Od „responsible AI” do konkretnej macierzy ryzyk

MLGdańsk

Jakub Szarmach

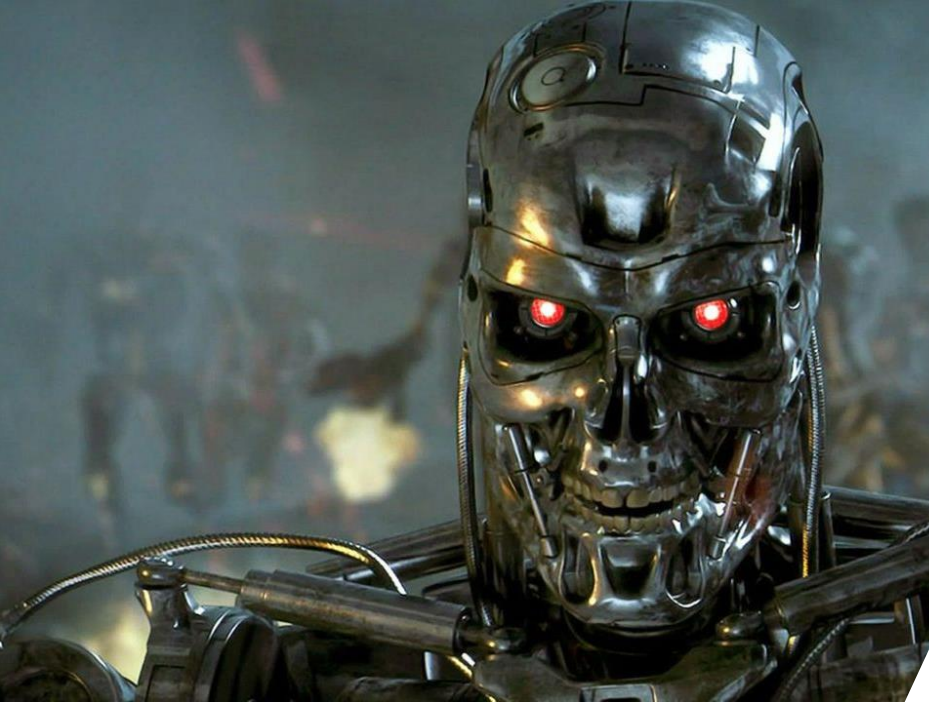
Advanced AI Risk & Compliance Analyst

Kurator AI Governance Library

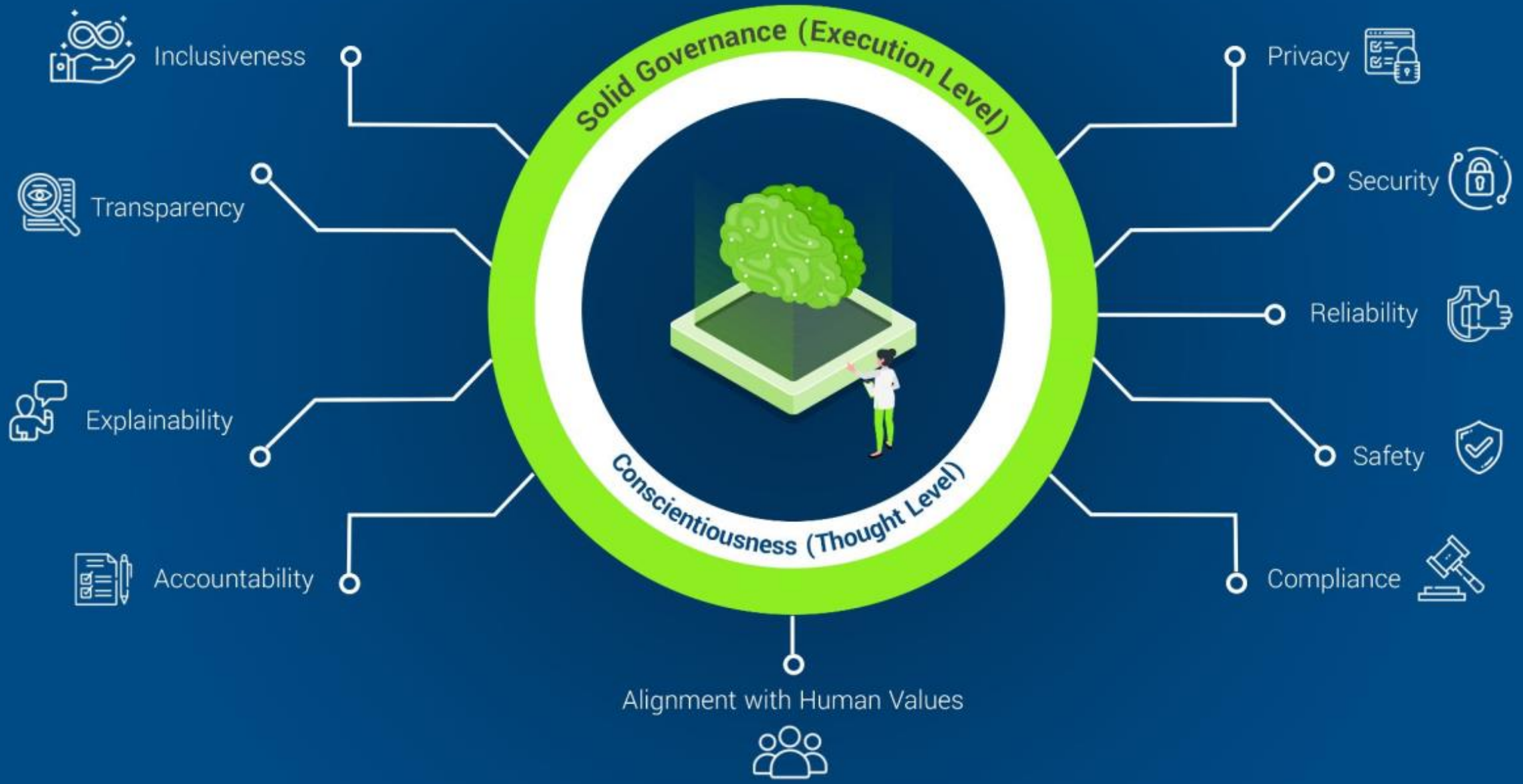
Certified AI Governance Professional AIGP iappai

Certified Information Privacy Manager CIPM iapp

Po co komu responsible AI?



Pillars of Responsible AI



Po co komu responsible AI?

DEKLARACJE ETYCZNE I „AI PRINCIPLES” SĄ POTRZEBNE, ALE NIE WYSTARCZAJĄ.

ZESPOŁY NAUKOWE PYTAJĄ: „OK, *ALE CO MAMY ZROBIĆ KONKRETNIE?*”

BRAKUJE NARZĘDZIA, KTÓRE ŁĄCZY TECHNOLOGIĘ, ETYKĘ I PRAKTYKĘ.

Cel projektu

Jeżeli nie wiesz dokąd zmierzasz ,
możesz dojść gdzieś gdzie nie
chciałbyś się znaleźć

Cel projektu

1. Konkretny efekt, jaki ma przynieść AI
2. Kontekst badawczy
3. Zakres działania modelu

„Model AI ma wspierać analizę sygnałów z czujników laboratoryjnych, wykrywając anomalie, które wymagają oceny człowieka. Celem jest skrócenie czasu analizy o 40% po zakończeniu wdrożenia projektu, bez wpływu na jakość interpretacji wyników.”

4. Ustal **apetyt na ryzyko**


2023
2024

Zarządzanie ryzykiem AI

Czym jest ryzyko?

Ryzyko to kombinacja **prawdopodobieństwa** wystąpienia zdarzenia i jego **wpływu** na cele organizacji

Ryzyko to nie “problem” — to informacja dla decydowania


Risk Management Lifecycle
Risk Identification
Risk Analysis
Risk Mitigation Planning
Risk Management Implementation
Review and Tracking

Identyfikacja Ryzyka

Zrób pełną listę ryzyk – bez cenzury

- Spisz wszystko, co może pójść nie tak: technicznie, prawnie, etycznie, operacyjnie.
- Nie oceniaj na tym etapie – celem jest **szerokie pole widzenia**, nie perfekcja.
- Włącz zespół: inni zobaczą ryzyka, których Ty nie widzisz.
- Korzystaj z gotowych frameworków.

