Severity Escalation Protocol

A Universal Severity Classification System
for the Harm Blindness Framework

Version 1.0 Draft
February 18, 2026

Real Safety AI Foundation
Travis Gilly, Executive Director
t.gilly@ai-literacy-labs.org

realsafetyai.org

Contents

Part 1: Existing Severity Classification Systems by Domain

The following catalog documents how major industries and regulatory domains classify severity of harm. Each system represents decades of refinement within its specific context. The SEP synthesizes these into a universal classification that the Harm Blindness Framework can apply across any domain.

1. Safety Engineering and Industrial Systems

IEC 61508 Safety Integrity Levels (SIL)

SIL is the international standard for functional safety of electrical, electronic, and programmable electronic systems. It classifies risk based on the probability of dangerous failure per hour.

- SIL 1 (Lowest): Minor injury possible. Probability of failure on demand 0.1 to 0.01.
- SIL 2: Serious injury possible. Probability 0.01 to 0.001.
- SIL 3: Single death or multiple serious injuries. Probability 0.001 to 0.0001.
- SIL 4 (Highest): Multiple deaths, catastrophic. Probability 0.0001 to 0.00001.

Classification method: Risk graph considering severity, frequency of exposure, probability of avoidance, and probability of unwanted occurrence.

Key insight for SEP: SIL separates severity of consequence from probability of occurrence. Both matter independently.

2. Automotive

ISO 26262 Automotive Safety Integrity Levels (ASIL)

The automotive-specific adaptation of IEC 61508, governing vehicle electronic systems.

- QM (Quality Management): No safety requirement beyond standard quality processes.
- ASIL A (Lowest): Light to moderate injuries, low probability.
- ASIL B: Severe to life-threatening injuries, moderate probability.
- ASIL C: Life-threatening to fatal injuries, high probability.
- ASIL D (Highest): Life-threatening to fatal injuries affecting multiple people, very high probability.

Classification method: Three-factor matrix of Severity (S0-S3), Probability of Exposure (E1-E4), and Controllability (C1-C3).

Key insight for SEP: Controllability (can the affected person avoid the harm?) is a critical dimension most other systems miss.

3. Aviation

DO-178C Design Assurance Levels (DAL)

Software assurance standard for airborne systems and equipment.

- DAL E (Lowest): No effect on aircraft operational capability or pilot workload.
- DAL D: Minor effect. Slight reduction in safety margins, slight increase in pilot workload.
- DAL C: Major effect. Significant reduction in safety margins, physical discomfort to occupants.
- DAL B: Hazardous effect. Serious or fatal injury to a small number of occupants.
- DAL A (Highest): Catastrophic. Failure conditions preventing continued safe flight and landing.

Classification method: Failure condition analysis tied to system function within the aircraft.

Key insight for SEP: Aviation classifies based on the consequence of system failure, not normal operation. The Harm Blindness Framework should consider both intended operation harms and failure mode harms.

4. Medical Devices

FDA Risk Classification (US) and EU Medical Device Regulation (MDR)

FDA classifies devices into three risk categories. EU MDR uses four classes with sub-categories.

- FDA Class I: Low risk. Tongue depressors, bandages. General controls sufficient.
- FDA Class II: Moderate risk. Powered wheelchairs, infusion pumps. Special controls, 510(k) clearance.
- FDA Class III: High risk. Implantable pacemakers, heart valves. Premarket approval with clinical evidence.
- EU MDR adds Class IIa (low-medium) and IIb (medium-high) between I and III.

Classification method: Based on invasiveness and duration of contact with the body.

Key insight for SEP: "How deeply embedded is this system in the user's life, and how hard is it to remove?" is directly relevant to AI systems.

5. Cybersecurity

NIST SP 800-60 Impact Levels and CVSS

NIST uses three impact levels. CVSS scores vulnerabilities on a 0-10 scale factoring both exploitability and impact.

- NIST Low: Limited adverse effect. Minor financial loss, minor mission degradation.
- NIST Moderate: Serious adverse effect. Significant harm to individuals not involving loss of life.
- NIST High: Severe or catastrophic adverse effect. Major mission degradation, loss of life.
- CVSS Critical (9.0-10.0): Trivially exploitable, complete system compromise.

Classification method: Separation of exploitability characteristics from impact assessment.

Key insight for SEP: The separation of exploitability from impact maps well to the Harm Blindness Framework's distinction between how likely a harm is versus how bad it is.

6. Nuclear

IEC 61226 Safety Categories and INES Scale

IEC 61226 classifies system roles in safety (A, B, C). INES classifies events on a 0-7 scale.

