ISD | Powering solutions to extremism, hate and disinformation

Identification, assessment and mitigation of systemic risks in the context of the Digital Services Act

4. July 2025

Agenda

1 Report Background

2 Systemic Risks & Categorisation

3 Risk Assessment Process

4 Challenges

5 Q&A

Report Background & ISD Team

ISD Team

Dr. Jan Beyer

Director of Digital Methods (EU)

Beatriz Saab

Digital Methods and Policy Manager

Leonie Oehmig

Digital Policy Associate

Alexander Hohlfeld

Digital Policy Analyst

Mauritius Dorn

Director of Public Affairs

Anna Katzy-Reinshagen

Analyst

Carolin von Bredow

Digital Policy Associate

About ISD's Report

- Report was written for the **German Digital Service Coordinator** (DSC): **Bundesnetzagentur** (BNetzA)

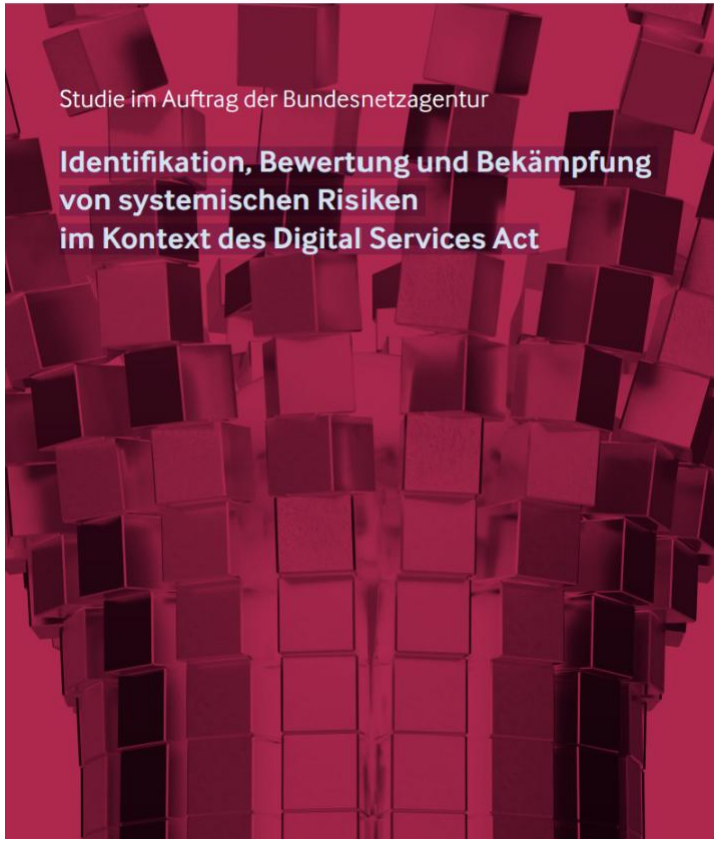
- The overall aim was to develop a coherent and practically applicable approach for the
  - ➤ **Identification**
  - ➤ **Assessment**
  - ➤ **Mitigation**

  of systemic risks on Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs).

- Timeframe: 01.10.2024-30.11.2024

ISD Institute for Strategic Dialogue

Im Auftrag von: Bundesnetzagentur



Identifying, assessing and combating systemic risks in the context of the Digital Services Act

Scope

Identification of Risks

Develop a clear approach to identify systemic risks on very large online platforms and search engines.

Assessment Framework

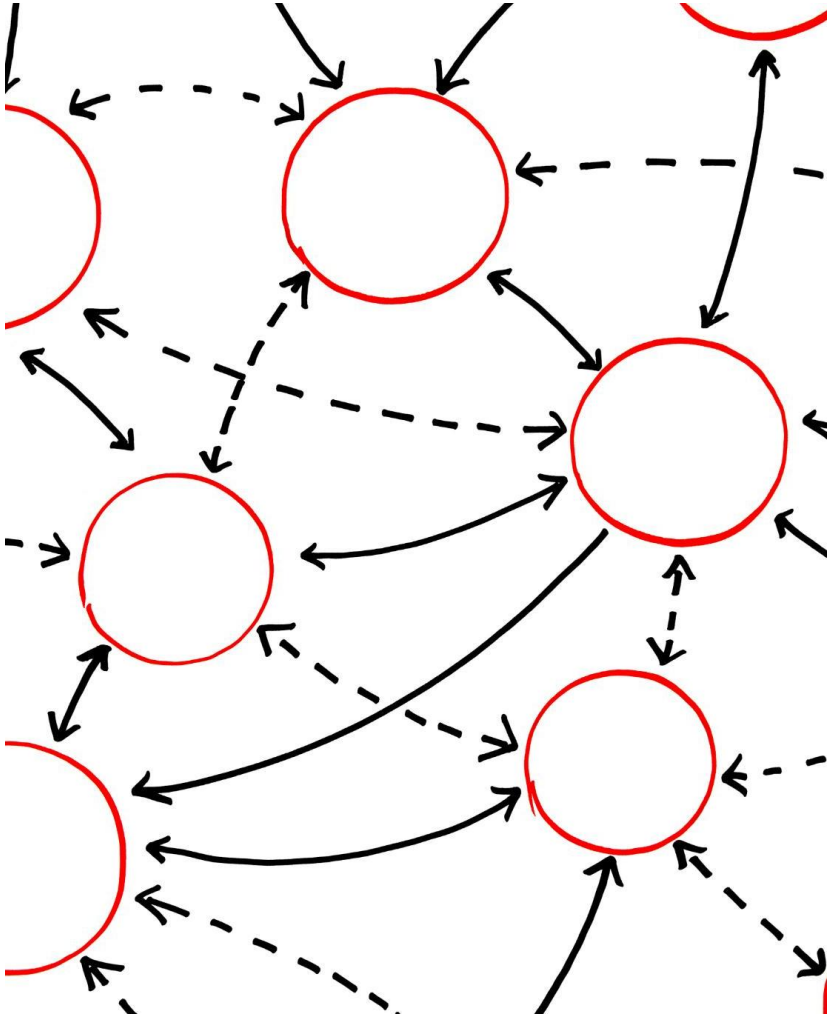
Provide a structured framework to assess these risks and determine their potential impact.