2025 OWASP Top 10 List for LLM and Gen AI

LLM01:25

Prompt Injection

This manipulates a large language model (LLM) through crafty inputs, causing unintended actions by the LLM. Direct injections overwrite system prompts, while indirect ones manipulate inputs from external sources.

LLM02:25

Sensitive Information Disclosure

Sensitive info in LLMs includes PII, financial, health, business, security, and legal data. Proprietary models face risks with unique training methods and source code, critical in closed or foundation models.

LLM03:25

Supply Chain

LLM supply chains face risks in training data, models, and platforms, causing bias, breaches, or failures. Unlike traditional software, ML risks include third-party pre-trained models and data vulnerabilities.

LLM04:25

Data and Model Poisoning

Data poisoning manipulates pre-training, fine-tuning, or embedding data, causing vulnerabilities, biases, or backdoors. Risks include degraded performance, harmful outputs, toxic content, and compromised downstream systems.

LLM05:25

Improper Output Handling

Improper Output Handling involves inadequate validation of LLM outputs before downstream use. Exploits include XSS, CSRF, SSRF, privilege escalation, or remote code execution, which differs from Overreliance.

LLM06:25

Excessive Agency

LLM systems gain agency via extensions, tools, or plugins to act on prompts. Agents dynamically choose extensions and make repeated LLM calls, using prior outputs to guide subsequent actions for dynamic task execution.

LLM07:25

System Prompt Leakage

System prompt leakage occurs when sensitive info in LLM prompts is unintentionally exposed, enabling attackers to exploit secrets. These prompts guide model behavior but can unintentionally reveal critical data.

LLM08:25

Vector and Embedding Weaknesses

Vectors and embeddings vulnerabilities in RAG with LLMs allow exploits via weak generation, storage, or retrieval. These can inject harmful content, manipulate outputs, or expose sensitive data, posing significant security risks.

