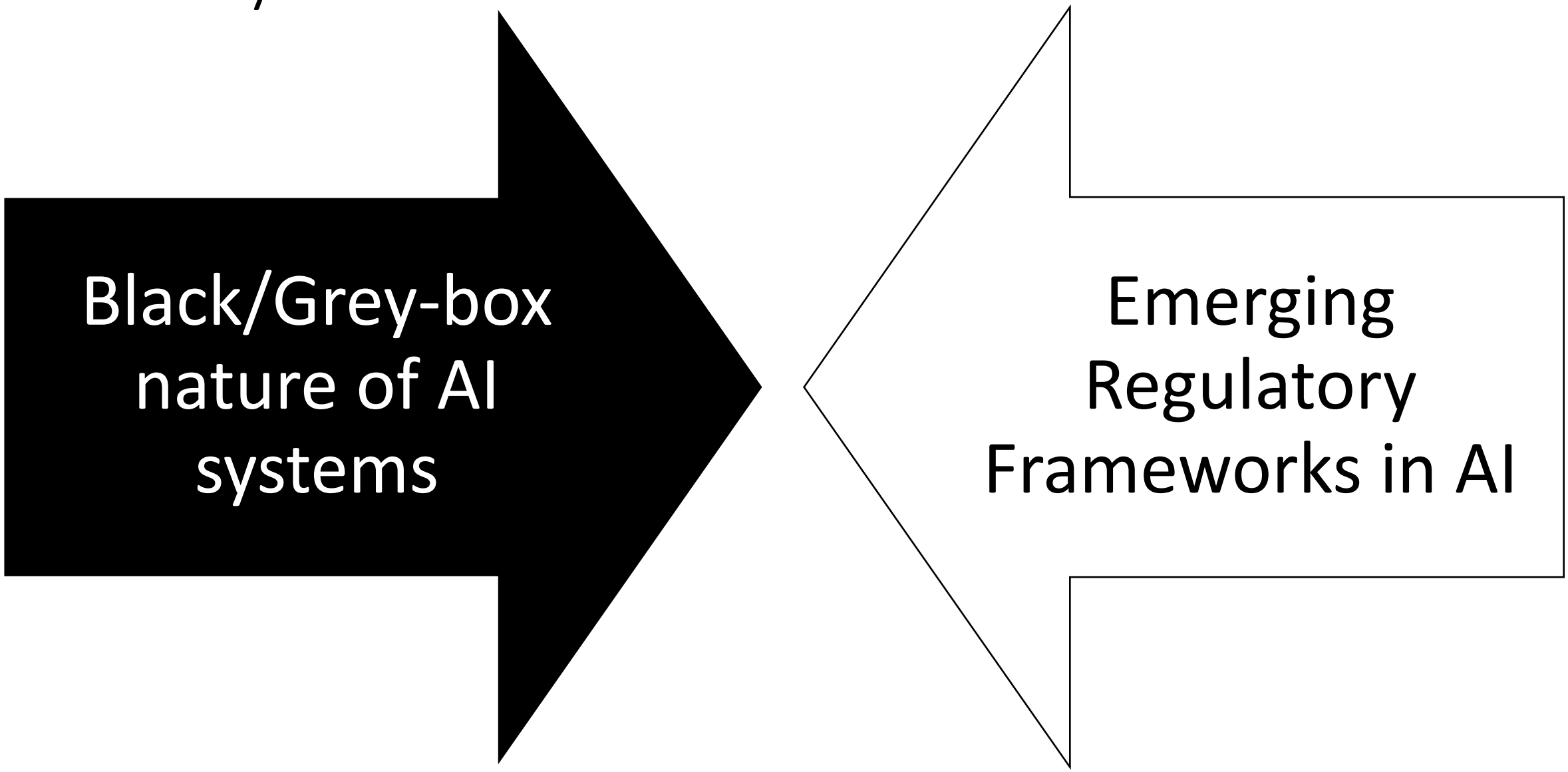
Evaluating Residual Risk of AI Systems

Ensuring Safety and Regulatory Compliance

Valentin Nikonov,
International Expert on Risk Management
Vice Chair UNECE WP.6 GRM

The Key Role of Residual Risk



Emerging/Existing Regulatory Frameworks

(Residual) Risk has become a horizontal issue in the context of AI regulation

Regulations require AI systems to have an acceptable/tolerable level of (residual) risk

(Residual) Risk should be acceptable across all hazards

OCTOBER 30, 2023

Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence

BRIEFING ROOM › PRESIDENTIAL ACTIONS

Proposal for a

REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS

{SEC(2021) 167 final} - {SWD(2021) 84 final} - {SWD(2021) 85 final}

COMMISSION IMPLEMENTING REGULATION (EU) 2022/1426

of 5 August 2022

laying down rules for the application of Regulation (EU) 2019/2144 of the European Parliament and of the Council as regards uniform procedures and technical specifications for the type-approval of the automated driving system (ADS) of fully automated vehicles

(Text with EEA relevance)

Residual Risk in Regulatory Frameworks

7.1.1. The manufacturer shall define the acceptance criteria from which the validation targets of the ADS are derived to evaluate the residual risk for the ODD taking into account, where available, existing accident data ([1]), data on performances from competently and carefully driven manual vehicles and technology state-of-the-art.

