Human Factors in Cybersecurity, Vol. 208, 2026, 1–10
https://doi.org/10.54941/ahfe1007407

A Human Factors-Cyber-Safety Framework for Risk and Requirements in Critical Infrastructure

Eylem Thron[1], Duncan Ki-Aries[2], Martin Freer[1], Huseyin Dogan[2], and Shamal Faily[3]

[1]Mima, London, UK
[2]Bournemouth University, Bournemouth, UK
[3]Defence Science Technology Lab, Wiltshire, UK

ABSTRACT

Cyber-attacks on critical infrastructure are increasing in scale and sophistication, yet cybersecurity practice remains dominated by technology-centric assessments that insufficiently represent human contributions to risk. In cyber-physical systems (CPS), non-malicious human actions -including slips, mistakes, workarounds, training gaps, and misaligned procedures- frequently create, amplify or fail to detect vulnerabilities. This paper presents an integrated socio-technical framework that combines Human Factors (HF) methods, safety analysis, and cybersecurity modelling within a Secure-by-Design approach.The framework models how human performance variability influences cyber vulnerability and safety outcomes, enabling structured, scenario-based risk assessment and the derivation of traceable engineering requirements. An illustrative application demonstrates how HF findings are translated into human error mechanisms, cyber effects, unsafe control actions, safety impacts, and prioritised Secure-by-Design controls. By operationalising HF methods as cybersecurity engineering tools, the approach reframes cybersecurity as a socio-technical reliability problem comparable to safety engineering.

Keywords: Human factors, Cybersecurity, Safety-critical systems, Critical infrastructure, Risk modelling, Secure-by-design

INTRODUCTION

Cybersecurity risk in critical infrastructure is increasingly shaped by how humans interact with complex cyber-physical systems. High levels of automation, tightly coupled safety-critical functions, and operational pressures mean that human actions often accelerate the transition from a cyber disturbance to a safety consequence. Slips, lapses, rule-based mistakes, workarounds, or misinterpretation of procedures - particularly under high workload, fatigue, or degraded system states - can create, expose, or amplify cyber vulnerabilities (Pollini, 2021; Khadka, & Ullah, 2025).

Human Factors (HF) research provides established methods for analysing task demands, cognitive workload, interface design and organisational pressures. However, structured pathways for translating HF outputs into cybersecurity engineering artefacts remain limited (Thron et al., 2024).

Received February 19, 2026; Revised April 5, 2026; Accepted April 24, 2026; Available online July 20, 2026

Cyber risk assessment typically assumes stable human performance, while safety analysis often treats human actions as simplified elements within control structures. As a result, socio-technical interactions that drive cyber risk escalation are under-represented in Secure-by-Design practice.

CONTRIBUTIONS

This work addresses this gap by providing:

- A modelling workflow that translates HF performance variability into cybersecurity and safety artefacts.
- A structured mapping between Human Reliability Analysis (HRA) outputs and System-Theoretic Process Analysis (STPA) unsafe control actions.
- Scenario-based modelling of performance degradation under cyber-degraded operational conditions.
- Tool-supported implementation in the open-source platform Computer Aided Integration of Requirements and Information Security (CAIRIS) enabling traceable Secure-by-Design requirements (Faily, 2018).

The framework treats human performance as a dynamic system component rather than a background factor, enabling cybersecurity to be analysed as a socio-technical reliability problem.

BACKGROUND

Human performance in CPS emerges from interactions between operators, technology, tasks, and organisational context. Errors are system phenomena rather than solely individual failings.

HF distinguishes between:

- Slips and lapses (execution failures)
- Rule-based mistakes (misapplication of procedures)
- Knowledge-based mistakes (errors in novel or complex situations)

Performance Shaping Factors (PSFs) such as workload, time pressure, fatigue, situation awareness, usability, training, and organisational constraints influence performance variability. Under adverse PSFs, variability can become performance degradation, increasing the likelihood of error mechanisms.

During cybersecurity incidents, these PSFs can rapidly degrade human performance, increasing the likelihood that human actions introduce vulnerabilities, delay detection or weaken safety barriers.

Disciplinary Modelling Traditions and Gaps

Cyber-physical systems are typically analysed using methods from distinct traditions: HF, safety, and cybersecurity.