LLM09:25

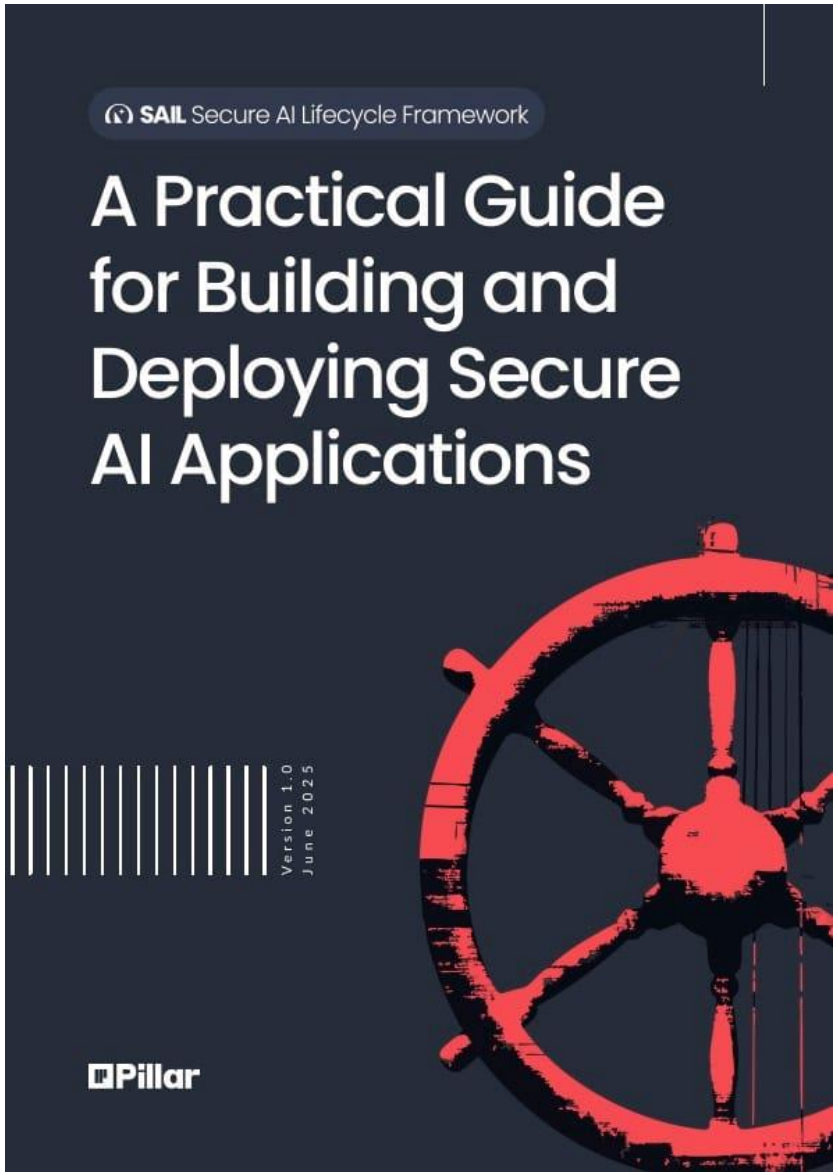
Misinformation

LLM misinformation occurs when false but credible outputs mislead users, risking security breaches, reputational harm, and legal liability, making it a critical vulnerability for reliant applications.

LLM10:25

Unbounded Consumption

Unbounded Consumption occurs when LLMs generate outputs from inputs, relying on inference to apply learned patterns and knowledge for relevant responses or predictions, making it a key function of LLMs.