(a) Artificial Intelligence must be safe and secure. Meeting this goal requires robust, reliable, repeatable, and standardized evaluations of AI systems, as well as policies, institutions, and, as appropriate, other mechanisms to test, understand, and mitigate risks from these systems before they are put to use. It also requires addressing AI systems' most pressing security risks — including with respect to biotechnology, cybersecurity, critical infrastructure, and other national security dangers — while navigating AI's opacity and complexity. Testing and evaluations, including post-deployment performance monitoring, will help ensure that AI systems function as intended, are resilient against misuse or dangerous modifications, are ethically developed and operated in a secure manner, and are compliant with applicable Federal laws and policies. Finally, my Administration will

Example of a Regulatory Framework: EU AI Act

1. Regulation sets out requirements for a risk management process:

2. The risk management system shall consist of a continuous iterative process run throughout the entire lifecycle of a high-risk AI system, requiring regular systematic updating. It shall comprise the following steps:

(a) identification and analysis of the known and foreseeable risks associated with each high-risk AI system;

(b) estimation and evaluation of the risks that may emerge when the high-risk AI system is used in accordance with its intended purpose and under conditions of reasonably foreseeable misuse;

(c) evaluation of other possibly arising risks based on the analysis of data gathered from the post-market monitoring system referred to in Article 61;

(d) adoption of suitable risk management measures in accordance with the provisions of the following paragraphs.

2. Regulation describes risk mitigation measures for developing AI systems, such as:

- Data and data governance,
- Technical documentation,
- Record keeping,
- Quality management system, etc.

3. Regulation establishes requirements for acceptability of the **residual** risk:

any residual risk associated with each hazard as well as the overall residual risk of the high-risk AI systems is judged acceptable, provided that the high-risk AI system is used in accordance with its intended purpose or under conditions of reasonably foreseeable misuse. Those residual risks shall be communicated to the user.

Acceptable Risk and Regulatory Approval: the concept is not new

- WP.6 Recommendation R (2011) describes a Risk-Based Regulatory Framework and presents regulation as a risk mitigation tool
- It recommends, among other things, that:
  - "**All functions** of the risk management process should be consistently described in legislation that lays out the regulatory framework at a general level or for a specific sector"
  - "Regulatory authorities should establish, implement and maintain, a process for determining, analyzing, reviewing and monitoring an **acceptable level** of risk within a regulatory framework"
- According to the GPSD, a product is deemed safe whenever it complies with a given European or national legislation

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE

Recommendations on regulatory cooperation and standardization policies

Working Party on Regulatory Cooperation and Standardization Policies

UNITED NATIONS

R. Managing Risk in Regulatory Frameworks*

The Working Party on Regulatory Cooperation and Standardization policies,

Recognizing that mitigating risk that may affect society and hamper economic development is an important goal for policy-making,

Underlining that risk management is an important tool for promoting regulatory convergence at international and regional levels,

Emphasizing the role of risk management in achieving sustainable development goals, Stressing that risk-management tools are essential to enhancing the efficiency of regulatory action and of regulatory systems,

Recognizing the need of regulatory authorities, standardization, conformity assessment and accreditation bodies, as well as market surveillance authorities, economic operators, consumers, as well as other regulatory stakeholders, in promoting coherent, consistent, efficient, effective and systemic application of risk management in regulatory systems,

Taking into account international standards related to the management of risk, such as ISO 31000:2009, ISO 9001:2008, ISO/IEC 17000:2004, and other standards, including sector-specific standards, such as ISO/IEC 27001:2005,

Underlining that regulation in many cases may not be the best response to risk, and that absolute safety cannot be a regulatory outcome, as it is impossible, as well as undesirable to make the world risk-free,

Regulatory Approval of an AI system: regulatory and conformity assessment challenges

Regulatory challenge:
How safe is safe enough?
Which standards should be applied

What is the acceptable level of the residual risk?

- Challenging task for any Regulatory Authority

Tolerable level of residual risk

Conformity Assessment Challenge:

How much testing will be enough to prove that a product is safe?

Which conformity assessment schemes should be applied?