- Human Factors (HF) methods focus on understanding and shaping human performance. Examples include task analysis (Kirwan, 1994), cognitive work analysis (Vicente, 1999), and human reliability analysis

(HRA) (Hollnagel, 1998), which quantify the likelihood of slips, lapses, rule-based mistakes, or knowledge-based errors under varying conditions. While these methods provide rich insights into operator behaviour, they rarely link directly to system-level hazards or cyber threats.

- Safety methods aim to identify hazards and unsafe system behaviours. STPA and its security extension STPA-Sec analyse unsafe control actions and the causal pathways linking system behaviour to hazards or security breaches (Leveson, 2011), while Bow-Tie analysis provides a structured visual overview of threat pathways and barrier effectiveness (de Ruijter, & Guldenmund, 2016). Although these methods effectively map risks and controls, human contributions are often simplified or assumed to be stable (Hulme et al., 2022; Hollnagel, 2017).
- Cybersecurity methods (e.g., attack trees, STRIDE (Howard & LeBlanc, 2003), PASTA (UcedaVelez & Morana, 2015), ISO/IEC 27005) focus primarily on technical vulnerabilities and assume relatively stable human behaviour rather than the socio-technical interactions that may amplify cybersecurity risk.

Because these approaches evolved from different disciplinary priorities, they are rarely integrated. Human error is not systematically represented in cybersecurity risk models, and HF findings rarely translate directly into engineering requirements. This separation constrains Secure-by-Design effectiveness in safety-critical systems.

Table 1 lists some of the analytical approaches in different disciplines together with their limitations when used alone and how this work integrates them to overcome their limitations.

Table 1: Different disciplines model parts of the socio-technical risk space.

Discipline	Focus	Limitation When Used Alone	Contribution of This Work
HF / HRA	Human performance variability	Rarely linked to system hazards or cyber threats	Connects HRA outputs to cyber vulnerabilities
Safety (STPA, Bow-Tie)	Hazards, control structures, barriers	Human performance often simplified	Integrates PSF-driven human error
Cybersecurity (STRIDE, ISO 27005)	Technical vulnerabilities and threats	Assumes normal human behaviour	Models degraded human performance states

PROPOSED HF-SECURITY-SAFETY MODELLING FRAMEWORK

To address this gap, we propose a structured workflow that links HF modelling, cybersecurity analysis, and safety reasoning into a unified requirement derivation process. The workflow is applied through a structured sequence linking task analysis, PSF characterisation, HRA-based error identification, STPA mapping to unsafe control actions, and requirement derivation within CAIRIS.

Task and Cognitive Modelling

Hierarchical Task Analysis (HTA) is used to define operator tasks, decompose operational goals, and identify task steps where cyber-relevant errors may occur (Stanton, 2006). Cognitive Task Analysis (CTA) captures operators' mental processes, situation awareness demands, and decision-making under uncertainty, revealing the cognitive actions and workload associated with complex tasks (Knisely, Joyner & Vaughn-Cooke, 2021). Performance shaping factors (PSFs) are explicitly represented to model workload, fatigue, time pressure, automation reliance, usability constraints, training adequacy, and organisational pressures that shape performance variability and risk in human-system interactions (Knisely, Joyner & Vaughn-Cooke, 2021; Stanton, 2006).

Human Error Characterisation

Human Reliability Analysis (HRA) is used qualitatively to characterise how changes in PSFs influence the relative likelihood of slips, rule-based mistakes, or knowledge-based errors. Rather than estimating numerical probabilities, the framework uses structured comparison across operational states (e.g., normal, cyber-degraded, prolonged degraded) to represent escalation in human error likelihood. This approach avoids reliance on precise probability values while preserving analytical rigour and traceability.

Linking to STPA and Cybersecurity

Human error mechanisms are mapped onto unsafe control actions using System-Theoretic Process Analysis (STPA), enabling analysis of how performance variability propagates through control structures to influence cyber system states and safety hazards (Leveson, 2011). STPA systematically models controllers, control actions, feedback loops, and causal scenarios so that inadequate or incorrect control actions - including those arising from human cognitive or procedural errors - can be linked to hazardous system states and potential losses, thus making risk pathways explicit for safety and cybersecurity integration.

