SaferAI

April 2025

Informing the Code of Practice: A Hierarchical Methodology of Defining Risk Tiers

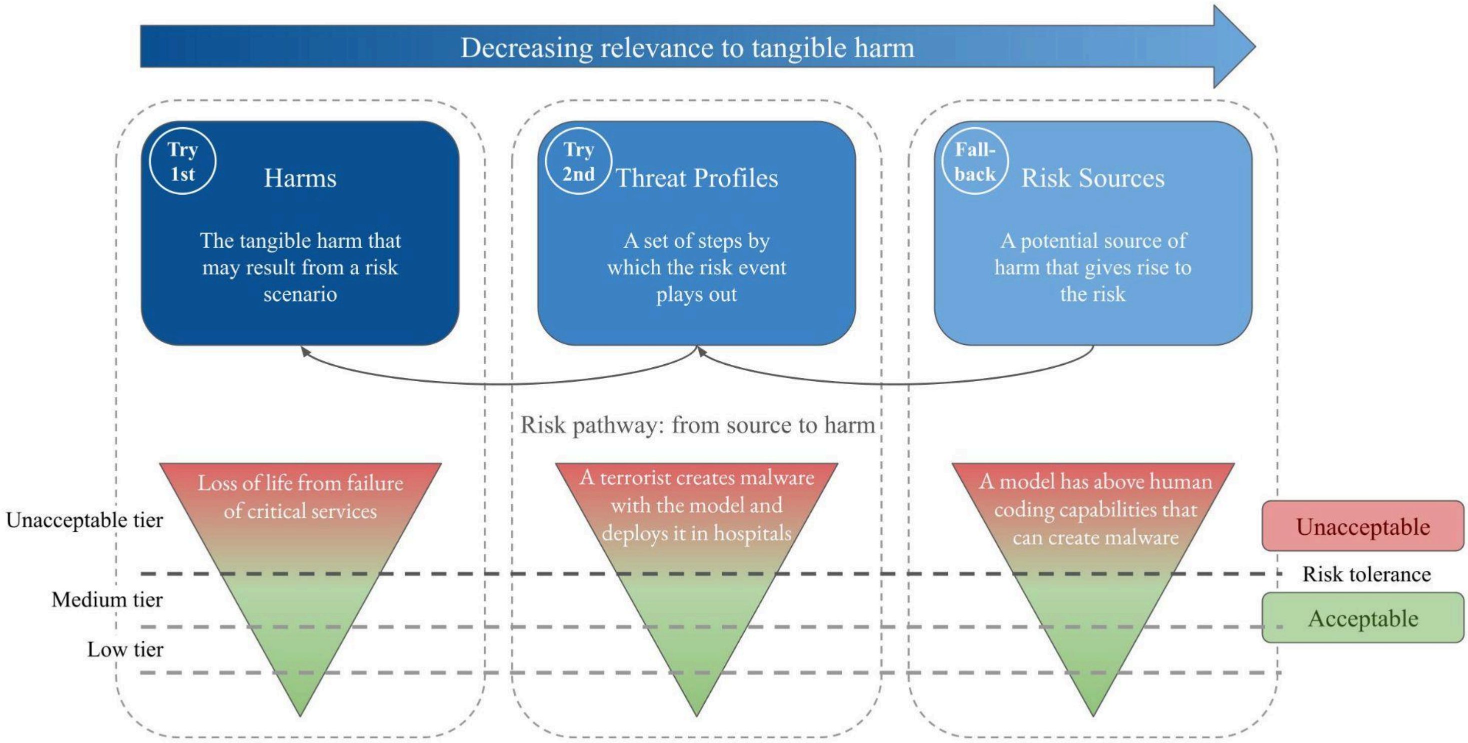
Daniel Kossack, James Gealy, Malcolm Murray,
Henry Papadatos, Siméon Campos, Otter Quarks

Introduction

Defining acceptable and unacceptable risk levels is a necessary step when managing risk. Our risk management framework (Campos et al., 2025) leverages two concepts: risk tolerance (maximum acceptable risk levels that should never be exceeded) and key risk indicator thresholds (thresholds on measurable signals such as model evaluations, that trigger specific mitigations to maintain risks below the risk tolerance). Within this framework, multiple key risk indicator thresholds can be defined for a single risk tolerance level. The third draft of the EU General Purpose AI (GPAI) Code of Practice[1] requires the definition of Systemic Risk Tiers that serve multiple functions, including specifying which mitigations to implement under different conditions (third draft Measure II.1.2(List 2)(1))) and establishing acceptable risk levels (third draft Measure II.1.2(List 1)(2)(c)). This memo provides guidance on defining and operationalizing these risk tiers. It is an expanded version of parts of the feedback that we provided on the second and third Code of Practice drafts.

To inform the drafting of the Code, we proposed a hierarchy of approaches to define risk tiers that helps organize harm levels, risk scenario levels, and capability levels for general-purpose AI models in a structured, ordered way. Risk tiers can in principle be defined by regulators and/or providers. Ideally, they should be the same for all models. For each risk, multiple tiers may be defined, but at minimum, one tier must be designated as "unacceptable." This unacceptable tier effectively establishes the risk tolerance.