- Challenging task for any Conformity Assessment Body

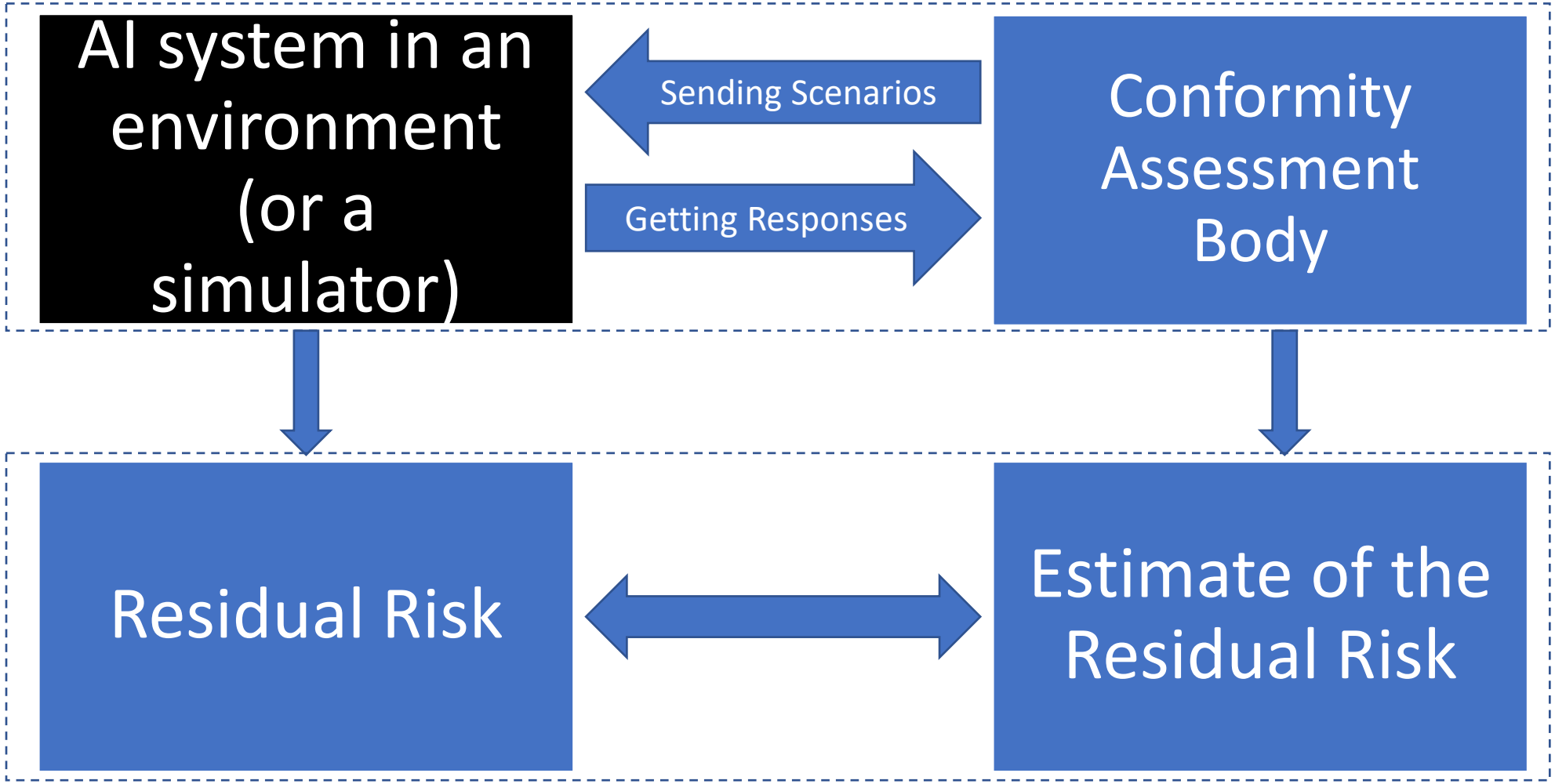
Traditional Products vs. AI Systems: Compliance with Standards is not sufficient

Traditional, Deterministic Products/Systems	**Regulating AI Systems – black/grey boxes**
• Product characteristics refer to attributes of a product (such as width, weight, etc.) • Regulation describes the regulated product itself • If a product is broken, it is broken	• Functionality is unknown/partly unknown • It is impossible to "look inside" to check how it works • System is stochastic, not deterministic
A Regulator can establish requirements for: • Products characteristics • Related processes • Production methods	• Regulations establish requirements for AI system provider **to mitigate risks** of a system • Regulations require **the residual risk of an AI system to be acceptable**
• Compliance with standards demonstrates that the risk is tolerable • Sufficient to make sure that safe products are placed on the market	• Compliance with standards demonstrates that the AI system has been developed in the risk mitigation conditions • Showing that the AI systems are safe should be based on the evaluation of the residual risk