The key distinction from standalone STPA-Sec is that unsafe control actions are not treated as static failure points; instead, they are explicitly linked to dynamically changing performance shaping factors. Thus, the framework extends STPA-Sec by:

1. Modelling how degraded human performance states emerge.
2. Explicitly linking PSFs to error mechanisms.
3. Using operational scenarios to prioritise derived requirements.

Unsafe control actions are therefore analysed not only structurally but also behaviourally. In this approach, HRA-identified error mechanisms are treated as causal contributors to unsafe control actions and are explicitly represented within CAIRIS to maintain traceability from human performance conditions to derived security requirements.

Tool-Supported Implementation (CAIRIS)

The workflow is implemented in CAIRIS, which supports representation of:

- Tasks and cognitive demands
- Performance shaping factors
- Vulnerabilities and threats
- Hazards and unsafe control actions
- Traceable Secure-by-Design requirements

This implementation allows HF-informed requirements to be embedded early in system design, rather than treated as retrospective mitigations. From these relationships, Secure-by-Design requirements are derived, including interface improvements, procedural controls, organisational safeguards, and training or decision-support measures. The framework does not depend on visual CAIRIS artefacts for validity. However, CAIRIS enables structured traceability across these elements, improving repeatability and auditability of the modelling process.

Figure 1 illustrates the HF–Security–Safety Requirement Flow proposed in this work. Human Factors modelling outputs are translated into human error mechanisms, which shape unsafe control actions, influence cyber system states, and lead to safety impacts from which explicit design requirements are generated.

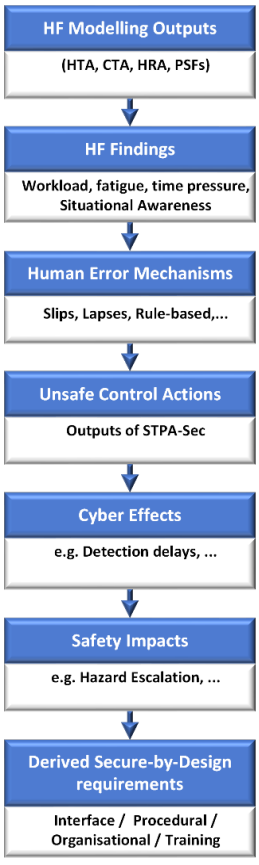

Figure 1: HF–security–safety requirement flow.

Table 2 illustrates how cyber-degraded conditions shift human performance states, altering key Performance Shaping Factors such as workload, time pressure, fatigue, and situational awareness.

Table 2: Human performance state shift under cyber-degraded operations.

Normal Operations	Human Performance State Shift	Cyber-Degraded Operations
Moderate	Workload	High
Low	Time pressure	High
Controlled (Low)	Fatigue	High (Extended shifts/ops)
High	Situational awareness	Reduced (Low)
Routine (Normal)	Interface use	Degraded (Emergency Mode)
Predictable	Automation behaviour	Degraded (Unexpected)
Low	Human error probability	High
Rare (Low)	Unsafe control actions	More Frequent (Moderate-high)
Limited (Normal)	Cyber vulnerability exposure	Expanded (High risk)
Maintained (Normal)	Safety barrier integrity	Degraded (High risk)

These changes increase the likelihood of human error and unsafe control actions, exposing systems to cyber vulnerabilities and weakening safety barriers. This pathway from cyber disturbance to safety consequence underpins the Human Error-Security-Safety Requirement Flow in Table 3, showing how HF outputs are translated into error mechanisms, cyber effects, safety impacts and traceable Secure-by-Design requirements.

Table 3: HF error-security-safety requirement flow.

HF Finding (From Modelling)	Error Mechanism	Cyber Effect	Safety Impact	Derived Requirement Type	Example Requirement
High workload + alarm overload increases misinterpretation risk	Knowledge-based mistake in alarm prioritisation	Delayed detection of intrusion	Prolonged degraded system state	Interface Design	The system shall prioritise and visually differentiate security-critical alerts from operational alarms.
Fatigue increases probability of incorrect configuration actions	Rule-based mistake	Misconfiguration of access control	Safety barrier weakened	Procedural Control	Configuration changes shall require confirmation prompts and automated validation checks during degraded operations.

(Continued)

Table 3: Continued.

HF Finding (from Modelling)	Error Mechanism	Cyber Effect	Safety Impact	Derived Requirement Type	Example Requirement
Time pressure leads to skipped escalation steps	Omission error	Incident response delay	Failure to mitigate escalating hazard	Process / Organisational inc. Training	Incident response procedures shall include automated escalation triggers when alerts remain unresolved beyond threshold time.