We think risk tiers should be defined using the following three approaches, in the order of preference listed. Developers should only use the next approach after they demonstrate that the previous approach is not feasible for the specific risk being assessed.


Decreasing relevance to tangible harm
Try 1st
Harms
The tangible harm that may result from a risk scenario
Try 2nd
Threat Profiles
A set of steps by which the risk event plays out
Fall-back
Risk Sources
A potential source of harm that gives rise to the risk
Risk pathway: from source to harm
Loss of life from failure of critical services
A terrorist creates malware with the model and deploys it in hospitals
A model has above human coding capabilities that can create malware
Unacceptable tier
Medium tier
Low tier
Unacceptable
Risk tolerance
Acceptable

[1] At the time of writing, only the third draft (i.e., the last public draft before the final version) of the Code of Practice has been published. While the final version may contain revisions, we anticipate that the core principles discussed in this paper will remain applicable to the final version.

1. Harm-based Tiers

Steps:

1) **Define:** Define which levels of risk (combinations of severity and likelihood of harm) are unacceptable in terms of traditionally-used harm indicators such as casualties, fatalities and economic damage. Optionally, also define what risk levels correspond to acceptable tiers such as "medium tier" and "low tier". These intermediate tiers provide greater granularity for determining the conditions under which specific mitigations should be implemented. For example, certain easy-to-implement safety measures might be triggered at "low tier", while more intensive interventions may be required at "medium tier". While these graduated tiers enable a more granular approach to risk mitigations, the critical requirement remains that, at all times, risks must stay below the unacceptable tier. This ensures that, regardless of the intermediate steps taken, the final risk profile of the system meets the fundamental safety objective established by the unacceptable risk tier.
    - *Example*: More than 1% chance per year of at least one serious injury (e.g., MAIS3+)[2]
2) **Connect:** For each GPAI model, connect risk sources[3] and risk scenarios[4] to harm estimates to know which could lead to unacceptable levels of harm[5]. Measurements of risk sources, such as capabilities evaluations, serve as key risk indicators, i.e., elements that can be measured and should be monitored because they indicate risk. Additionally, connect mitigations to risk sources, scenarios and harms.[6] This results in a complete risk model.
3) **Demonstrate:** For each GPAI model, demonstrate how mitigations reduce post-mitigation risk below the unacceptable risk tier accounting for all risk sources, risk scenarios and harms.

This is the preferred approach when potential harm can be directly estimated because harm is ultimately what we aim to mitigate—it represents the actual negative impacts we care about preventing. Additionally, harm based tiers are most useful for governance because they allow societal discourse on unacceptable levels of risk and standardisation of them across providers.

This approach is used in various industries as well as EU guidelines and regulations such as:

- for risks to health and safety from consumer products: Commission Implementing Decision (EU) 2019/417...on general product safety and its notification system (Table 4)
- for *catastrophic consequences* in railway safety: Commission Implementing Regulation (EU) No 402/2013...on the common safety method for risk evaluation and assessment (Annex 1, 2.5.4)

Harm is difficult to measure in two cases: where it is hard to characterize by a metric, and where the potential risk scenarios are unclear at present. Therefore, sometimes the harm-based approach is not feasible, in which case the scenario-based approach should be tried instead.

2. Scenario-based Tiers

Steps:

1) **Define:** Define what risk scenarios are unacceptable based on the likelihood of their occurrence. Optionally, also define acceptable risk scenarios of different acceptable tiers.
    - *Example*: More than 1% chance per year that a model enables an expert to develop a dangerous novel biological agent

[2] Actual definitions should be more detailed.

[3] Risk sources include model capabilities, propensities, affordances and contextual factors, or combinations thereof.

[4] The causal chains that connect risk sources to harms.

[5] This is otherwise known as risk modeling.

[6] To help minimize part of the burden of risk assessment, batches of mitigations (instead of individual ones) can be defined for each tier.

2) **Connect:** For each model, connect risk sources that could lead to those risk scenarios. Additionally, connect mitigations to risk sources and scenarios.
3) **Demonstrate:** For each model, demonstrate how mitigations reduce the likelihood of risk scenarios to acceptable tiers.

This approach should only be used when direct harm is hard to measure or foresee. An approach similar to this is currently used (though not with probabilities) in the risk management policies of most developers of advanced AI systems. In practice, the approach of numerous developers is in between scenario-based tiers and source-based tiers. The characterization of the thresholds is often less connected to harms in the real-world than we would recommend, but more connected to what the capabilities enable.

3. Source-based Tiers

Steps:

1) **Define:** Define tiers based on risk sources that could lead to harm, but how (i.e., in which risk scenarios) they could lead to harm is unclear at present.
   - *Example*: Autonomous AI research and development capabilities (e.g., RE-Bench)
2) **Connect:** For each model, directly connect mitigations to risk sources that could lead to harm.
3) **Demonstrate:** For each model, demonstrate how mitigations reduce risk sources below acceptable levels.

This approach should only be used when harm-based and scenario-based approaches are infeasible. While easiest to measure and implement, this approach is furthest removed from actual harm and should be used as a last resort.