Scenario-Based Approach for the Evaluation of the Residual Risk

Applicable for any AI system

Scenario-based Approach for Evaluating the Residual Risk of an AI system



Conformity Assessment Challenge: so many scenarios to check

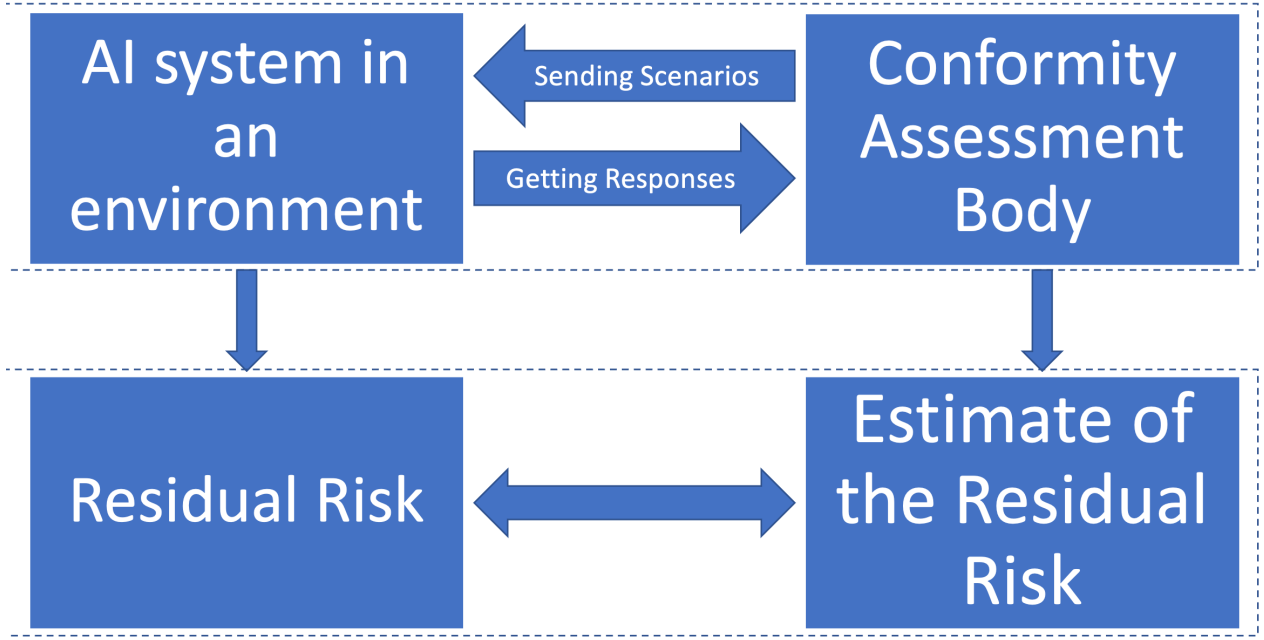
AI Systems are **Complex Systems**

AI Systems operate in **Complex Environments**

Infinity of scenarios to check

Critical considerations/key questions in Conformity Assessment of AI systems

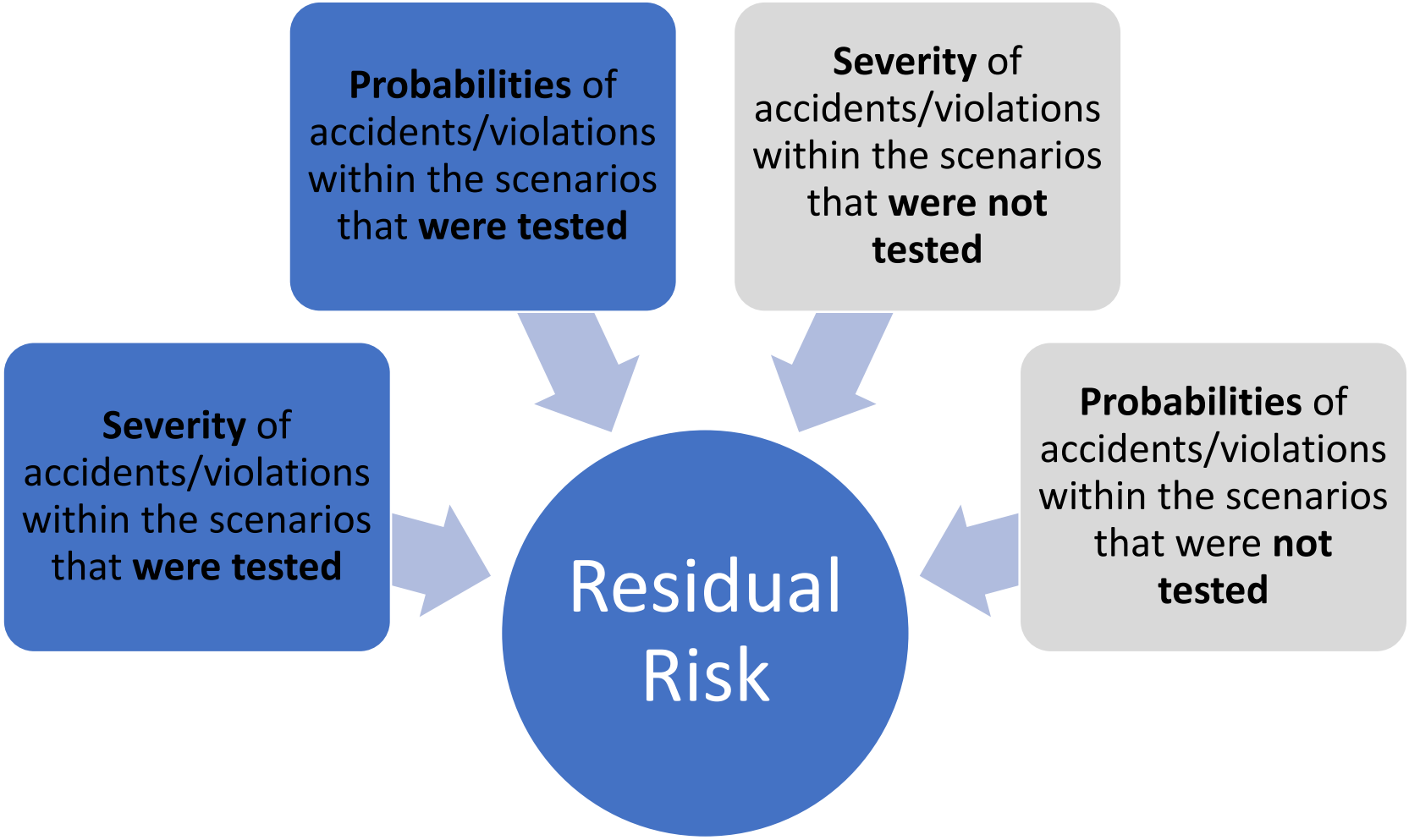
- **How to test a product:**
  - Physical test or simulation in a lab?