Figure 2 shows the workflow and underlying mechanics of the system and the various components / elements (i.e., in Table 3) demonstrating the operational integration of HF, safety and security analysis to generate requirement / outcomes.

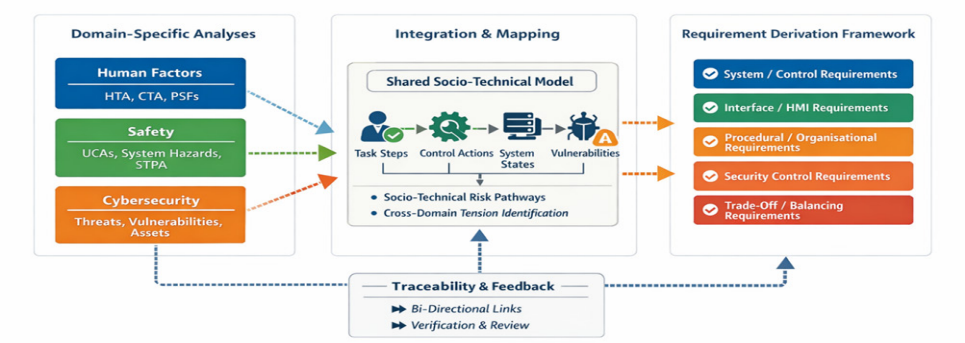

Figure 2: Integrated secure-by-design workflow.

Together, the tables and the figure demonstrate how transient human performance degradation is translated into structured risk artefacts and actionable design controls. This mechanism is illustrated with a scenario, as below.

ILLUSTRATIVE APPLICATION OF THE HF-SECURITY-SAFETY FRAMEWORK

Scenario Context

An operator supervising a safety-critical CPS receives multiple simultaneous alarms during a suspected cyber intrusion. Operational and security alerts share a common display format. The operator is working under elevated workload and time pressure and must determine whether the alert pattern

represents system fault or malicious activity. Table 4 shows how operational states alter PSFs in this scenario.

Table 4: Operational states vs PSFs.

Operational State	Workload	Alarm Density	Time Pressure	Cognitive Effect
Normal	Moderate	Manageable	Adequate	Stable interpretation
Cyber-degraded	High	Dense	High	Diagnostic strain
Prolonged degraded	High + fatigue	Sustained density	Persistent	Uncertainty, reduced verification

Propagation to Error and Risk

Under cyber-degraded conditions, elevated workload and alarm density increase cognitive demand. This creates a higher likelihood of a knowledge-based mistake in alarm interpretation. Within the control structure, this manifests as failure to initiate timely incident response an unsafe control action. The system remains in a degraded state, increasing exposure to safety hazards.

Figure 3 below illustrates how the scenario propagates across normal, cyber-degraded and prolonged degraded states, showing how performance shaping factors influence error likelihood and ultimately generate the derived Secure-by-Design requirement.

Figure 3: Escalation of performance shaping factors across normal, cyber-degraded, and prolonged degraded states and their propagation through human error mechanisms, unsafe control actions, cyber effects and safety exposure, generating the derived Secure-by-Design requirement.

Derived Secure-by-Design Requirement

The system shall prioritise, visually differentiate, and enforce acknowledgement of security-critical alerts using distinct colour coding and escalation logic to prevent misinterpretation under high workload and alarm-dense conditions.

DISCUSSION

The framework explicitly maps the dynamic behavioural pathways connecting cyber disturbances to safety consequences. Unlike standalone STPA-Sec analyses, which focus mainly on control structure adequacy, this approach captures how performance shaping factors (PSFs) - such as workload, fatigue, and time pressure - degrade over time and influence unsafe control actions. By linking PSFs to human error mechanisms, cyber effects, and operational hazards, the framework provides a structured basis for deriving and prioritising Secure-by-Design controls under degraded conditions, rather than assuming stable human performance.

The implementation within CAIRIS strengthens traceability by representing tasks, PSFs, vulnerabilities, hazards, and requirements within a unified socio-technical model. This improves transparency, repeatability, and scalability of the analysis. The use of CAIRIS to operationalise similar socio-technical modelling approaches has been demonstrated in prior work (e.g., Altaf et al., 2024), and its application here provides structured support for integrating HF, cybersecurity, and safety artefacts within a single traceable environment.