Mitigation Strategies

Develop effective indicators and strategies to mitigate the identified systemic risks.

Multi-Stakeholder Enforcement

Support a comprehensive enforcement structure involving multiple stakeholders for effective implementation.



ISD's Approach to Systemic Risks

What are systemic risks?

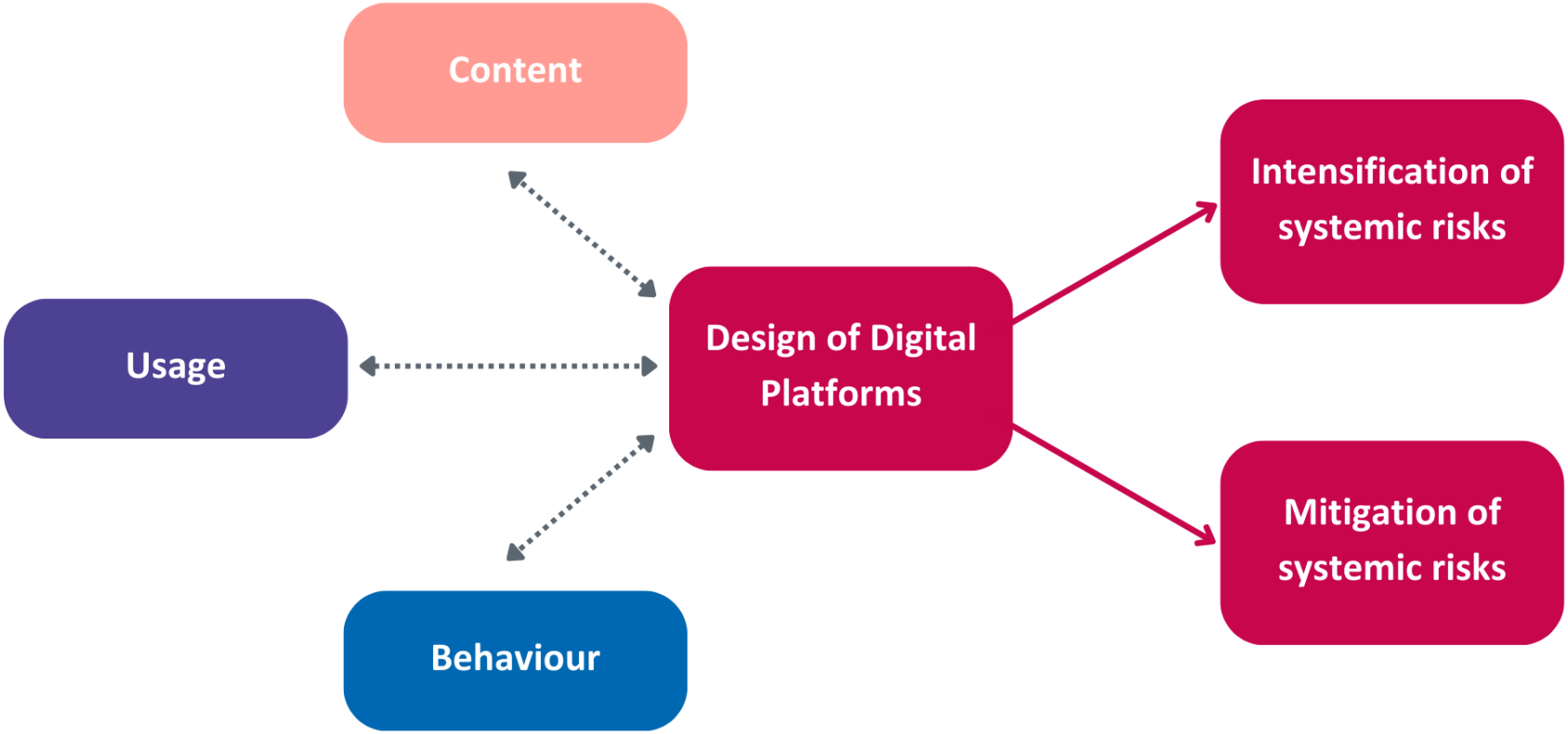
- Although Art. 34.1 describes different categories of systemic risks, there is no clear definition in the DSA of what is meant by "systemic" risks.
- The DSA introduces the criteria of "severity" and “likelihood" of systemic risks as an assessment standard. However, it remains unclear which specific elements constitute the "systemic" nature of risks.
- ISD’s working definition of systemic risks:

> *Systemic risks refer to potential online damage (e.g. viral disinformation) that is essentially caused by the **functionalities** of services, their **use** or deliberate **manipulation**. Their **probability is high,** and their social **impact** goes far **beyond individual damage.***

ISD's Risk Categorisation

Content-Related	Usage-Related	Behaviour-Related
Content-related risks refer to risks arising from the creation, dissemination or amplification of illegal content.	Usage-related risks arise from the way in which online platforms are used, regardless of whether the design of the platform intends this or not.	Behavioural risks focus on actors who exploit vulnerabilities or terms of use of online platforms to carry out illegal or harmful activities.