- **How to choose which scenarios to test:**
  - Which scenarios are most likely to happen in reality?
  - Which scenarios are most dangerous?
- **How to evaluate the responses of the tested product:**
  - How to “translate” the behavior of an AI system in metrics?
- **Can we trust the results:**
  - How can we know that we tested enough?
  - Can we trust our estimates of the residual risk?



Requirements for a Residual Risk Evaluation Framework

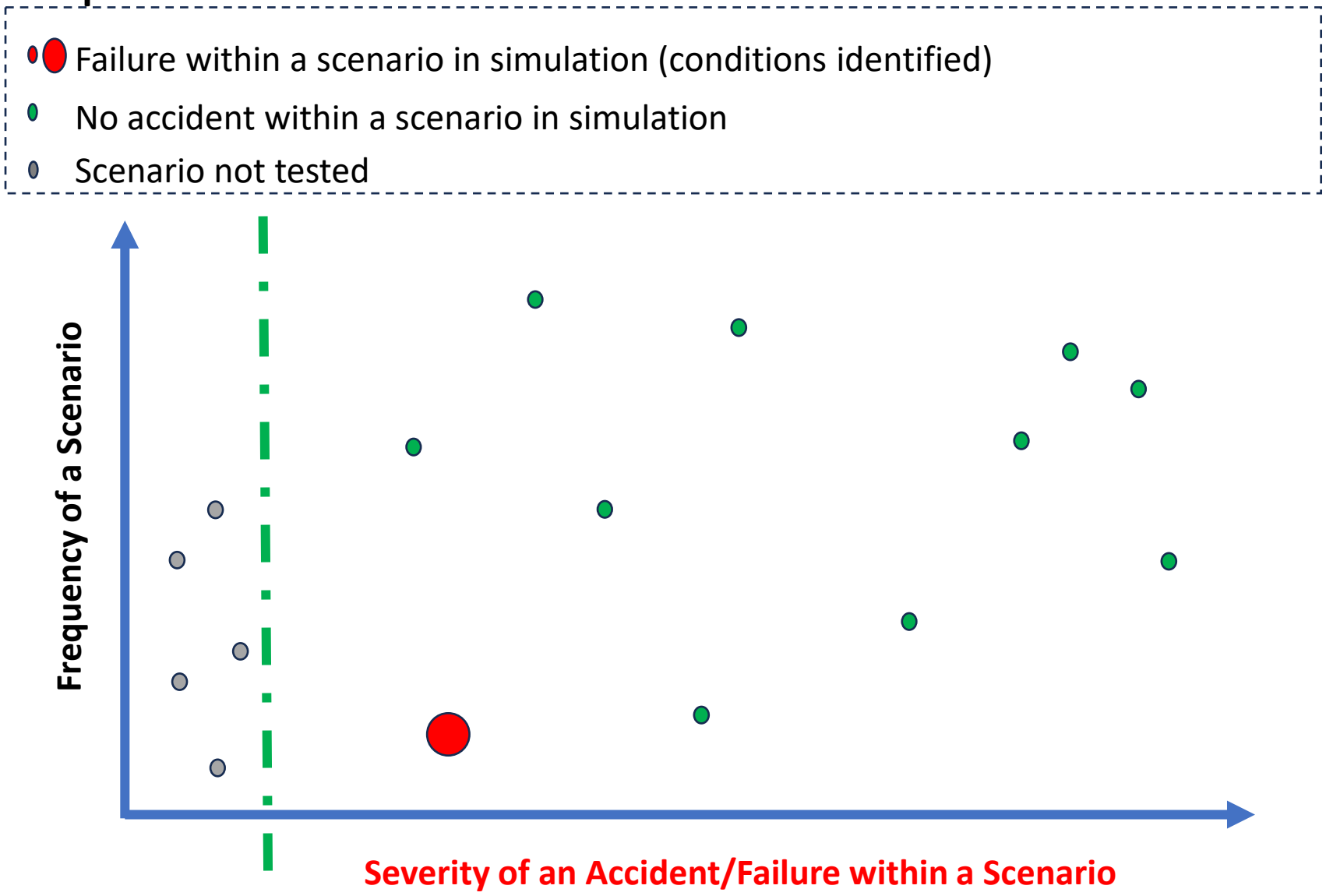
Residual Risk depends on both what was found during testing and what was tested