- Category A: Systems playing a principal role in achieving safe state.
- INES 0-1: Deviation or anomaly, no safety significance.
- INES 2-3: Incident with near-accident conditions.
- INES 4-5: Accident with limited to significant off-site risk (e.g., Three Mile Island).
- INES 6-7: Major accident (e.g., Chernobyl, Fukushima).

Classification method: Distinguishes between system role (how important it is to safety) and event severity (what actually happened).

Key insight for SEP: A system's centrality to harm prevention matters independent of whether harm has occurred yet.

7. Chemical and Hazardous Materials

GHS (Globally Harmonized System of Classification and Labelling)

Uses hazard categories within each hazard class, classified by empirical data.

- Category 1 (Highest): Fatal, causes serious damage. Signal word "Danger."
- Category 2: Toxic, causes damage.
- Category 3: Harmful. Signal word "Warning."
- Category 4 (Lowest): May cause harm.

Classification method: Purely outcome-based; classified by empirical data (LD50 values, test results).

Key insight for SEP: Intent does not change severity. A product that kills people is Category 1 whether it was designed to kill or designed to help.

8. Food Safety

HACCP (Hazard Analysis and Critical Control Points)

Identifies Critical Control Points where monitoring is essential to prevent, eliminate, or reduce hazards.

- Life-threatening: Botulism, allergic anaphylaxis, chemical poisoning.
- Serious: Salmonella, E. coli infections requiring medical treatment.
- Moderate: Minor illness, short duration.
- Low: Unlikely to cause illness in normal populations.

Classification method: Identifies specific points in a process where things go wrong, not just overall product risk.

Key insight for SEP: HACCP's checkpoint-based approach mirrors the Harm Blindness Framework exactly.

9. Construction and Civil Engineering

Eurocode Consequence Classes

Classifies structures by consequence of failure, not by structure type.

- CC1 (Low): Low consequence for loss of human life. Agricultural buildings, rarely occupied structures.
- CC2 (Medium): Medium consequence. Residential, office buildings, moderate occupancy.
- CC3 (High): High consequence. Grandstands, concert halls, high-rise structures, bridges.

Classification method: Classification by occupancy and consequence of failure.

Key insight for SEP: Scale of affected population is a primary classifier. A poorly built house differs from a poorly built stadium.

10. Financial Services

Basel Framework Risk Categories

Classifies credit, market, and operational risk. SIFI designation identifies systemically important institutions.

- Operational Risk Low: Minimal financial impact.
- Operational Risk Moderate: Significant financial impact.
- Operational Risk High: Major impact, potential contagion.
- Systemic (SIFI): Threatens financial system stability.

Classification method: Explicitly classifies systemic risk based on interconnectedness, size, substitutability, and complexity.

Key insight for SEP: The Harm Blindness Framework should capture cascade potential, not just direct harm.

11. Environmental

EPA Risk Assessment and Environmental Impact Assessment

Risk characterization based on hazard identification, dose-response, exposure assessment, and risk characterization.

- Negligible: No measurable environmental change.
- Minor: Localized, reversible change within natural variation.
- Moderate: Measurable change requiring mitigation, reversible with intervention.
- Major: Significant change, difficult to reverse, widespread impact.
- Critical: Irreversible change, ecosystem collapse, species extinction.

Classification method: Four-step risk characterization process.

Key insight for SEP: Environmental assessment explicitly tracks reversibility. A harm you can undo is fundamentally different from one you cannot.

12. Pharmaceutical

Adverse Event Classification (CTCAE)

Five-grade adverse event classification used in clinical trials and post-market surveillance.

- Grade 1 (Mild): Asymptomatic or mild symptoms, no intervention needed.
- Grade 2 (Moderate): Minimal, local, or non-invasive intervention needed.
- Grade 3 (Severe): Significant but not immediately life-threatening; hospitalization.
- Grade 4 (Life-threatening): Urgent intervention indicated.
- Grade 5 (Death): Death related to adverse event.

Classification method: Direct outcome classification with explicit death category.

Key insight for SEP: Pharmaceutical classification names death directly, not as euphemism. The Harm Blindness Framework should be equally direct.

13. Failure Analysis (Cross-Domain)

FMEA (Failure Mode and Effects Analysis)

Three-factor risk model using Severity (1-10), Occurrence (1-10), and Detection (1-10). Risk Priority Number = Severity x Occurrence x Detection.

- Severity 1-3: No effect to minor disruption.
- Severity 4-6: Moderate effect, customer dissatisfied.