ISD's Risk Categorisation

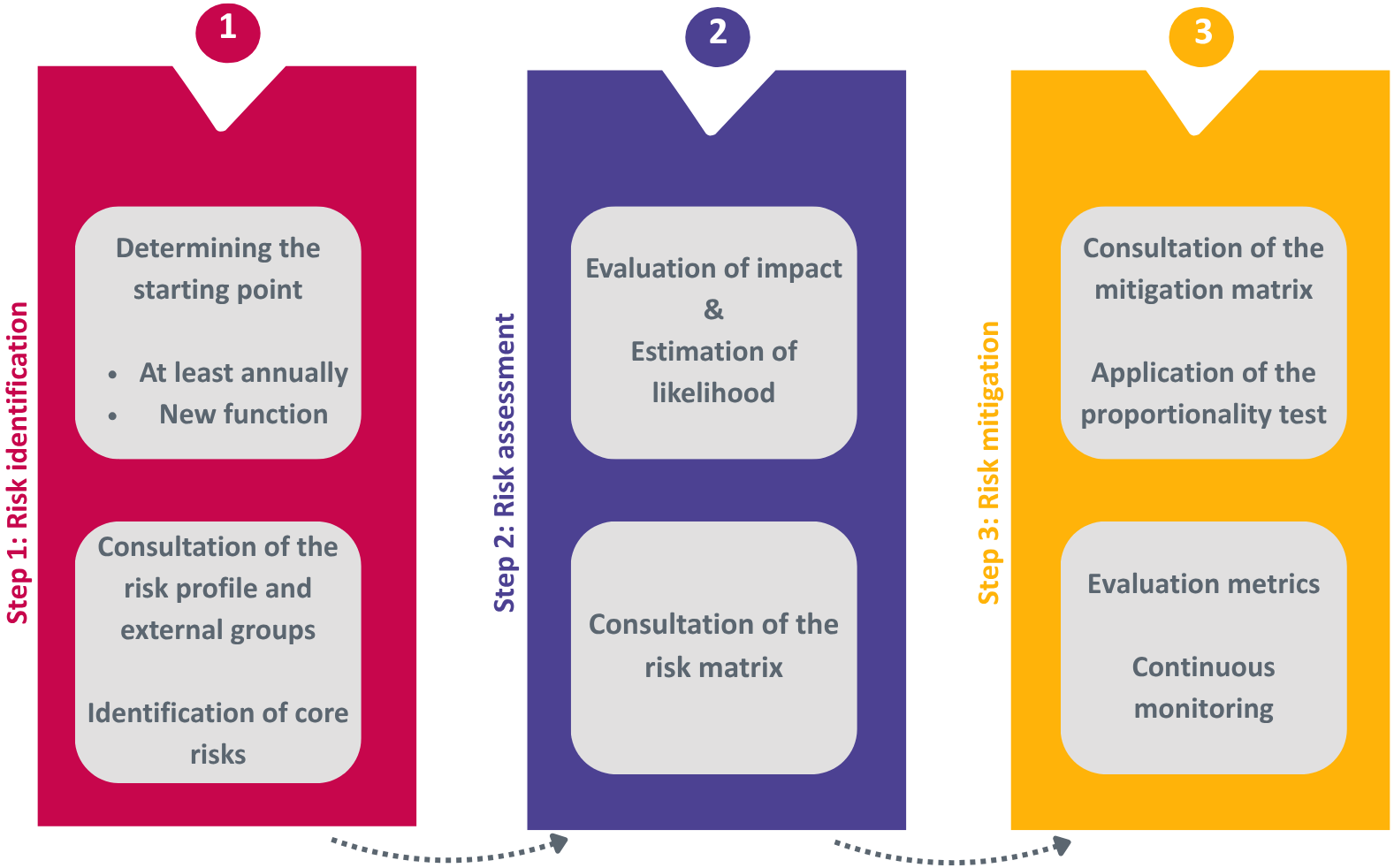


Risk Assessment Process

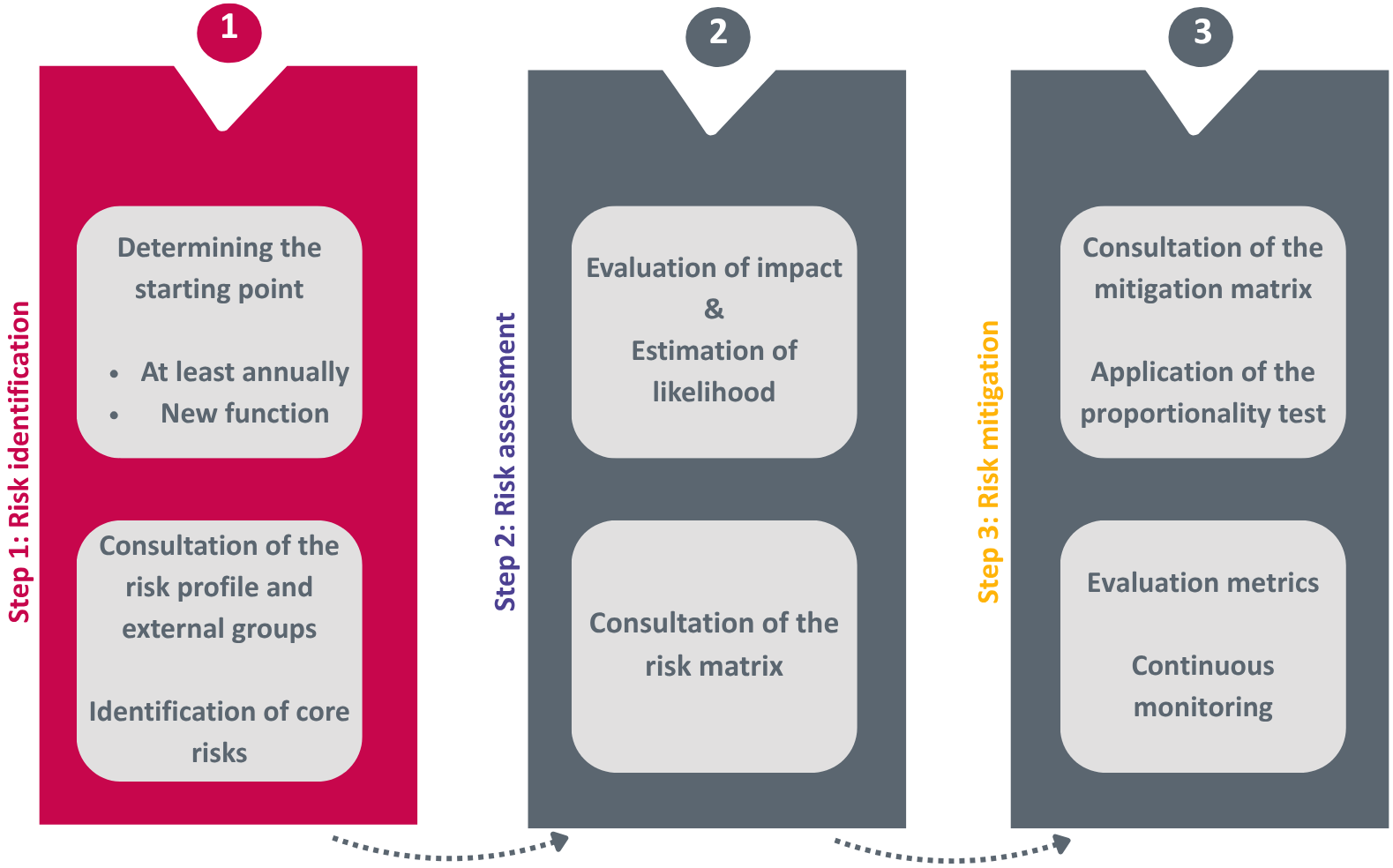
Adaptation from Human Rights Due Diligence Framework

- The risk assessment of VLOPs and VLOSEs must be fundamentally linked to human rights considerations due to the profound impact that these online platforms have on public debate, privacy and individual well-being.
- The concept of human rights due diligence, set out in the UNGPs, serves as the cornerstone of the ISD's risk assessment methodology

ISD's Risk Assessment Process



ISD's Risk Assessment Process



Step 1: Risk profile

Step 1: Risk factors

Risk Factors	Guiding Questions
Type of Platforms	Which services are offered? (e.g., social media platform, video-sharing platform, adult platform, online marketplace, app store, travel and accommodation platform, online encyclopedia, search engine)
Size and User Base	How many average monthly active users does the service have in the EU? (continuous measurement) What demographic characteristics define the users?
Business Model	Does the service use any of the following design options to generate revenue?
Design of Recommendation Systems and Other Relevant Algorithmic Systems (34(2a))	Does the service contain recommendation systems (e.g., feed, search)? Does the service consider feedback signals? Does the service generate personalized or non-personalized recommendations? Are the service's recommendations of content, products, or services based on a strategy of exploitation or discovery?
Content Moderation Systems (34(2b))	Does the service include the following measures for content moderation? Does the service include the following measures for dealing with accounts?
General Terms and Conditions (34(2c))	Does the service allow vulnerable consumers access to its services? Does the service allow content for adults?
Enforcement of General Terms and Conditions (34(2c))	Does the service have any of the following functions that relate to how users identify each other? Does the service offer any of the following security measures?
Systems for Selection and Display of Advertising (34(2d))	Does the service offer advertising and monetization opportunities? Does the service offer any of the following functions regarding advertising?
Data-Related Practices (34(2e))	Does the service process personal data in accordance with the GDPR?