- Residual Risk is an important characteristic of any system placed on the market
- Residual risk is not strictly defined in the regulations that require it to be acceptable
- Residual Risk makes more sense when the “before mitigation” risk is known
- The format of the Residual Risk should be interpretable



Residual Risk: a possible format

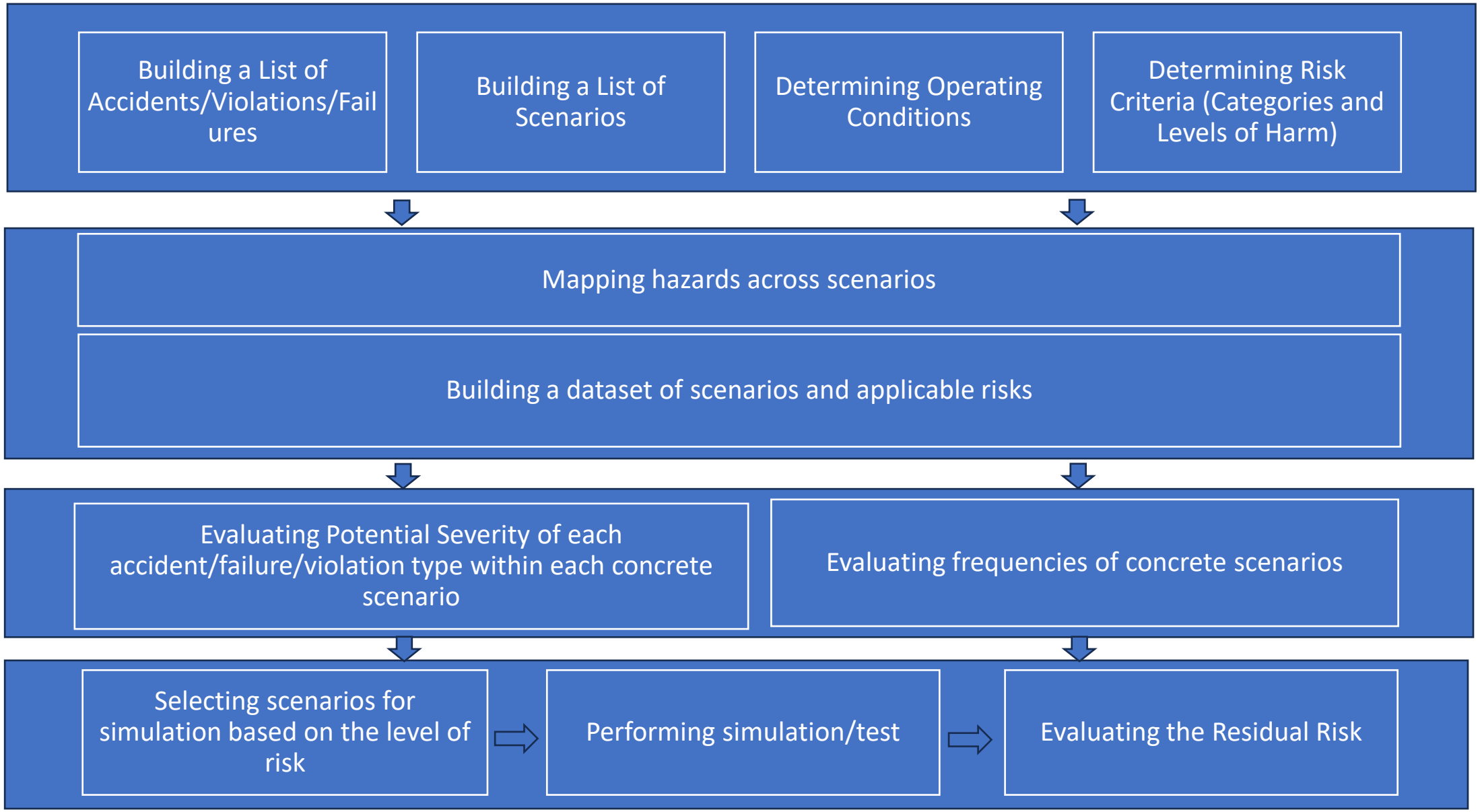
- To get the Residual Risk estimate, the risk of an accident/violation (because of the system misbehavior) should be the basis for selecting scenarios for testing
- We need:
  - Assumptions on the severity of accidents/failures within the scenarios
  - Assumptions on the expected frequencies of scenarios
- By changing assumptions, we can get different estimates of the residual risk