Overall, the approach reframes cybersecurity as a socio-technical reliability problem and produces structured engineering outputs rather than advisory observations.

This work provides a proof-of-concept demonstration through an illustrative scenario. Future research will focus on empirical validation in operational environments to assess scalability, usability and the impact of the framework on the quality and effectiveness of derived requirements.

CONCLUSION

Human actions are a central, and often overlooked, source of cyber risk in critical infrastructure. The proposed HF-Security-Safety framework models how human error propagates into cyber vulnerabilities and safety consequences, generates traceable Secure-by-Design requirements, and supports scenario-based resilience planning.

The derived requirements are Secure-by-Design because they are based on early modelling of human–system interaction rather than post-incident fixes. The framework enables scenario-based assessment by modelling how cyber incidents change workload, fatigue, and situational awareness, and therefore human error likelihood. This supports proactive resilience planning by anticipating socio-technical risk conditions rather than reacting to known failures. While demonstrated through an illustrative scenario, the framework establishes a foundation for future empirical validation in real-world safety-critical systems.

REFERENCES

Altaf, A., Faily, S., Dogan, H., Thron, E. and Mylonas, A. (2021) 'Integrated design framework for facilitating systems-theoretic process analysis', in *European Symposium on Research in Computer Security*, Cham: Springer International Publishing, pp. 58–73.

de Ruijter, A. and Guldenmund, F., 2016. The bowtie method: A review. *Safety science*, *88*, pp. 211–218.

Faily, S. (2018) *Designing usable and secure software with IRIS and CAIRIS*. New York, NY: Springer International Publishing.

Hollnagel, E. (1998) *Cognitive reliability and error analysis method (CREAM)*. Elsevier.

Hollnagel, E. (2017) *Barriers and safety management*. CRC Press.

Howard, M. and LeBlanc, D. (2003) *Writing secure code*. Microsoft Press.

Hulme, S., et al. (2022) 'Reliability of STPA in complex systems hazard analysis', *Safety Science*, 150, 105699.

ISO/IEC 27005:2022 (2022) *Information security risk management*. ISO.

Khadka, K. and Ullah, A.B., 2025. Human factors in cybersecurity: an interdisciplinary review and framework proposal: K. Khadka, AB Ullah. *International Journal of Information Security*, *24*(3), p. 119.

Kirwan, B. (1994) *A guide to practical human reliability assessment*. Taylor & Francis.

Knisely, B.M., Joyner, J.S. and Vaughn-Cooke, M. (2021) 'Cognitive task analysis and workload classification', *MethodsX*, 8, p. 101235. Available at: https://doi.org/10.1016/j.mex.2021.101235 (Accessed: [11/01/2026]).

Leveson, N. (2011) *Engineering a safer world: Systems thinking applied to safety*. Cambridge, MA: MIT Press.

Pan, X. and Wu, Z. (2020) 'Performance shaping factors in the human error probability modification of human reliability analysis', *International Journal of Occupational Safety and Ergonomics*, 26(3), pp. 538–550. Available at: https://doi.org/10.1080/10803548.2018.1498655 (Accessed: [11/01/2026]).

Pollini, A., Callari, T.C., Tedeschi, A., Ruscio, D., Save, L., Chiarugi, F. and Guerri, D., 2022. Leveraging human factors in cybersecurity: an integrated methodological approach. Cognition, Technology & Work, 24(2), pp. 371–390.

Stanton, N.A. (2006) 'Hierarchical task analysis: developments, applications, and extensions', *Applied Ergonomics*, 37(1), pp. 55–79. Available at: https://doi.org/10.1016/j.apergo.2005.06.003 (Accessed: [18/01/2026]).

Thron, E., Faily, S., Dogan, H. and Freer, M. (2024) 'Human factors and cybersecurity risks on the railway – the critical role played by signalling operations', *Information & Computer Security*, 32(2), pp. 236–263.

UcedaVelez, T. and Morana, M.M., 2015. Risk Centric Threat Modeling: process for attack simulation and threat analysis. John Wiley & Sons.

Vicente, K.J. (1999) *Cognitive work analysis: Toward safe, productive, and healthy computer-based work*. CRC Press.