Step 1: Risk Profile based on ISD's framework

Risk factor: Type of platforms

Social Media Platforms

- **Usage-Related Risks**
  - Behavioural addiction (doomscrolling, auto-play, likes)
  - Mental health

Content-Related Risks

- Illegal content
- Influential content (self-perception, self-harm, misinformation)

Behaviour-Related Risks

- Disinformation campaigns
- Grooming

Adult Platforms

- **Usage-Related Risks**
  - Use by minors
  - Mental health

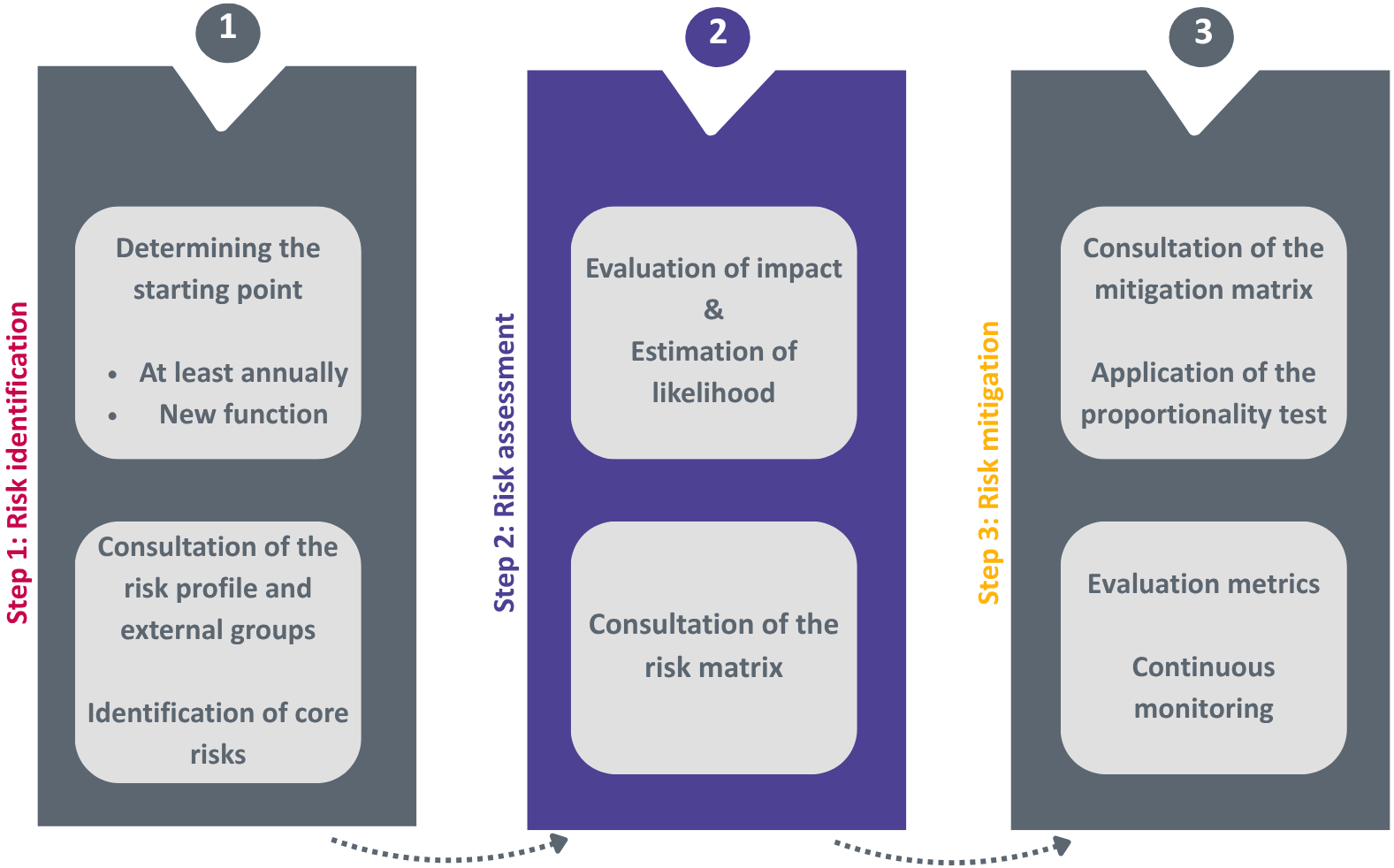
Content-Related Risks

- Illegal content
- Violent content
- Non-consensual content

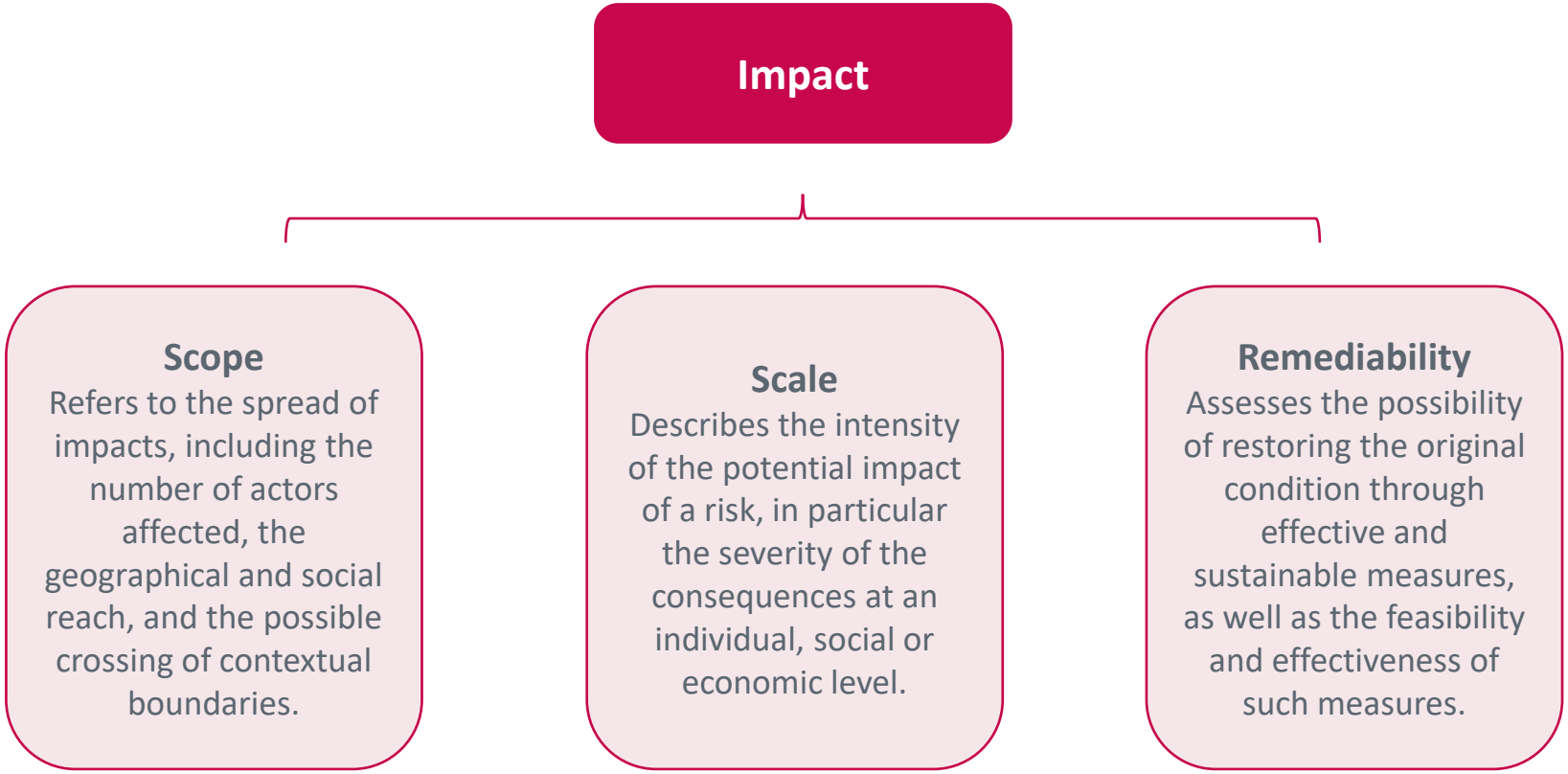
Behaviour-Related Risks

- Revenge pornography
- Doxing
- Extortion

ISD's Risk Assessment Process



Step 2: Evaluation of the impact



Impact Indicators

Type of Risk / Indicator	Scope	Scale	Remediability
Usage-Related Risks	• Number of vulnerable users exposed to harmful content • Cross-platform impact on people at risk • Demographic and geographical distribution of affected users	• Severity of impact on mental health • Long-term effects on vulnerable users • Socio-economic impacts	• Availability of psychosocial support • Speed and simplicity of restoring security
Content-Related Risks	• Number of users who meet illegal content on the platform • Scope of illegal content shared across communities or regions	• Severity of psychological or physical harm • Impact on public health/safety • Socio-political impact	• Restoring user trust • Support for psychological recovery • Institutional/legal remedies
