

# The Full Set - From Linear Risk to Emergent Safety Approaches in System Safety Analysis

David Slater

## ABSTRACT

*Safety science has undergone a steady evolution from the analysis of mechanical failure to the modelling of emergent behaviour in complex socio-technical systems. Early quantitative methods such as Fault Tree Analysis (FTA) and Probabilistic Risk Assessment (PRA) established the foundations of analytical rigour, but their deterministic assumptions limited their capacity to explain human and organisational performance. The subsequent development of Task Analysis, Human Reliability Analysis (HRA), and the Human Factors Analysis and Classification System (HFACS) extended the scope to human variability but retained a linear, reductionist logic.*

*The systems-thinking movement, beginning with Reason's Swiss Cheese Model, Rasmussen's AcciMap, and Leveson's system-theoretic STAMP framework, introduced the ideas of feedback, hierarchy, and constraint. Hollnagel's Functional Resonance Analysis Method (FRAM) completed this conceptual progression by modelling how variable functional interactions produce emergent outcomes. Together, these methods trace the transition from failure analysis to resilience analysis—from explaining what went wrong to understanding why things usually go right.*

*In modern safety assessment, static or purely qualitative tools such as Bow-Ties, risk matrices, and LOPA are no longer sufficient. The integration of the quantitative precision of FTA and HRA, the systemic structure of STAMP, and the dynamic variability modelling of FRAM—augmented by metadata and AI reasoning—offers a unified, predictive framework. This convergence of control logic and functional resonance defines the next stage of system safety science.*

## Keywords

*System Safety; Fault Tree Analysis (FTA); STAMP; STPA; FRAM; Functional Resonance; Resilience Engineering; Human Reliability; Probabilistic Risk Assessment (PRA); Safety-II; Socio-Technical Systems; AI-Assisted Safety Modelling; Large Language Models (LLMs)*

## INTRODUCTION

There has been growing interest in combining established safety assessment methodologies to exploit their complementary strengths. The central motivation lies in uniting structured, control-oriented approaches such as **STPA** with adaptive, variability-focused approaches such as **FRAM**. STPA offers a disciplined framework for identifying unsafe control actions, safety constraints, and causal mechanisms within defined control hierarchies (Leveson, 2011). FRAM, by contrast, focuses on how variability propagates through interacting functions to produce emergent outcomes (Hollnagel, 2012). Together they provide a richer and more complete depiction of how hazards arise and how systems succeed or fail under changing conditions.

In complex domains such as hydrogen refuelling or AI-enabled systems, such hybridisation helps bridge traditional linear safety logic with resilience-engineering perspectives (Hollnagel et al., 2006). **Kaya et al. (2025)** have shown that researchers are increasingly exploring integrated analyses supported by AI tools such as ChatGPT and Gemini. The intention is not to replace analysts but to accelerate model development and provide structured first-pass analyses that experts can refine. This reflects a broader recognition that neither FRAM nor STPA alone captures all facets of complex socio-technical systems: STPA excels in modelling control and constraint reasoning, while FRAM reveals adaptive performance and resonance effects.

Current thinking emphasises that while large language models can produce useful material rapidly, they still lack true systems reasoning and internal consistency. Their analyses remain fragmented and linear unless guided and validated by expert oversight. Consequently, combining complementary methods—supported, but not replaced, by AI—is increasingly seen as the most promising route toward balanced, multi-perspective safety assessments that integrate causal control logic with emergent variability analysis.

### From Components to Probabilities

The formal study of engineering risks and safety began with **Fault Tree Analysis (FTA)** in the early 1960s, developed at Bell Laboratories and Boeing to model missile-system reliability (Bell Laboratories, 1962). FTA decomposed system failure into logical combinations of basic events, using Boolean algebra to compute the probability of a top-level failure. It was elegant, rigorous, and readily quantifiable, but fundamentally mechanistic: systems were treated as collections of independent parts whose states were either functional or failed.

