, compute intensity, and system integration*
- Regulatory > *Non-compliance with laws and regulations, and intellectual property challenges*
- Financial > *Unexpected cost increases, such as using agentic workflows*
- Supply Chain > *External sources affecting partners and the organization*
- Environmental > *High energy consumption and generation of harmful gases*

AI HARM TAXONOMY

- *Discrimination > Social stereotypes, unfair discrimination, exclusionary norms and toxic language - These harms involve biased treatment, exclusion of certain groups, and promotion of harmful language or behavior*
- *Representational > Stereotyping, demeaning social groups, erasing or alienating social groups, denying self-identification, and reinforcing essentialist social categories - These harms affect how groups are portrayed or represented, potentially reinforcing negative stereotypes or erasing identities*
- *Allocative > Opportunity loss and economic loss. These harms involve unfair distribution of resources or opportunities, leading to economic disadvantages for certain groups.*
- *Quality-of-Service > Alienation and increased labor - These harms affect user experience, potentially causing feelings of isolation or requiring more effort from user*
- *Inter- intrapersonal >Service or benefit loss, loss of agency, social control, technology-facilitated violence, diminished health and well-being, privacy violations - These harms affect individuals' personal lives, relationships, and overall well-being*

AI HARM TAXONOMY

- Societal > Information harms, cultural harms, political and civic harms, macro socio-economic harms, and environmental harms - These harms have broader impacts on society, culture, politics, and the environment
- Information > Lower performance for some languages or groups, privacy compromises through leaking or inferring information - These harms relate to unequal system performance and risks to personal privacy
- Misinformation > Disseminating false or misleading information, causing material harm through poor information, and leading users to unethical or illegal actions - These harms involve the spread of inaccurate or harmful information and its consequences
- Malicious uses > Facilitating disinformation, fraud, scams, cyber attacks, weapons development, illegitimate surveillance, and censorship - These harms involve the deliberate misuse of AI systems for harmful purpose
- Human-computer interaction > Overreliance due to anthropomorphization, exploiting user trust, and manipulation - These harms arise from how users interact with and perceive AI systems, potentially leading to misuse or exploitation

DOCUMENTATION

AI System Information

- *Describe the AI system's capabilities and how it works*
- *Include technical requirements, demonstrations, and proof of concepts*

AI System Features

- *Identify and describe the AI system's features*
- *Consider predictions, data types, algorithms, user interaction, and configurations*

AI System Purpose

- *Explain why the AI system was created and its objectives*
- *Document any relationships with other systems or products*

Intended Uses

- *Identify specific scenarios for which the AI system is intended*
- *Consider potential impacts on users and societies*

Potential Misuse

- *Identify possible misuses of the AI system*
- *Consider both intentional abuses and unintentional misuses*

DOCUMENTATION

Use identification

- *Recognize foreseeable misuse and intentional abuse*
- *Understand potential beneficial uses*
- *Use ISO/IEC 42001 for guidance on acceptable and prohibited uses*

Data information

- *Include a variety of data, from none to comprehensive*
- *Consider data type relevance during assessment*
- *Address data availability, format, volume, and quality*

Data quality

- *Apply documentation to trained models for output data*
- *Ensure datasets meet requirements to avoid incorrect outputs and bias*
- *Address issues like bias or unfairness in training datasets*

Quality model life cycle

- *Ensure AI system development meets data quality requirements*
- *Address gaps where requirements are not yet met*

DOCUMENTATION

Information on used algorithms

- *Evaluate algorithm alignment with business goals and tasks*
- *Document modifications and reasons*
- *Record real-world performance and undesirable outcomes*

Algorithm development

- *Ensure requirements are met before deployment*
- *Document plans for unmet requirements*
- *Record approval process for algorithm use*

Models information

- *Detail data (training, testing, validation) and development algorithms*
- *Avoid data sample reuse between training and validation/test*
- *Document model selection criteria*

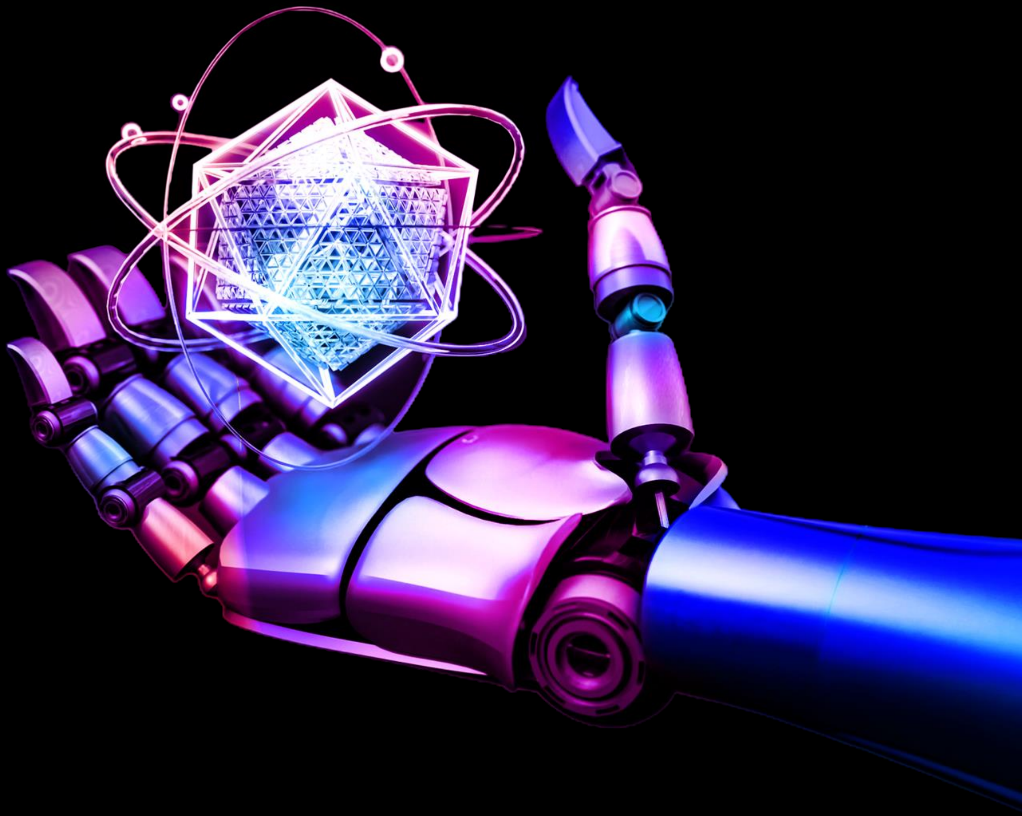
Deployment environment complexity and constraints