Behaviour-Related Risks	• Number of users affected by harassment or abusive behaviour on the platform • Cross-platform reach of harmful • campaigns	• Extent of stress or damage • Impact on public trust and the reputation of the platform • Legal consequences of abuse campaigns	• Reversal of the effects of campaigns • Mitigation of artificial behaviours • Availability of support/recovery for affected users

Step 2: Estimation of the likelihood

Likelihood Indicators

Type of Risk / Indicator	Platform Features	Historical data and trends	Contextual factors
Usage-Related Risks	• Type of services • Size and user base • Business model • Design of recommendation systems and other relevant algorithmic systems • Content moderation systems • Terms and conditions • Systems for selecting and displaying advertising • Data-related practices	• Patterns indicating frequent exposure of at-risk users to triggering or harmful content. • Rates of reported mental health crises, distress, or self-harm incidents related to platform use. • High engagement with specific harmful communities	• Socio-economic or political instability that influences the behaviour of vulnerable users • Trends increased use of the platform by vulnerable users during a crisis or instability. • Demographic susceptibility to manipulation
Content-Related Risks		• Frequency and scope of uploading illegal content • Frequency of extremist or hateful content in the past	• Increased creation of illegal content in times of social unrest or crisis events • Influence of external events
Behaviour-Related Risks		• Frequent occurrence of abusive, harassing or trolling behaviour • Recorded data of disinformation or manipulation attempts.	• External events that reinforce abusive or trolling behaviour • Coordinated disinformation/FIMI or harassment campaigns in response to current events