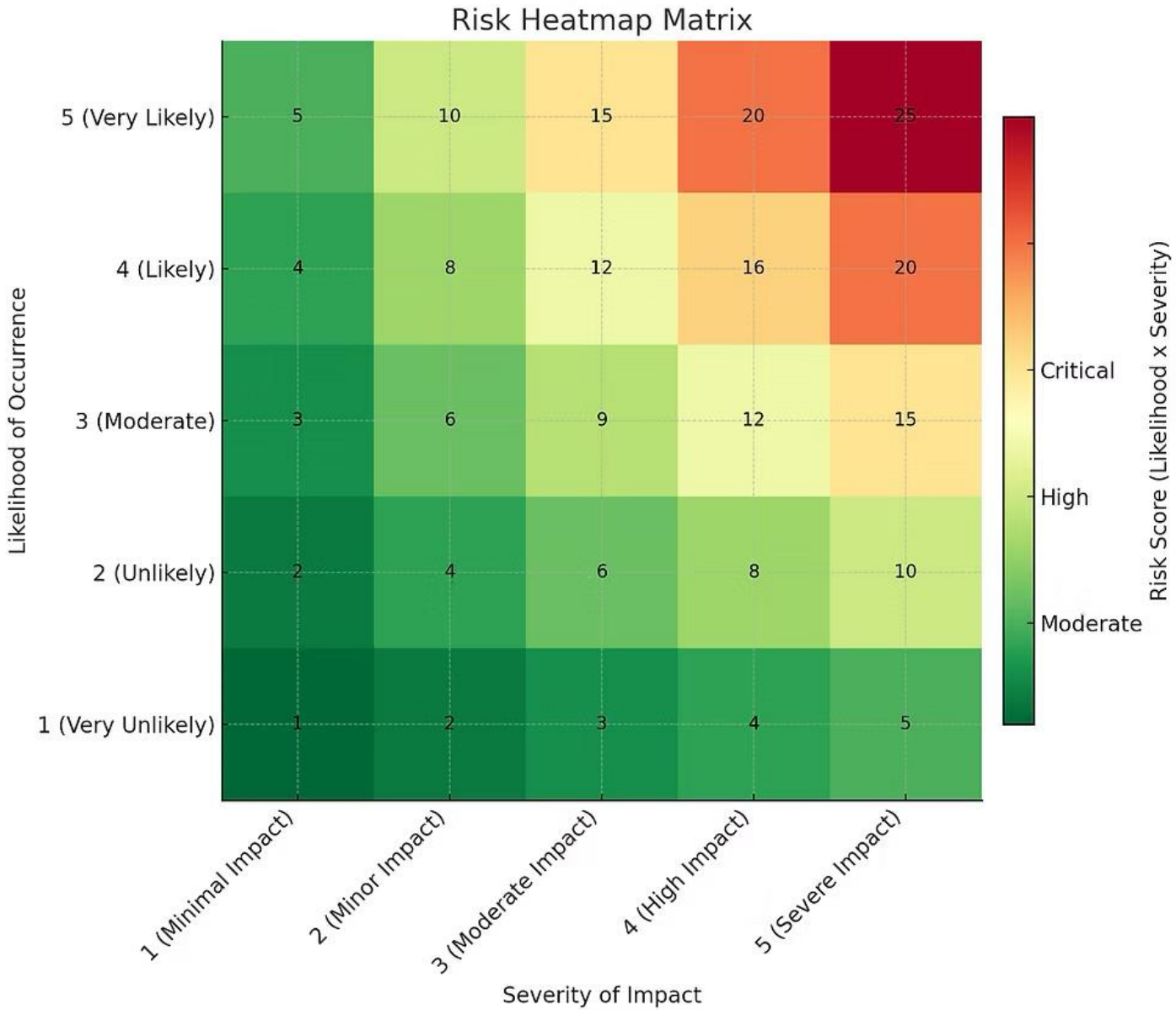

SAIL Secure AI Lifecycle Framework
A Practical Guide for Building and Deploying Secure AI Applications
Version 1.0
June 2025
Pillar


Map
Measure
AI Risk Management Framework
Govern
Manage

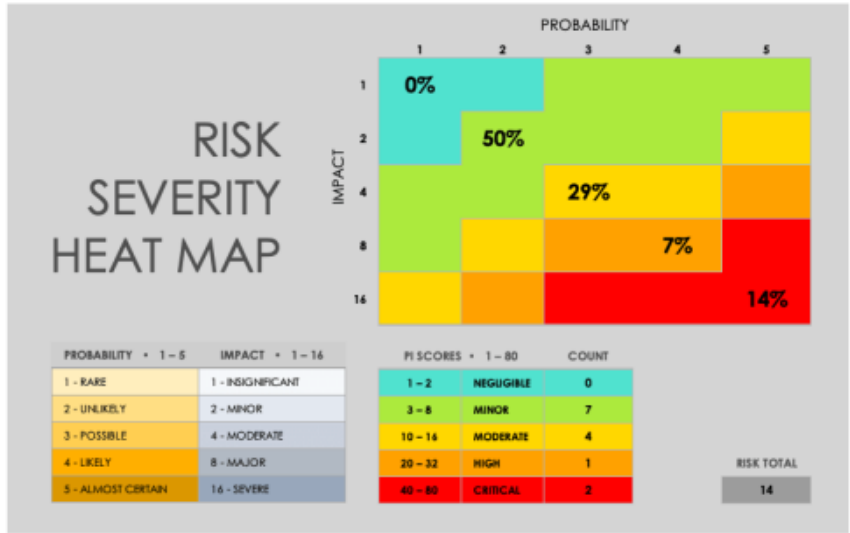
Identyfikacja Ryzyka

Nadaj parametry

- Jak bardzo prawdopodobne, że to się wydarzy? (niske / średnie / wysokie)
- Jaki będzie wpływ, jeśli się wydarzy? (lokalny / projektowy / poważny dla ludzi / reputacyjny)
- Stwórz wspólny język nazywania ryzyka



RISK ASSESSMENT HEAT MAP TEMPLATE FOR EXCEL



RISK ASSESSMENT DATA TABLE

REF ID	DATE RAISED	RISK DESCRIPTION	LIKELIHOOD 1 – 5	IMPACT 1 – 16	RISK SEVERITY SCORE Prob x Impact	NOTES
R1	00/00/00	R1 Risk Description	4	4	16	
R2			3	2	6	
R3			5	8	40	
R4			4	2	8	
R5			5	2	10	
R6			3	4	12	

Identyfikacja Ryzyka

Odróżnij szum od tego, co krytyczne

- Ryzyka o **niskim wpływie i niskim prawdopodobieństwie** → zapisujemy, ale nie poświęcamy im dużo czasu.
- Ryzyka o **wysokim wpływie lub wysokim prawdopodobieństwie** → trafiają na „listę priorytetową”.
- Jeśli ryzyko dotyczy **uczestników badań, danych wrażliwych albo reputacji uczelni** – traktujemy je z definicji jako istotne.