The development and refinement of this framework led to **Probabilistic Risk Assessment (PRA)** in the 1970s and 1980s, notably in the *Reactor Safety Study* (U.S. Nuclear Regulatory Commission, 1975). PRA integrated event and fault trees, combining initiating-event frequencies with barrier reliabilities to estimate overall system risk. It provided a means to express safety numerically but remained constrained by its reductionism. Human performance entered only as a parameter; adaptation and learning lay outside its scope. The pursuit of quantitative precision yielded insight but at the expense of context.

### The Human as an Agent

In parallel with the growth of PRA, attention turned to the human contribution to safety. Early **Task Analysis** and **Human Reliability Analysis (HRA)** extended probabilistic logic to operator behaviour (Swain & Guttman, 1983; Kirwan, 1994), treating actions and decisions as measurable failure points. Techniques such as THERP and HEART translated complex cognitive processes into numerical modifiers compatible with fault trees. This created a bridge between human and mechanical domains but portrayed people chiefly as sources of error to be mitigated.

By the 1990s, this view was giving way to a more interpretive understanding. Research into expertise, situation awareness, and team coordination revealed that human variability was not merely a liability but also a resource. **Crew Resource Management (CRM)** and **Naturalistic Decision Making (NDM)** emphasised adaptation under uncertainty rather than procedural compliance. These insights culminated in the emergence of **Safety-II** and **Safety Differently**, which define safety as the ability to succeed under varying conditions (Hollnagel, 2014). Associated movements such as **Human and Organisational Performance (HOP)** and the promotion of **psychological safety** encouraged openness, reporting, and continuous learning. The human was re-centred—not as an unreliable component, but as a dynamic, sense-making participant whose variability sustains resilience (Woods et al., 2010).

## The Systems Turn

This widening perspective was mirrored by a shift in how analysts conceptualised the system itself. **James Reason's Swiss Cheese Model** reframed human error as the product of latent system deficiencies (Reason, 1990; 1997). Organisational defences were depicted as layers, each containing potential weaknesses whose momentary alignment could permit failure to propagate. The model introduced the idea that safety emerges from the collective performance of defences rather than from individual perfection.

**Jens Rasmussen** transformed Reason's static metaphor into a **dynamic socio-technical network**. His **AcciMap** framework represented government, regulators, management, operators, and environment as interacting strata linked by feedback loops (Rasmussen, 1997). Accidents arose through "migration toward the boundary," where local adaptations and pressures gradually erode safety margins. Rasmussen thus established the foundations of modern systems thinking in safety: adaptation, coupling, and feedback as generative mechanisms of risk.

**Nancy Leveson's System-Theoretic Accident Model and Processes (STAMP)** formalised these ideas through control theory (Leveson, 2004; 2011). In STAMP, each level of the socio-technical hierarchy is a controller that issues constraints and receives feedback from the controlled process. Failures occur when control actions are missing, inappropriate, or delayed. By recasting accidents as control problems rather than barrier failures, STAMP provided a coherent analytical structure linking organisational management to system dynamics. Its derivatives—**CAST** and **STPA**—translated the theory into practical methods for investigation and hazard analysis (Fleming, Leveson & Thomas, 2014).

**Erik Hollnagel's Functional Resonance Analysis Method (FRAM)** advanced this thinking by shifting the analytical unit from "component" or "controller" to **function** (Hollnagel, 2012). Each function is characterised by six aspects—inputs, outputs, preconditions, resources, controls, and time—whose performance inevitably varies. When such variabilities interact across functions, they can amplify or dampen one another, producing emergent outcomes that define system behaviour. FRAM generalises the logic of **HAZOP** beyond equipment to include human and organisational functions, making it uniquely suited to represent genuinely complex systems where cause and effect are non-linear and context-dependent.

## The Plateau of Simplification

By the early 2000s, fully fledged PRA had become costly and resource-intensive, leading analysts to adopt more accessible tools such as **Bow-Tie Analysis**, **risk matrices**, and **Layers of Protection Analysis (LOPA)**. These methods preserved fragments of PRA's logic while simplifying its application. Bow-Ties combined fault and event trees around a central "top event," mapping causes and barriers in a visually intuitive format (Hollnagel, 2004). Matrices

and LOPA converted probabilities into categorical likelihoods and standardised the evaluation of protective layers. Such techniques improved communication and decision-making but froze the dynamics of complex systems into static diagrams. They described potential pathways to failure without capturing the continual adaptation through which real systems evolve.