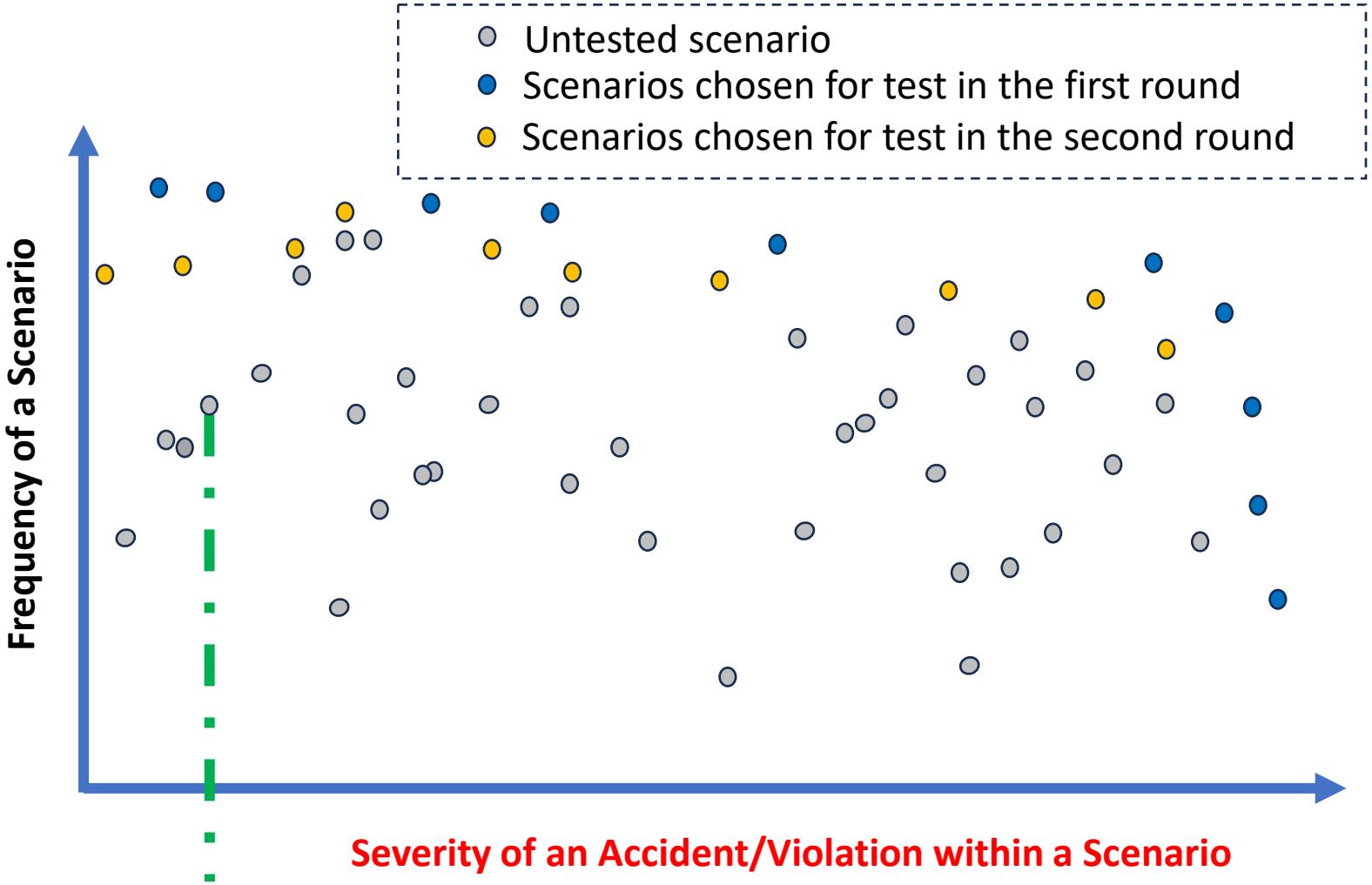
Basic requirements

- Identifying all possible hazards and risk events that could materialize during the functioning of an AI system and cause harm;
- Building a list of situations/scenarios that a system can face;
- Identifying which hazards can occur in each scenario;
- Evaluating Potential Severity of hazards in scenarios and their frequencies;
- Selecting scenarios for testing based on the level of risk: ensuring coverage of the most probable and most dangerous scenarios;
- Performing simulation/test and evaluating the residual risk.