Wynik: krótka lista ryzyk, którymi naprawdę zarządzamy

RISK and INCIDENT RANKING MATRIX

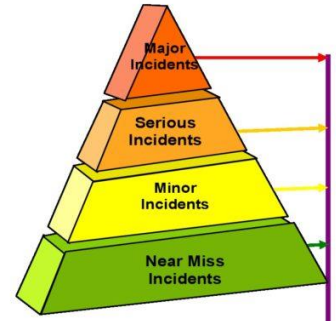
NOTES:

1. "Serious" risks to people and environment have higher priority than serious risks to asset or production.
2. When choosing Likelihood, consider:
   * Near Miss incidents
   * Incidents which have happened on-site or at similar facilities
   * Design of the Facility
     → Compared to intended operation
     → Over-design
     → Design safety margins
     → Materials of construction
   * Publications and other industry documentation
   * PHA's performed on similar operating facilities

Risk Ranking / Priority Action Setting

Critical	**Risk is unacceptable to the company. WORK SHALL NOT PROCEED until risk is reduced.** Immediate Action Plan is required before proceeding.
Serious	**Risk is undesirable. May proceed with approval from responsible Management.** Consider additional safeguards to reduce risk. Short-Term Action Plan is required.
Moderate	Reasonably practicable risk reduction measures and safeguards must be used. Long-Term Action Plan should be considered.
Acceptable	Normally accepted controls should be in place. No additional action is required.

Relationship to Incident Reporting Pyramid



Likelihood Of Occurrence

	IMPROBABLE	REMOTE	OCCASIONAL	PROBABLE	FREQUENT
Frequency	Once in 1000 yrs	Once in 100 yrs	Once in 10 years	Once in 5 years	Once per year
	1.0 E-03 /yr	1.0 E-02 /yr	1.0 E-01 /yr	2.0 E-01 /yr	1.0 E-00 /yr
Qualitative Description	Very unlikely to occur in lifetime of this facility; Rare occurrence in the industry.	Unlikely, but could occur once in lifetime of this facility. Rare occurrence in the industry.	Infrequent occurrence in the industry.	May occur several times at facility. Common occurrence in the industry.	Very likely to occur. Regular, repeated occurrence in the industry.

Consequence / Severity Ranking	A	B	C	D	E
5	Moderate	Serious	Critical	Critical	Critical
4	Acceptable	Moderate	Serious	Critical	Critical
3	Acceptable	Acceptable	Moderate	Serious	Serious
2	Acceptable	Acceptable	Acceptable	Moderate	Serious
1	Acceptable	Acceptable	Acceptable	Acceptable	Moderate
0	Acceptable	Acceptable	Acceptable	Acceptable	Acceptable