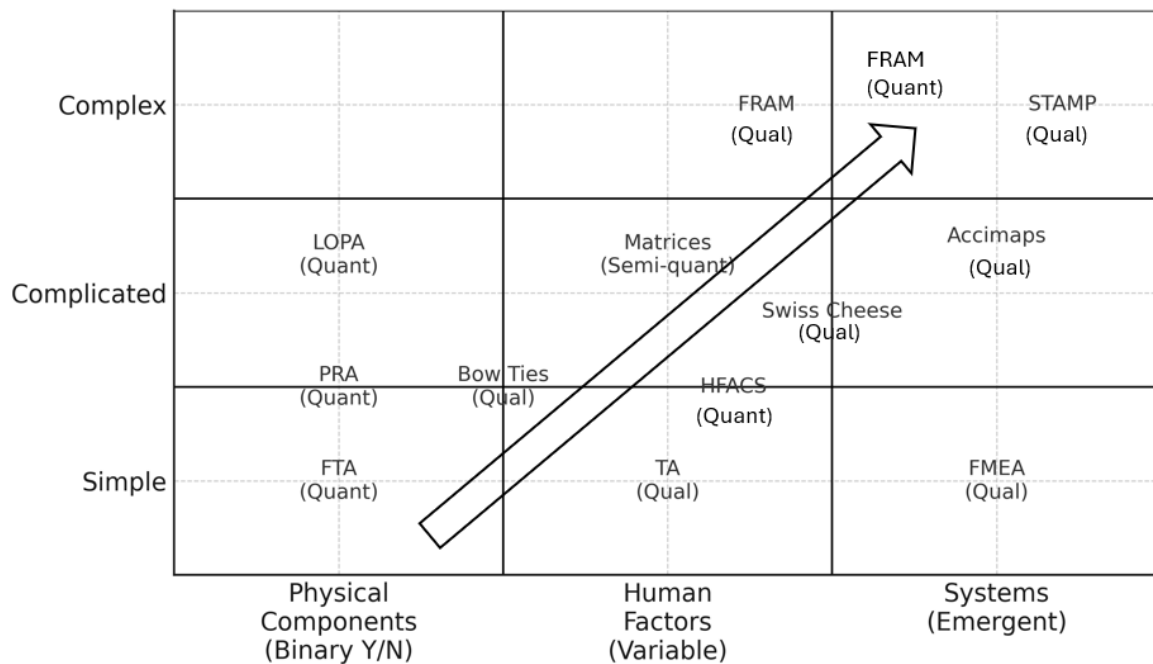
## Towards Integrated and Quantitative System Understanding

Today’s socio-technical systems—digitally connected, data-rich, and adaptive—demand methods that are both rigorous and dynamic. Analyses intended to support real-time learning, prediction, or digital-twin simulation require a foundation in quantitative modelling that traditional Bow-Tie or matrix approaches cannot provide alone.

An integrated framework must therefore draw on the complementary strengths of multiple traditions. The quantitative discipline of **FTA** and **HRA** provides the baseline performance expectations of components and human tasks (Kirwan, 1994). The **control-structural insight of STAMP** captures how oversight and feedback govern those performances across organisational levels (Leveson, 2011). The **dynamic coupling and variability modelling of FRAM**, especially when extended with **metadata** to represent probabilistic or parametric aspects of each function, enables the simulation of how small perturbations can resonate through the system to yield emergent behaviour (Hollnagel, 2012).

This synthesis allows analysts to represent safety as a continuum: quantitative reliability defines the boundary conditions; control theory maps the governance of those boundaries; and functional resonance models reveal the fluid interdependencies that determine actual outcomes. The result is an analytical architecture (Figure 1), capable of describing, measuring, and predicting the behaviour of complex socio-technical systems with both explanatory and numerical power.