- Severity 7-8: High severity, system inoperable, potential safety issue.
- Severity 9: Potential safety hazard with warning.
- Severity 10: Safety hazard without warning, regulatory noncompliance.

Classification method: Three-factor model (severity, occurrence, detection) producing Risk Priority Number.

Key insight for SEP: Detection (can we catch this before it causes harm?) maps directly to checkpoint timing. Early checkpoints increase detection; late checkpoints decrease it.

14. Privacy and Data Protection

GDPR Data Protection Impact Assessment (DPIA)

Required when processing is "likely to result in a high risk to the rights and freedoms of natural persons."

- Low: Processing unlikely to affect individuals.
- Medium: Processing could affect individuals but harm is limited.
- High: Processing could significantly affect individuals, particularly vulnerable groups.
- Very High: Physical harm, discrimination, financial loss, damage to reputation.

Classification method: Risk assessment with explicit vulnerable group consideration.

Key insight for SEP: GDPR explicitly calls out vulnerable groups as an escalation factor. The same processing that is "medium" for general population becomes "high" for children or disabled people.

Part 2: Synthesis - Universal Dimensions of Harm Severity

Across all 14 classification systems, the following seven dimensions consistently determine severity. These are the universal axes that the SEP captures.

Dimension 1: Consequence Severity

What is the worst realistic outcome?

- **Negligible**: No measurable harm.
- **Minor**: Inconvenience, temporary discomfort, small financial loss. Fully self-correcting.
- **Moderate**: Significant discomfort, meaningful financial loss, temporary impairment. Correctable with intervention.
- **Serious**: Hospitalization, major financial harm, sustained psychological trauma, loss of livelihood.
- **Severe**: Permanent disability, destruction of livelihood, forced displacement, systemic discrimination.
- **Fatal/Catastrophic**: Death, mass casualty, civilizational or ecosystem-level irreversible harm.

Dimension 2: Reversibility

Can the harm be undone?

- **Fully reversible**: Harm resolves completely without intervention.
- **Reversible with intervention**: Harm can be undone but requires active remediation.
- **Partially reversible**: Some lasting effects remain despite remediation.
- **Irreversible**: Harm cannot be undone. Death, permanent disability, species extinction.

Dimension 3: Scale of Affected Population

How many people are harmed?

- **Individual**: One person or household.
- **Group**: Dozens to hundreds.
- **Community**: Thousands to tens of thousands.
- **Population**: Hundreds of thousands to millions.
- **Civilizational**: Billions or existential.

Dimension 4: Vulnerability of Affected Population

Are the people harmed already in a disadvantaged position?

- **General population**: No specific vulnerability.
- **Sensitive population**: Some members have heightened risk (e.g., elderly, pregnant).
- **Vulnerable population**: Structural disadvantage (e.g., children, disabled, economically disadvantaged).
- **Highly vulnerable**: Multiple intersecting vulnerabilities with limited agency or recourse.

Dimension 5: Cascade Potential

Can this harm trigger harm in connected systems?

- **Contained**: Harm stays within the immediate context.
- **Spreading**: Harm affects adjacent systems or populations.
- **Cascading**: Harm triggers chain reactions across multiple systems.
- **Systemic**: Harm threatens the stability of the broader system or society.

Dimension 6: Controllability

Can the affected person avoid or mitigate the harm themselves?

- **Fully controllable**: Affected person can easily avoid or mitigate harm.
- **Partially controllable**: Affected person can reduce but not eliminate harm.
- **Difficult to control**: Specialized knowledge or resources needed to avoid harm.
- **Uncontrollable**: Affected person has no ability to avoid or mitigate harm.

Dimension 7: Detection Likelihood

Can the harm be identified before it reaches the affected person?

- **Highly detectable**: Standard monitoring will catch this.
- **Detectable with effort**: Targeted testing or monitoring required.
- **Difficult to detect**: Specialized analysis required; easy to miss.
- **Undetectable until harm occurs**: No known method to identify before impact.

Part 3: The SEP Tier System

For any project, policy, or system undergoing Harm Blindness Framework analysis, the facilitator evaluates each of the seven dimensions during checkpoint review. The highest-severity dimension determines the minimum SEP tier, with additional escalation based on combinations.

Tier 1: Standard Review

Trigger: All seven dimensions score at their lowest levels. Harms are negligible to minor, fully reversible, affecting general population at individual scale, fully controllable, highly detectable, and contained.

Examples: Redesigning a consumer app's color scheme. Updating internal documentation templates. Changing a meeting scheduling policy.

Requirements:

- Standard checkpoint completion.
- Document identified stakeholders and potential harms.
- Assign mitigation owner.
- Proceed with project owner sign-off.