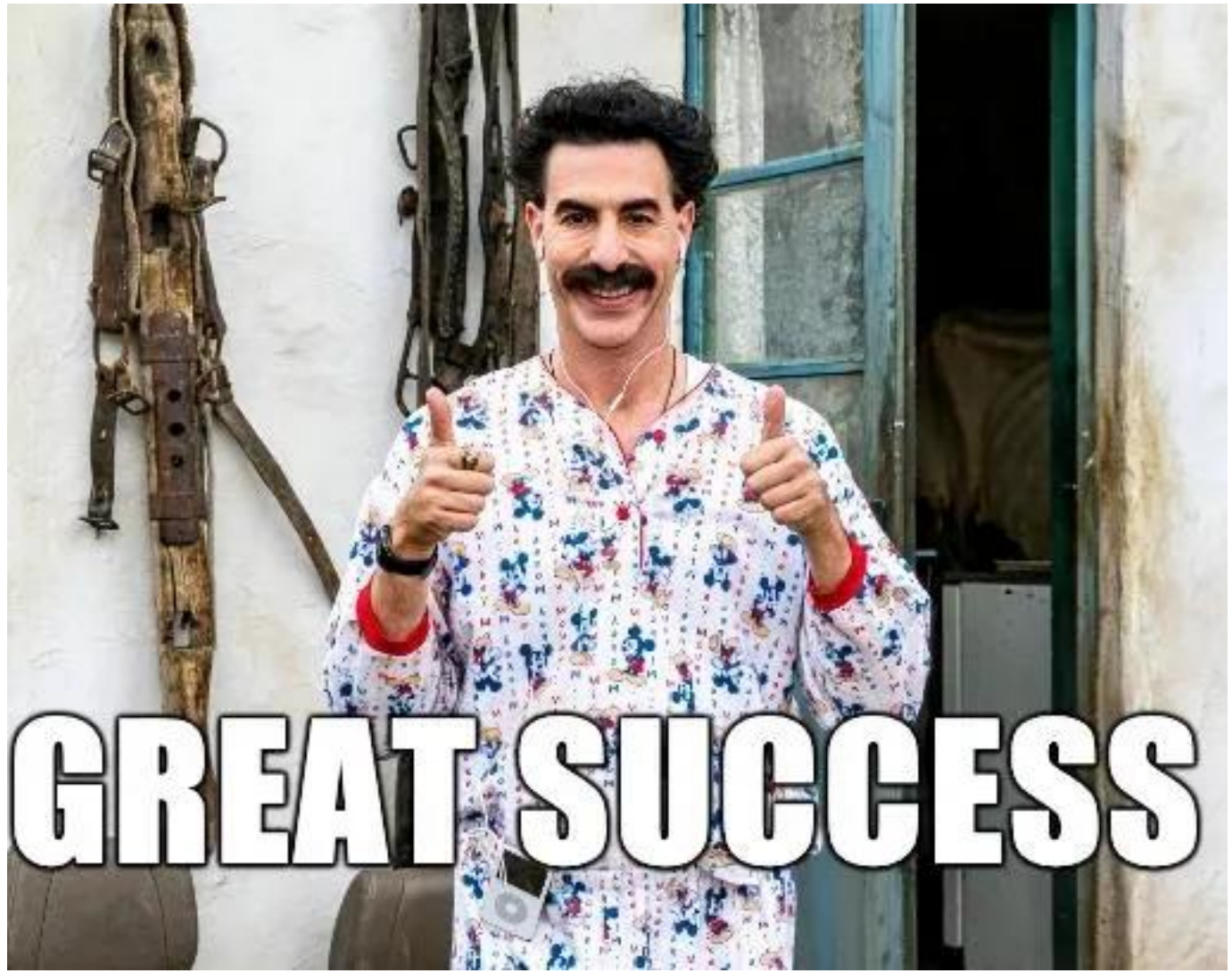
Consequence / Severity Description (without safeguards)

	People	Environmental	Asset Damage or Losses	Unplanned Downtime for Client	Reputation	
5	**Fatality or Permanent Disability**	**Reportable** spill or release resulting in severe ecological impact. Direct impact on public. Prosecution.	Greater than $ 1 Million	Greater than 90 days, or well must be abandoned	Negative international publicity, significant impact on market share or investor valuation	Major Incident
4	**Hospitalization**	**Reportable** spill or release requiring external remedial measures. Regulatory restriction or enforcement action.	$ 100,000 to $ 1 Million	15 days to 90 days	Negative national or regional publicity, transitory impact on market share or investor valuation	
3	**Lost Time Injury/Illness (LTI)**	**Reportable** spill or release not contained within client facility and requiring activation of facility's remedial actions or measures. **Non-reportable** spill or release in company facility requiring outside mitigation services.	$ 10,000 to $ 100,000	3 days to 15 days	Local media coverage, Community complaint	Serious Incident
2	**Medical Aid and/or Restricted/Modified Work Case**	**Reportable** spill or release contained within client facility, or **small release** on company facility not requiring activation of any remedial measures.	$ 2,500 to $ 10,000	1 shift to 3 days	Little or no local media coverage	
1	**First Aid**	**Non-reportable** spill or release contained within company or client facility.	Less than $ 2,500	Less than one shift	Negative public complaint	Minor Incident
0	**No injury / illness; or NEAR MISS**	No environmental impact	No cost impact	No downtime	No Impact	Near Miss Incident