**Domains of Safety Analysis Methods by Complexity and Analytical Focus**



**Figure 1. Domains of safety analysis methods by complexity and analytical focus.**

**Methods are positioned according to their epistemic domain—from simple, component-based quantitative analyses (FTA, PRA) to complex, emergent system approaches (STAMP, FRAM)—showing the historical and conceptual evolution of safety methodology.**

## CONCLUSION

Across six decades, system safety has progressed from the binary logic of mechanical reliability to the systemic modelling of emergent performance. Each methodological generation has addressed the shortcomings of the one before it—extending from component failures to human cognition, from human error to organisational control, and from control to functional interaction. The convergence now taking place—linking FTA, STAMP, and FRAM through quantitative metadata—marks the next stage in this evolution. It offers the possibility of unifying deterministic, probabilistic, and emergent paradigms into a single, coherent approach. In such a framework, safety is no longer the absence of failure but a measurable, adaptive property of systems that learn, coordinate, and evolve within their operating environments.

### Why This Matters

The convergence of FRAM and STPA—especially when mediated by LLMs—marks an effort to build AI-assisted, systems-theoretic safety frameworks capable of spanning both mechanistic and emergent domains. This hybridisation aims to improve early-stage hazard recognition, capture “unknown unknowns,” and ensure AI tools operate safely within complex, adaptive systems (Kaya et al., 2025).

## REFERENCES

- Bell Laboratories. (1962).** *Fault Tree Handbook for Systems Safety Engineering*. Bell Systems Technical Memorandum, Bell Laboratories, New Jersey.
- Fleming, C. H., Leveson, N. G., & Thomas, J. (2014).** *Safety assurance in complex systems: The system-theoretic process analysis (STPA) method*. MIT Engineering Systems Division Working Paper Series.
- Hollnagel, E. (2004).** *Barriers and Accident Prevention*. Aldershot: Ashgate.
- Hollnagel, E. (2012).** *FRAM: The Functional Resonance Analysis Method – Modelling Complex Socio-technical Systems*. Aldershot: Ashgate.
- Hollnagel, E. (2014).** *Safety-I and Safety-II: The Past and Future of Safety Management*. Farnham: Ashgate.
- Hollnagel, E., Woods, D. D., & Leveson, N. G. (Eds.). (2006).** *Resilience Engineering: Concepts and Precepts*. Aldershot: Ashgate.
- Kaya, K., Goktas, M., Yildiz, A., & Basar, C. (2025).** *Large Language Models Powered System Safety Assessment: Applying STPA and FRAM*. *Proceedings of the 2025 International Conference on Safety and Systems Engineering*, IEEE (in press).
- Kirwan, B. (1994).** *A Guide to Practical Human Reliability Assessment*. London: Taylor & Francis.
- Leveson, N. G. (2004).** *A new accident model for engineering safer systems*. *Safety Science*, 42(4), 237–270.
- Leveson, N. G. (2011).** *Engineering a Safer World: Systems Thinking Applied to Safety*. Cambridge, MA: MIT Press.
- Rasmussen, J. (1997).** *Risk management in a dynamic society: A modelling problem*. *Safety Science*, 27(2–3), 183–213.
- Reason, J. (1990).** *Human Error*. Cambridge: Cambridge University Press.

**Reason, J. (1997).** *Managing the Risks of Organisational Accidents*. Aldershot: Ashgate.

Shappell, S. A., & Wiegmann, D. A. (2000). *The Human Factors Analysis and Classification System (HFACS)*. *Aviation, Space, and Environmental Medicine*, 71(10), 997–1006.

**Swain, A. D., & Guttman, H. E. (1983).** *Handbook of Human Reliability Analysis with Emphasis on Nuclear Power Plant Applications (NUREG/CR-1278)*. U.S. Nuclear Regulatory Commission, Washington, D.C.

**U.S. Nuclear Regulatory Commission. (1975).** *The Reactor Safety Study: An Assessment of Accident Risks in U.S. Commercial Nuclear Power Plants (WASH-1400)*. Washington, D.C.: U.S. Government Printing Office.

**Woods, D. D., Dekker, S., Cook, R., Johannesen, L., & Sarter, N. (2010).** *Behind Human Error* (2nd ed.). Aldershot: Ashgate.