A step-by-step process

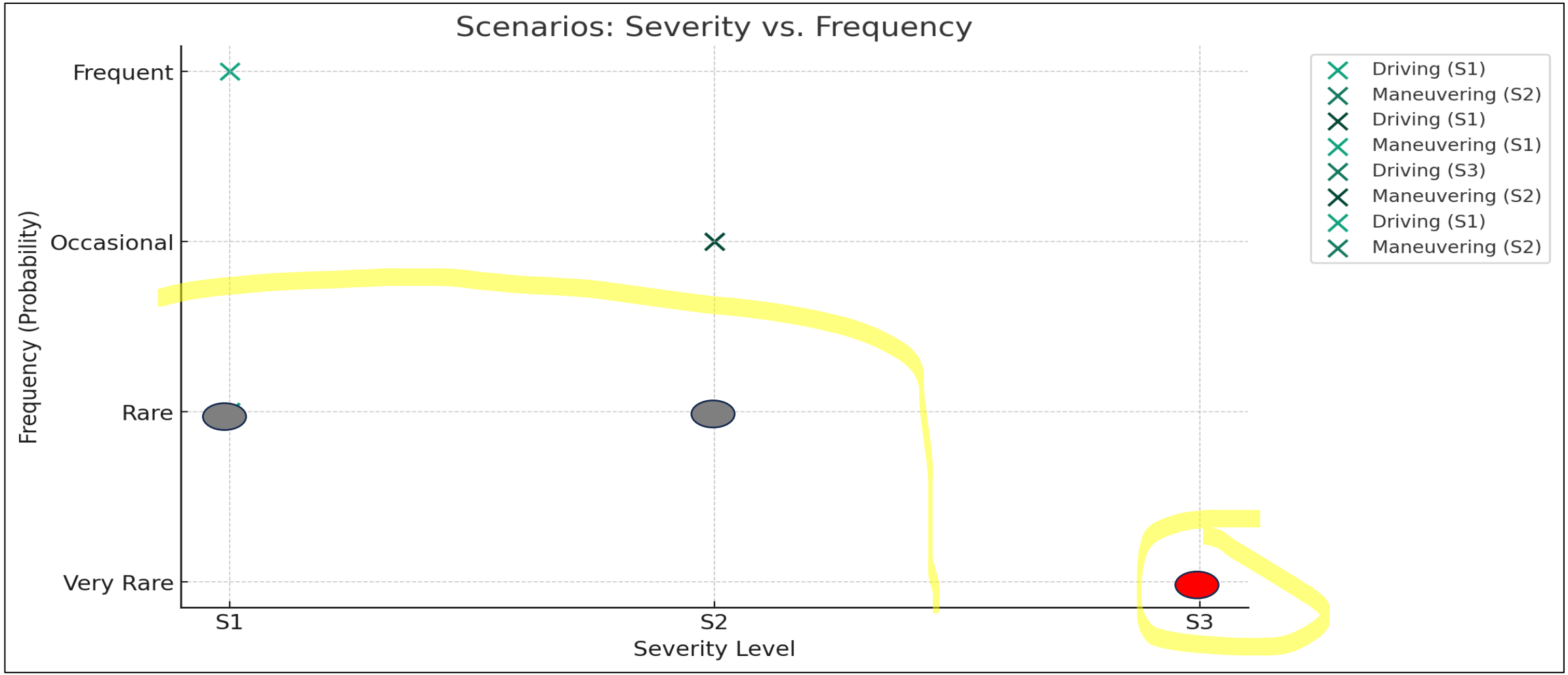


Selecting scenarios for simulation based on the level of risk



- We can test a limited number of scenarios
- By choosing scenarios that are Pareto optimal in terms of severity and frequency, we ensure that what has been tested is of higher risk than what hasn't been tested
- Picking all optimal scenarios, we cover both what is most likely and most dangerous
- Tolerable level of risk can be shown as an area on the graph

Estimating Residual Risk based on the results of the simulation



Conclusion and next steps

- Developing a comprehensive framework for evaluation of the residual risk of AI systems is essential for ensuring safety and facilitating trade
- Recommendations developed by WP.6 GRM (especially R and S) can be used in the development of the required methodologies and tools
- GRM can be a platform for international cooperation in the field and the advancement of these techniques.