Analogous to: SIL 1, QM (automotive), DAL E (aviation), FDA Class I.

Tier 2: Enhanced Review

Trigger: Any dimension reaches moderate level. Harm is meaningful but manageable, reversible with intervention, affecting groups, involving sensitive populations, or partially controllable.

Examples: Launching a SaaS product collecting user data. Implementing a hiring algorithm. Redesigning a school curriculum.

Requirements:

- Full checkpoint completion with expanded stakeholder identification.
- External stakeholder input required (at least one affected person not on the project team).
- Written mitigation plan with measurable outcomes.
- Review by someone with authority to halt the project.
- Proceed with project owner and reviewer sign-off.

Analogous to: SIL 2, ASIL A-B, DAL C-D, FDA Class II, NIST Moderate.

Tier 3: Elevated Review

Trigger: Any dimension reaches serious or severe. Harm could be permanent, affect communities or populations, involve vulnerable groups, cascade across systems, or be difficult to detect or control.

Examples: Deploying AI in criminal justice. Releasing social media for minors. Implementing predictive policing. Deploying autonomous systems in public spaces.

Requirements:

- Full checkpoint completion with comprehensive stakeholder mapping.
- Mandatory external review panel including affected vulnerable populations.
- Independent risk assessment by a party with no financial interest.
- Documented kill switch and exit strategy.
- Written precedent analysis.
- Executive-level sign-off with documented personal accountability.
- Post-launch monitoring plan with defined escalation triggers.
- Cannot proceed without completing all Tier 3 requirements at every remaining checkpoint.

Analogous to: SIL 3, ASIL C-D, DAL A-B, FDA Class III, NIST High, INES 4-5.

Tier 4: Critical Review

Trigger: Fatal/catastrophic consequence possible in any scenario, or irreversible harm at population scale, or uncontrollable and undetectable harm to vulnerable populations, or systemic cascade potential affecting critical infrastructure or societal stability.

Examples: AI companion apps with no crisis escalation. Autonomous weapons systems. AI making unsupervised life-or-death medical decisions. Infrastructure control systems. Frontier AI with emergent unpredictable behavior at scale.

Requirements:

- Full stop at every checkpoint until all Tier 4 requirements are satisfied.
- Exhaustive stakeholder mapping including second-order and third-order effects.
- Independent safety audit by qualified external experts.
- Mandatory red team analysis for failure modes that cause death or irreversible harm.
- Documented crisis response plan including automated escalation.
- Kill switch with defined activation criteria and tested execution.
- Legal review of liability and accountability chain.
- Board-level sign-off with personal liability acknowledgment.
- Regulatory compliance verification for all applicable jurisdictions.
- Post-launch monitoring with real-time alerting and defined shutdown triggers.
- Scheduled review intervals (no less than quarterly) with authority to halt.
- Cannot launch without unanimous sign-off from all required reviewers.
- Any single reviewer can halt the project at any time.

Analogous to: SIL 4, ASIL D, DAL A, INES 6-7, Pharmaceutical Grade 4-5.

Part 4: Domain Translation Tables

The following tables allow practitioners in specific domains to map their existing classification systems to SEP tiers. This enables organizations already using domain-specific standards to integrate the Harm Blindness Framework without replacing their existing compliance infrastructure.

Safety Engineering (IEC 61508)

SIL Level	SEP Tier	Notes
SIL 1	Tier 1-2	Depending on population vulnerability
SIL 2	Tier 2-3	Depending on reversibility and cascade
SIL 3	Tier 3	Single death possible
SIL 4	Tier 4	Multiple deaths possible

Automotive (ISO 26262)

ASIL Level	SEP Tier	Notes
QM	Tier 1	Standard quality management
ASIL A	Tier 2	Light to moderate injury
ASIL B	Tier 2-3	Severe injury possible
ASIL C	Tier 3	Life-threatening possible
ASIL D	Tier 4	Fatal, multiple occupants

Aviation (DO-178C)

DAL Level	SEP Tier	Notes
DAL E	Tier 1	No safety effect
DAL D	Tier 1-2	Minor effect
DAL C	Tier 2	Major effect
DAL B	Tier 3	Hazardous
DAL A	Tier 4	Catastrophic

Medical Devices (FDA)

FDA Class	SEP Tier	Notes
Class I	Tier 1-2	Low risk, general controls
Class II	Tier 2-3	Moderate risk, special controls
Class III	Tier 3-4	High risk, depends on failure mode

Cybersecurity (NIST)