- *Include technical environment details and constraints*
- *Consider online service specifics, security protocols, and infrastructure*

RECOMMENDATIONS

- *Define a process for creating and maintaining documentation*
- *Appoint specific owners for each document, ensuring they have the necessary skills and technical knowledge*
- *Define guidelines for when model cards, assessments and other documents are necessary, such as for models used by over multiple AI specialists and or in production and testing*
- *Involve cross-functional teams in the creation process to ensure comprehensive coverage*
- *Use a standardized template to ensure consistency and ease of use*
- *Leverage automation tools to generate documentation, reducing manual effort and increasing accuracy*
- *Utilize version control systems to track changes and maintain a clear record of updates*
- *Establish a centralized repository for documentation , ensuring easy access and management*

RESPONSIBLE AI PRINCIPLES



PRINCIPLES AT RISK

Accountability

- *Be responsible for actions, decisions, and performance related to AI systems*
- *Refer to existing accountability frameworks*
- *Analyze potential benefits and harms related to transparency*

Transparency

- *Communicate information about AI systems clearly*
- *Ensure relevant parties understand system capabilities*
- *Address transparency gaps to avoid unintended consequences*

Fairness

- *Impartial behavior without discrimination*
- *Treat all groups fairly during data collection and system development*
- *Avoid biases or unfairness in AI systems*

Privacy

- *Protect personal identifiable information*
- *Address risks like unauthorized access, discrimination, and accuracy issues*

PRINCIPLES AT RISK

Reliability

- *AI should work correctly and consistently*
- *Analyze benefits and harms related to reliability*
- *Consider updates and their impact on performance*

Safety

- *AI should not endanger people or property*
- *Assess safety risks during use*
- *Address unsafe performance or changes*

Explainability

- *Humans should understand AI decision-making*
- *Consider complexity challenges in deep neural networks*
- *Ensure sufficient information for understanding*

Environmental Impact

- *Evaluate energy consumption of AI systems*
- *Most systems run on electricity, raising sustainability concerns*

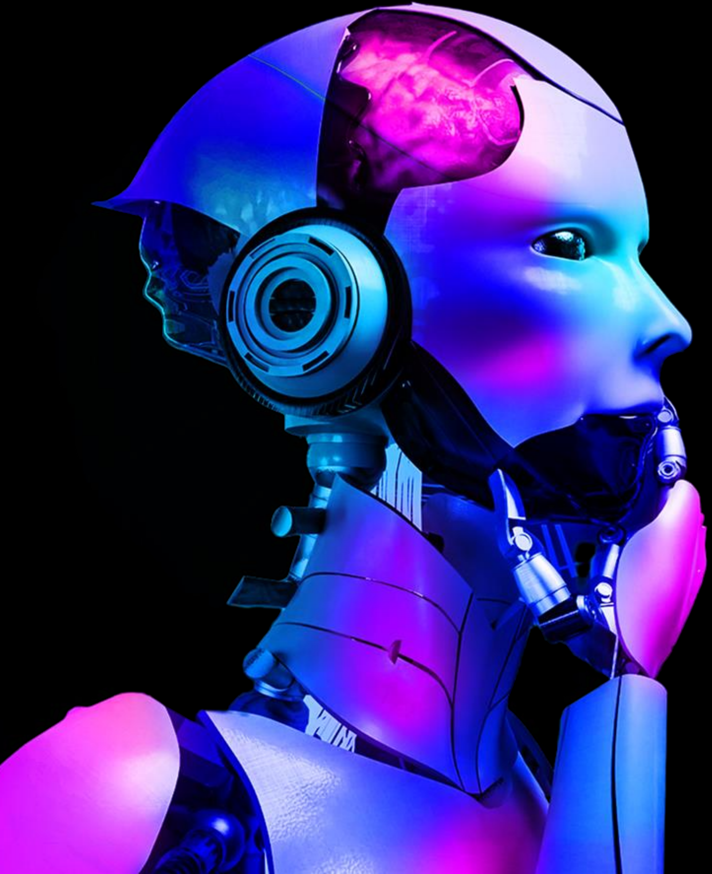
IMPACT CRITERIA

CRITERIA

- *Compliance requirements such as responsible AI commitments and policy, privacy and eprofiling, controlled uses, and contractual terms*
- *Expectations of interested parties*
- *Limitations of the models and technology*
- *Cultural norms and societal values*

PROCESS

- *Weigh benefits against potential risks*
- *Establish review processes for sensitive or restricted use cases*
- *Involving diverse perspectives to assess potential impacts*
- *Cover from design and development to deployment and monitoring*



PROF. HERNAN HUWYLER, MBA CPA

ie EXECUTIVE EDUCATION
COMPLIANCE, RISKS, CONTROLS
CYBER SECURITY AND GOVERNANCE

/hernanwyler

/hewyler