Open Ethics

01–30

Really Simple AI Risk Framework

v1.0.0

A shared register for identifying, scoring, and tracking what can go wrong when you build, ship, or run AI — thirty risks, four lifecycle stages, one ID space.

RSAIRF · Open Ethics Initiative github.com/OpenEthicsAI/RSAIRF

WHY IT EXISTS

Every organization deploys and uses AI.
Few build it. Almost none monitor it well.

The AI Risk Register is curated from public sources and the internal incident learnings of real client engagements — not written from theory. It exists because the biggest exposures usually aren't in the model itself. They're in the gaps around it: tools adopted without approval, and monitoring that was never built.

AIR01–30	4	5
traceable IDs	lifecycle stages	mapped to existing governance

THE LIFECYCLE

Every initiative moves through four stages

Development
Model Development

→

Deployment
Model / Component Deployment

→

Use
AI Use

Continuous Monitoring & Evaluation — runs across all three stages, and feeds findings back into any of them

Development

You design or materially change AI behavior — data selection, training, tuning, evaluation.

Deployment

You don't build the model, but you integrate, configure, and release AI components into production.

Use

You consume a deployed AI solution and make decisions or take actions based on its outputs.

Monitoring

You track risk, quality, and incidents over time — continuous, and running across every other stage.

THE PROCESS

Four steps, repeated at every transition

01 Check which lifecycle stages apply to your initiative — run a discovery session with your technical team if it's unclear.

02 Map risks using register.csv. Stage tags show where each risk is typically introduced or first exploitable.

03 For every applicable AIR## risk, record likelihood, impact, an owner, and a mitigation — and reference that ID in your project's risk log or audit.

04 Found a risk that isn't represented? Open a repository issue — your contribution may help other teams too.

> Re-check the register at each stage transition — a risk that was out of scope earlier may now apply.

↻ treat this as a living list, not a one-time checklist

How probable it is

For every applicable risk, rate **likelihood** — how probable it is that the risk materializes in your context.

Score	Likelihood	Meaning
1	**Rare**	**Not expected during the initiative's lifetime.**
2	**Unlikely**	**Could occur, but not expected.**
3	**Possible**	**May occur occasionally.**
4	**Likely**	**Expected to occur at some point.**
5	**Almost certain**	**Expected to occur often, or already observed.**

How severe, if it happens

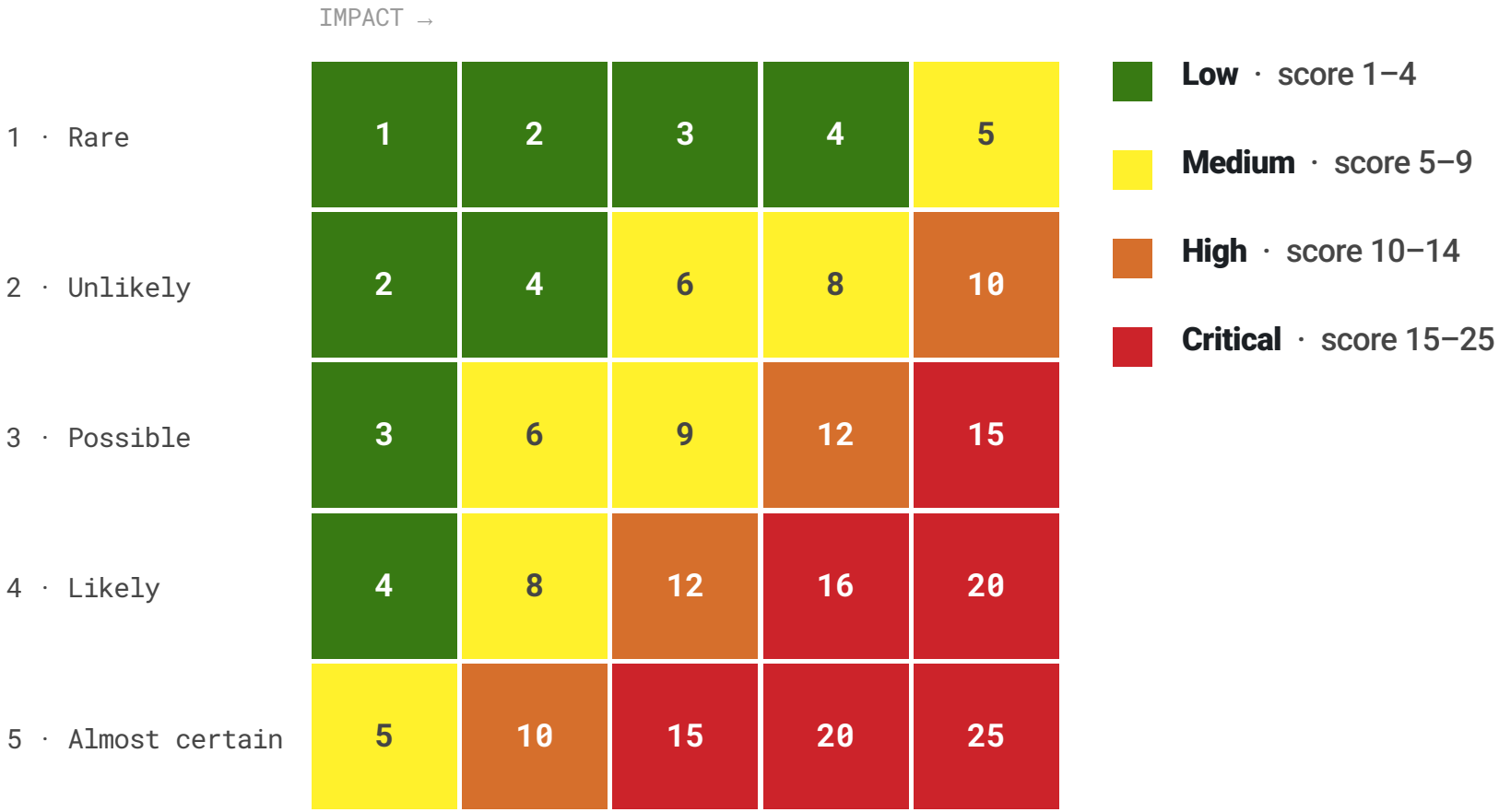
Then rate **impact** — the severity of the consequence if the risk does materialize.