NIST Impact	SEP Tier	Notes
Low	Tier 1	Limited adverse effect
Moderate	Tier 2	Serious adverse effect
High	Tier 3-4	Severe/catastrophic, depends on cascade

Nuclear (INES)

INES Level	SEP Tier	Notes
0-1	Tier 1	Deviation/anomaly
2-3	Tier 2-3	Incident, near-accident
4-5	Tier 3	Accident, limited off-site
6-7	Tier 4	Major accident

Environmental (EIA)

EIA Severity	SEP Tier	Notes
Negligible	Tier 1	No measurable change
Minor	Tier 1-2	Localized, reversible
Moderate	Tier 2	Requires mitigation
Major	Tier 3	Difficult to reverse
Critical	Tier 4	Irreversible, ecosystem collapse

Pharmaceutical (Adverse Events)

Grade	SEP Tier	Notes
1 (Mild)	Tier 1	No intervention needed
2 (Moderate)	Tier 2	Non-invasive intervention
3 (Severe)	Tier 3	Hospitalization
4 (Life-threatening)	Tier 3-4	Urgent intervention
5 (Death)	Tier 4	Fatal outcome

Financial Services

Risk Level	SEP Tier	Notes
Low	Tier 1	Minimal financial impact
Moderate	Tier 2	Significant financial impact
High	Tier 3	Major impact, potential contagion
Systemic (SIFI)	Tier 4	Threatens financial system stability

Data Protection (GDPR DPIA)

DPIA Risk	SEP Tier	Notes
Low	Tier 1	Unlikely to affect individuals
Medium	Tier 2	Limited, manageable impact
High	Tier 3	Significant impact, vulnerable groups
Very High	Tier 3-4	Physical harm, discrimination

Part 5: Escalation Rules

Automatic Escalation Triggers

Regardless of initial tier classification, the following conditions automatically escalate to the next tier:

1. Children or minors are among the affected population. (Minimum Tier 2; escalate +1 from initial classification.)
2. Affected population cannot opt out or has no alternative. (Escalate +1.)
3. System operates without human oversight in harm-relevant decisions. (Escalate +1.)
4. Previous similar system has caused documented harm. (Escalate +1. If documented deaths, minimum Tier 4.)
5. System handles data or decisions involving multiple vulnerability dimensions. (Escalate +1.)
6. Organization has previously been found negligent or liable for harm in a related domain. (Escalate +1.)

Maximum Tier Cap

Tier 4 is the maximum. Multiple escalation triggers do not create a Tier 5. Instead, they compound the requirements within Tier 4 (additional review panels, broader stakeholder representation, more frequent post-launch review intervals).

De-escalation

A project may be de-escalated by one tier if, and only if, all of the following conditions are met:

- A full checkpoint review at the higher tier has been completed.
- All identified harms have documented mitigations with verified effectiveness.
- An independent reviewer (not on the project team) concurs with de-escalation.
- De-escalation reasoning is documented for audit.
- De-escalation cannot reduce below Tier 2 if vulnerable populations are affected.
- De-escalation cannot reduce below Tier 3 if death is a possible outcome regardless of probability.

Supplementary Notes

Relationship to Existing Standards

The SEP does not replace domain-specific safety standards. Organizations subject to IEC 61508, ISO 26262, FDA regulation, or other compliance requirements must continue to meet those requirements independently. The SEP provides an additional layer of stakeholder harm analysis that complements (never substitutes for) regulatory compliance.

The translation tables allow organizations to use their existing classification as a starting point for SEP tier determination, then apply the Harm Blindness Framework's broader stakeholder analysis to identify harms that domain-specific standards may not capture. For example, an automotive ASIL assessment

may correctly classify a vehicle system at ASIL B from a crash safety perspective, but analysis might reveal that the same system disproportionately affects disabled drivers, triggering escalation to SEP Tier 3 based on the vulnerability dimension.

The Universality Principle

The most severe plausible harm in bathroom design (slip-and-fall death of an elderly person) triggers SEP Tier 4 through the fatal consequence dimension. However, a routine bathroom renovation for a healthy adult triggers Tier 1. The SEP correctly distinguishes these because it evaluates population vulnerability and consequence severity independently. The domain does not determine the tier; the specific project within that domain determines the tier.

This is what makes the SEP universal. It does not say "all AI projects are Tier 3" or "all construction projects are Tier 2." It says, "evaluate this specific project against these seven dimensions, and the dimensions tell you what tier it is."

Document Status: DRAFT v1.0

Next Steps: Integration into Harm Blindness Framework Checkpoint Templates (replacing Death Gate Protocol references and binary Critical Death Risk Screening)