USE FOR SEVERITY RANKING OF ACTUAL INCIDENTS

USE FOR RISK RANKING OF POTENTIAL INCIDENTS

Document No.	Revision	Date	Reason for Issue	Prepared	Revised	Endorsed	Approved
			Issued for Review		**Operations Manager**	**HSE Manager**	**VP Operations**


GREAT SUCCESS

Zarządzanie ryzykiem

Dlaczego zarządzamy ryzykiem?

- Aby **zmniejszyć** ryzyko inherentne do poziomu, który jest **akceptowalny** dla zespołu, uczelni i uczestników badań.
- Aby **świadomie decydować** co robim z ryzykiem.
- Aby uniknąć sytuacji, w której ryzyka „same się wydarzą”, bo nikt się nimi nie zajął.

Krótko:

- **Ryzyko inherentne** = startowy poziom ryzyka.
- **Zarządzanie ryzykiem** = proces, dzięki któremu nie działamy w ciemno.

Risk mitigation strategies

Four basic ways how to treat the risk

Accept

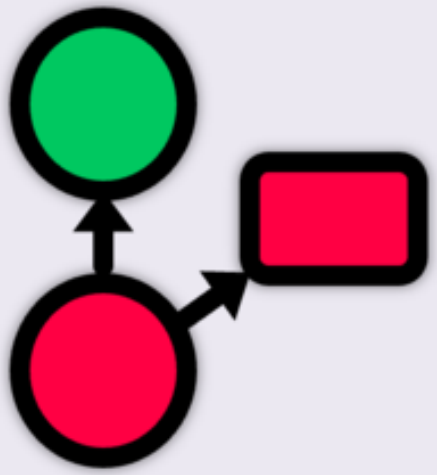
Hope it doesn't happen

Avoid

Cancel the source of the risk

Transfer

Move risk to someone else



Reduce

Decrease probability or impact

Zarządzanie ryzykiem

- **Wybierz jedną strategię dla każdego ryzyka**
- **Wyznacz właściciela ryzyka**
- Owner odpowiada za wdrożenie kontroli, monitoring i aktualizację statusu
- Bez właściciela macierz ryzyk jest **nieoperacyjna**


RISK
RISK
RISK
RISK
RISK
RISK
RISK
CONTROLS
RESIDUAL RISK

Monitorowanie ryzyka

Ryzyko w projektach AI **nie jest stałe** — zmienia się, gdy zmieniają się:

- dane, modele, narzędzia, kontekst badawczy, integracje, decyzje zespołu.
- Monitoring to sposób, by **nie działać na autopilocie**.

Co monitorujemy?

- **Skuteczność kontroli** – czy faktycznie redukują ryzyko?
- **Incydenty i anomalie** – błędne wyniki, halucynacje, nieprzewidziane działania
- **Poziom ryzyka rezydualnego** – czy nadal mieści się w poziomie akceptacji?
- **Trend ryzyka** – urus, zmalau?

Monitorowanie ryzyka

- Monitorujemy to, co ma znaczenie dla **NAS**
- Brak jednego uniwersalnego zestawu wskaźników.
- **Częstotliwość zależy od dynamiki projektu.**
- Kontrole zmieniają się wraz z projektem
- **Monitoring to adaptacja, nie „odhaczanie”.**
- Owner ryzyka ma obowiązek dostosować monitoring

Gratulacje!

Zamieniliśmy **Responsible AI** w *czynne decyzje projektowe*.

jszarmach.pl
AIGL.blog
jakub.szarmach@gmail.com
+48 668 234 204

Dziękuję za uwagę!