Score	Impact	Meaning
1	**Negligible**	**Minimal effect; no material harm.**
2	**Minor**	**Limited harm, easily remediated.**
3	**Moderate**	**Noticeable harm, cost, or user impact.**
4	**Major**	**Significant financial, regulatory, or reputational harm.**
5	**Severe**	**Critical or potentially irreversible harm to people, rights, or the organization.**

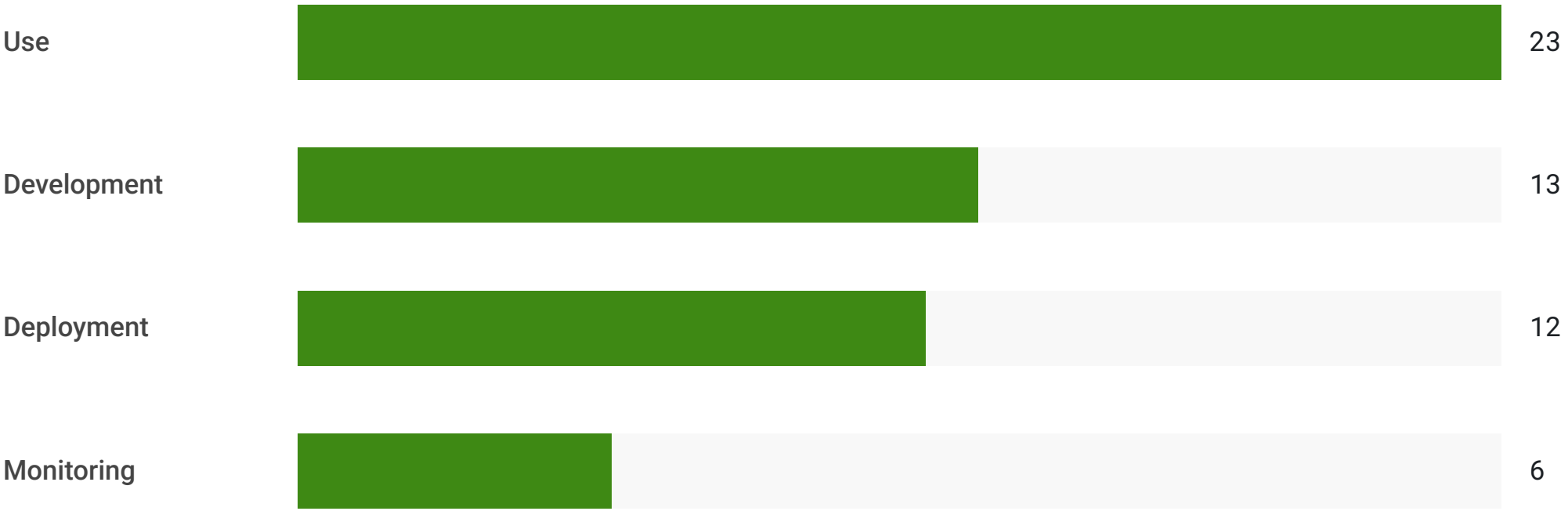
THE MATRIX

Likelihood × impact, at a glance

The same 1–25 score, banded into four priority levels — consistent across initiatives, without heavy tooling.



Thirty risks, mapped across four stages



SPOTLIGHT

A sample from the register

AIR01

Shadow AI Deployments

AI tools or models deployed and used by internal teams without documentation, approval, or oversight from security or governance.

DEPLOYMENT · USE

AIR03

Data Leakage

Sensitive data exposed through third-party tools, API calls, or model interactions, leading to loss of confidentiality or regulatory breach.