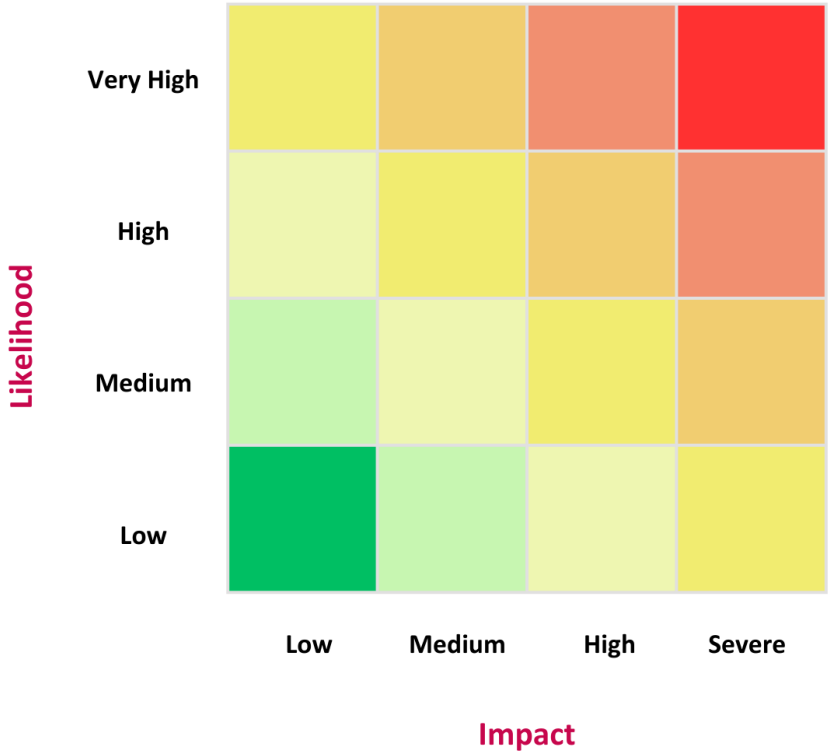
Step 2: Risk Matrix

Efficient risk mitigation:

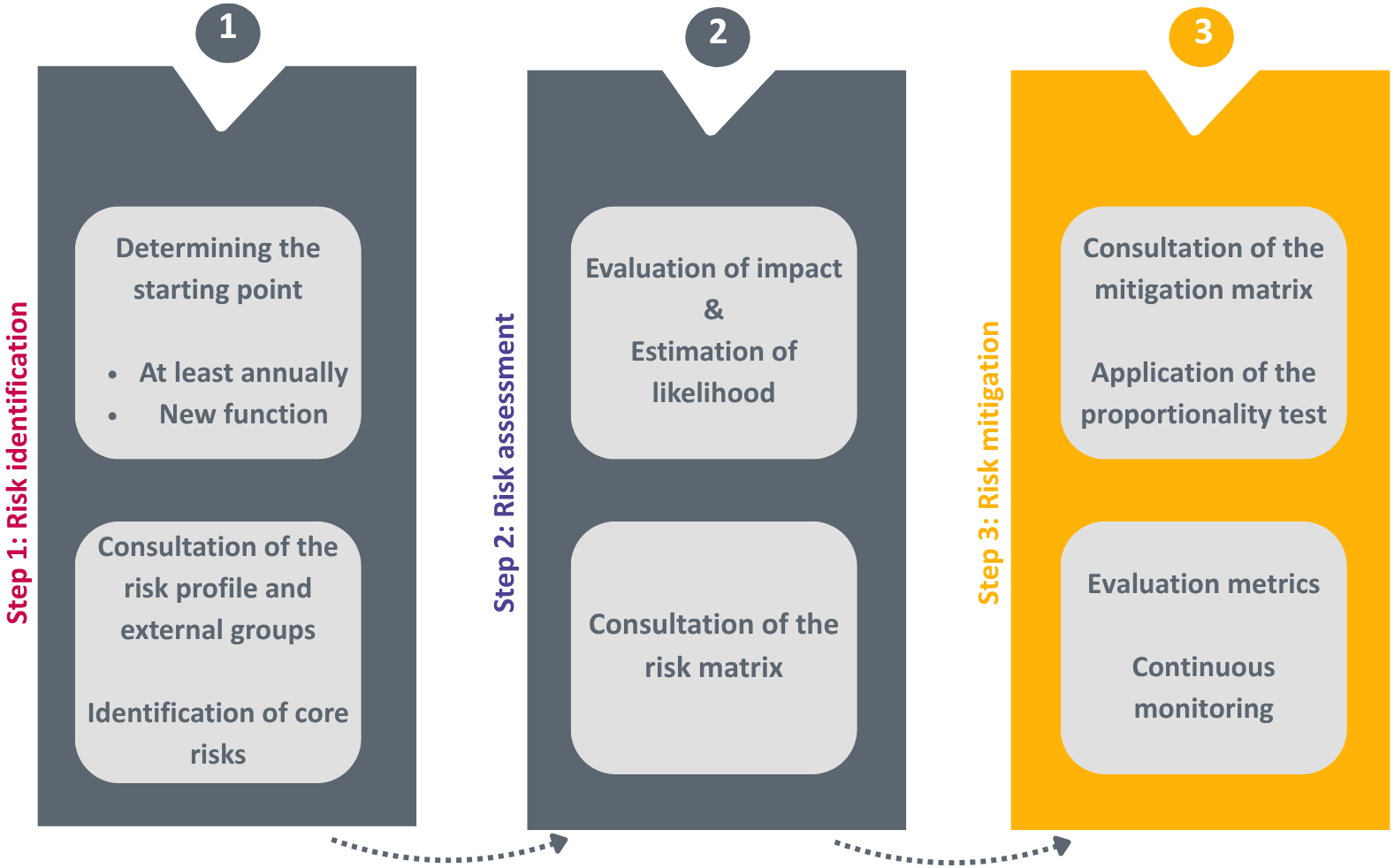
The matrix supports platforms in prioritising risks according to impact and probability. Supervisory authorities can use it for targeted monitoring and compliance with regulations.

Dynamic adjustment of limit values:

Limit values are continuously reviewed and adapted to new data. Scientific validation and stakeholder involvement ensure that they are up to date.



Step 3: Risk Mitigation



Mitigation measures

Goal	Mitigation measure	Specific mitigation measure	Category	Risk (Excel)	Risk dimension	Source
Protection of the integrity of electoral processes	Internal incident response plans for emergency situations	Establishment, coordination, and testing of an emergency plan, including red-teaming exercises; involvement of top management, as well as mapping of stakeholders within the organization involved in the emergency plan	Internal processes	25, 26	Usage-related	Guidelines on the Mitigation of Systemic Risks for Electoral Processes

Step 3: Mitigation Measures - Selection and Assessment

Are the measures suitable to achieve the intended goal?	**Suitability**
Are the measures necessary to achieve the intended goal?	**Necessity**
Will individuals bear a disproportionate burden compared to the intended goal?	**Proportionality**
Are the mitigation measures effective?	**Effectiveness**

Assessment of mitigation effectiveness

1. **Attribution challenge:** Mitigation measures may overlap, address multiple risks, cause new ones, or be influenced by external factors, making it hard to isolate their actual effect.
2. **Lack of scientific consensus:** Due to complex subject matter and limited data access, there is often no clear agreement on which measures are effective or what criteria define effectiveness.
3. **Secondary risks and trade-offs:** Measures like deplatforming may reduce harmful content but also risk infringing on fundamental rights such as freedom of expression, raising ethical and legal concerns.

Challenges

Study Limitations

Framework Requires Refinement

The exploratory nature of the study means that the proposed framework is a first attempt and requires further refinement.

Measurement Challenges

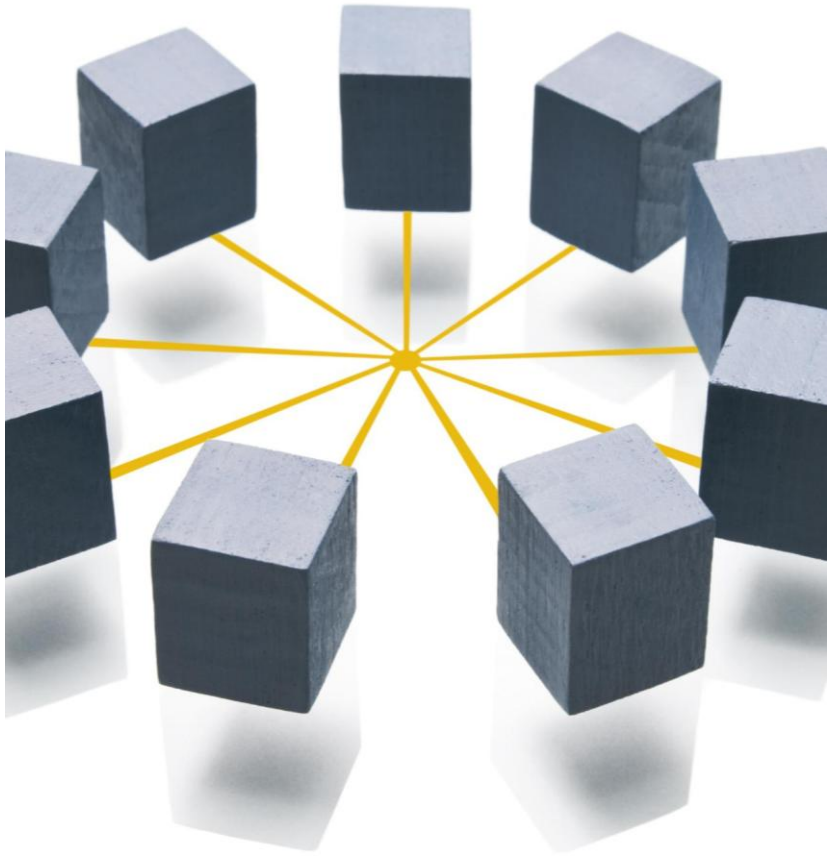
Challenges in measurement arise due to the absence of unified metrics, indicating a need for further testing of proposed indicators.

Lack of Process Insights

The lack of public risk assessments resulted in missing practical insights, affecting the applicability of the study.

Data Access Constraints

Limited data access has constrained the full validation of the proposed framework, posing a significant limitation to the study.



Discussion

Importance of Standardization

Standardization in public risk assessments is crucial to ensure safety and consistency in various practices.

Future Standard Setters

Entities providing detailed guidance on risk assessments may become the future standard setters in the industry.

Balancing Benchmarks and Flexibility

Creating strict benchmarks while allowing flexibility in the framework is a significant challenge in risk assessments.

Thank you.

Beatriz Saab
Digital Methods and Policy Manager
bs@isdglobal.org