DEVELOPMENT · USE

AIR15

Adversarial Attacks

An attacker deliberately alters input data to mislead the model — from prompt injection to imperceptible perturbations.

USE

AIR17

Bias & Discrimination in Outputs

Skewed or discriminatory results from biased training data or model design, including unequal performance across demographic groups.

DEVELOPMENT · USE

AIR27

Unbounded Consumption

Uncontrolled or excessive inference requests degrade availability or generate runaway compute costs — a.k.a. Denial of Wallet/Service.

USE · MONITORING

AIR29

Multi-Agent & Agentic AI Risks

Risks specific to autonomous agents acting with tools and memory via an agent harness — unconstrained tool access, unsafe actions, and privilege escalation.

DEVELOPMENT · DEPLOYMENT · USE

FULL REGISTER

All thirty, AIR01–AIR30

ID	RISK	
AIR01	Shadow AI Deployments	Deployment, Use
AIR02	Supply Chain Vulnerabilities	Development, Deployment
AIR03	Data Leakage	Development, Use
AIR04	Cross-Border Data Transfer	Deployment, Use
AIR05	No Validation / Improper Output	Use
AIR06	Insecure Deployment Pipelines	Deployment
AIR07	Critical 3rd-party Dependency	Development, Deployment
AIR08	Excessive Agency / Over-Reliance on AI	Use
AIR09	Model Poisoning	Development
AIR10	Model Inversion / Membership Inference / Theft	Deployment, Use
AIR11	Poor Monitoring	Monitoring
AIR12	Feedback Loop Contamination	Use, Monitoring
AIR13	Alert Fatigue	Monitoring
AIR14	Insider Threats	Development, Deployment, Use
AIR15	Adversarial Attacks	Use

ID	RISK	
AIR16	Model Drift	Use, Monitoring
AIR17	Bias & Discrimination in Outputs	Development, Use
AIR18	Critical Unintended Consequences	Development, Deployment, Use
AIR19	Censorship/Guardrail interference	Use
AIR20	Deployment misfit	Deployment, Use
AIR21	Lack of Explainability / Provenance opacity	Development, Use
AIR22	No Accountability	Development, Deployment, Monitoring
AIR23	Unclear Model Ownership/IP	Development, Use
AIR24	Output Integrity Tampering	Use
AIR25	System Prompt / Instruction Leakage	Use
AIR26	RAG / Vector Store Vulnerabilities	Development, Use
AIR27	Unbounded Consumption	Use, Monitoring
AIR28	Malicious Use / Out-of-Scope Exploitation	Deployment, Use
AIR29	Multi-Agent & Agentic AI Risks	Development, Deployment, Use
AIR30	Environmental / Sustainability Impact	Development, Use

STANDARDS

Mapped to existing governance

Every AIR## risk is cross-referenced so findings plug straight into the programs you already run.

NIST AI RMF	OWASP LLM Top 10	EU AI Act	ISO/IEC 42001	MITRE ATLAS
Govern · Map · Measure · Manage	2025 edition	Articles 5–95	Annex A controls	AML.T#### techniques

AIR	Risk	OWASP LLM	EU AI Act	MITRE ATLAS
AIR03	**Data Leakage**	**LLM02 Sensitive Info**	**Art 10**	**AML.T0057**
AIR15	**Adversarial Attacks**	**LLM01 Prompt Injection**	**Art 15**	**AML.T0051**
AIR26	**RAG / Vector Store**	**LLM08 Vector Weakness**	**Art 10 & 15**	**AML.T0070**
AIR27	**Unbounded Consumption**	**LLM10 Unbounded**	**Art 15**	**AML.T0029**

Full mapping in standards-map.csv.

IN PRACTICE

From register to risk log

Record two ratings, an owner, and a control per applicable risk in your project's log — traceable back to the register by AIR ID.

Initiative	AIR	Risk	L	I	Score	Level	Owner	Status
Support chatbot	AIR05	No Validation / Improper Output	3	4	12	High	ML Lead	Mitigating
Support chatbot	AIR27	Unbounded Consumption	2	3	6	Medium	Platform	Accepted

Template fields: likelihood, impact, score, level, controls, owner, sign-off, KRI / monitoring signal, next review, status — see risk-log.template.csv.

WHAT'S INSIDE

Four files, ready to use

register.csv

The 30-risk AI Risk Register with IDs, lifecycle tags, and descriptions.

standards-map.csv

Cross-references to NIST, OWASP, EU AI Act, ISO 42001 & MITRE ATLAS.

risk-log.template.csv

Ready-to-fill log for scoring, ownership, sign-off, and review cadence.

README.md

Lifecycle model, usage workflow, and scoring guidance.

Open Ethics

GET INVOLVED

Treat it as a living framework.

If you identify a risk not represented here, propose it by opening an issue in the repository — your contribution may help other teams facing the same gap.

REPOSITORY	LICENSE
github.com/OpenEthicsAI/RSAIRF	CC BY 4.0