Document id
D-2024-367959

Final Report

Safety Assessment 2030

(Säkerhetsbedömning - Ny Teknik 2030)

REVISION HISTORY

<Redovisa vad som skiljer mellan de olika utgåvorna genom att övergripande ange vad som har ändrats, syfte och orsak till ändringen av rapporten.>

VERSION	DATE	PREPARED BY	COMMENTS
0.1	2023-08-22	Lothar Meyer	Initial Version
0.5	2024-02-29	Lothar Meyer	First complete draft
0.6	2024-03-06	Lothar Meyer	First iteration comments
0.7	2024-03-07	Lothar Meyer	Second iteration commencts

AUTHORS OF THIS DOCUMENT

Lothar Meyer, LFV

Christian Bjursten Carlsson, LFV

Maximilian Peukert, LFV

Lars Danielson, SAAB

Eivind Martinsen, LFV

TABLE OF CONTENTS

SUMMARY (ENGLISH)

Flight deck operations enjoy the benefits of automation by increased safety margins and cost-efficiency. Air traffic management (ATM) and related air traffic control (ATC) has so far not followed this emerging trend to the same extent, which is increasingly expected to change with the deployment of AI-supported decision-making tools. High functional automation comes at the cost of phenomena in which the operator may expect something different from what automation may provide, known as “automation surprise” (AS). It was assumed so far that AS is specifically symptomatic for higher levels of automation, such as decision selection and action implementation, as it might be found on flight deck. But even nowadays, new techniques are making the work environment of air traffic controllers (ATCO) more complex, involving interactions and dependencies among multiple actors and systems and interfaces that increasingly include incorporating automation. ATM thus follows in the footsteps of the flight deck. Current safety assessment methods and related models do not support this phenomenon which makes it impossible for safety assessors to identify automation-related risks prospectively. In the trade-off between risks and efficiency, added certainty in the safety assessment has a beneficial effect. The more certain the safety margins are, the greater the scope gained for reducing costs and increasing efficiency. Investing in the further development of safety assessment methods therefore clearly pays off.

The study's objective was to set the groundwork for a new safety assessment approach that addresses new risk of higher levels of automation. The primary focus is understanding "automation surprise" and to assess the risks associated with this phenomenon. The project systematically attempted to integrate the phenomenon into prevailing safety-I models and processes. The project began with a literature review, exploring different perspectives of explanations and established assessment methods. In conjunction with the literature review, evidence supporting the existence of AS was gathered through interviews, a survey study, and an incident database analysis. Based on the available data and analysis results, requirements could be specified for a method capable of identifying risks involving automation surprise. In the search for a solution, an alternative term, "Discrepancy of Expectation," was proposed. Upon shifting towards an "expectation-centered" approach, it became more straightforward to elaborate on why an operator's expectations might deviate from the actual automation behavior. A guideline was proposed that base on the theory of corrupted critical cues as a primary reason for expectation discrepancies, offering a step-by-step approach for safety assessors to identify automation, related functions, and associated risks arising from such discrepancies. The guideline has been tested in the scope of a series of workshops using three exemplart cases considered innovative and reasonable candidates concerning the arousal of Automation Surprise. The cases cover different levels of automation from 0 to 5, from information acquisition to partly autonomous action implementation: AI-supported decision-making in-route (Level 2 automation), Multi Remote Tower (Level 0 automation) and Digital Tower Assistance (RESKILL/DITA) (Level 5 automation). Paradoxically, the cases with the higher levels of automation showed more ease in the approach than the level 0 automation of multi-remote towers. In the latter, operational experts found it difficult to find the automation, how to apply the theory, and find something that can be considered an "expectation". The case's contradictory results illustrates the effects of implicit versus explicit knowledge. Recommendations include refining the guideline to assist assessors and operational experts in identifying expectations that may lead to discrepancies in automation behavior.

SAMMANFATTNING

Flight deck-operations nyttjar fördelarna med automation genom ökade säkerhetsmarginaler och kostnadseffektivitet. Air Traffic Management (ATM) och relaterad flygtrafikledning (Air Trafic Control, ATC) har hittills inte följt denna framväxande trend i samma utsträckning, vilket alltmer förväntas förändras med införandet av AI-stödda beslutsfattande assistenter. Hög funktionell automation kommer med kostnaden för fenomen där operatören kan förvänta sig något annat än vad automationen kan erbjuda, känt som "automation surprise" (AS). Det antogs hittills att AS är specifikt symptomatiskt för högre nivåer av automation, såsom beslutsval och handlingsimplementering, som det kan hittas på flight deck. Men även nu för tiden gör nya tekniker arbetsmiljön för flygledare (ATCO) mer komplex, med interaktioner och beroenden mellan flera aktörer och system och gränssnitt som alltmer inkluderar automation. ATM följer således i flight deck:ens fotspår. Nuvarande flygsäkerhetsbedömningsmetoder och relaterade modeller stödjer inte detta fenomen, vilket gör det omöjligt för säkerhetsbedömare att identifiera automationrelaterade risker på ett prospektiv sätt. I avvägningen mellan risker och effektivitet har ökad kvalitet och noggranhet en gynnsam effekt. Ju mer exakta flygsäkerhetsmarginalerna är, desto större rörelsefrihet får man för att minska kostnaderna och öka effektiviteten. Att investera i ytterligare utveckling av säkerhetsbedömningsmetoder lönar sig därför tydligt.

Studiens mål var att lägga grunden för ett nytt tillvägagångssätt för flygsäkerhetsbedömning som hanterar de nya riskerna med högre nivåer av automation. Huvudfokus är förståelsen för "automation surprise" och att bedöma de risker som är förknippade med detta fenomen. Projektet försökte systematiskt integrera fenomenet i befintliga Safety-I-modeller och processer. Projektet inleddes med en litteraturstudie, där olika perspektiv på förklaringar och etablerade bedömningsmetoder utredes. I samband med litteraturstudien samlades bevis som stöder existensen av AS genom intervjuer, en enkätstudie och en analys av LFV incident/felrapporterings-databas. Baserat på tillgänglig data och analysresultat kunde krav specificeras för en metod som kan identifiera risker relaterade till automationsöverraskning. I sökandet efter en lösning föreslogs termen "Discrepancy of expectation". Genom att övergå till ett "förväntningscentrerat" tillvägagångssätt blev det lättare att förklara varför en operatörs förväntningar kan avvika från det faktiska automationsbeteendet. En riktlinje föreslogs som bygger på teorin om korrumperade kritiska ledtråd (eng. ”critical cue”) som en primär orsak till förväntningsdiskrepanser och erbjuder en steg-för-steg-plan för flygsäkerhetsbedömare att identifiera automationsrelaterade funktioner och associerade risker som uppstår från sådana discrepanser. Riktlinjen har testats inom ramen för en serie safety workshops med tre exempelartade fall som betraktas som innovativa och rimliga kandidater med avseende på framkallande av automation surprise. Fallen täcker olika nivåer av automation från 0 till 5, från informationsinhämtning till delvis autonom handlingsimplementering: AI-stödd beslutsfattande in-route (nivå 2 automation), Multi Remote Tower (nivå 0 automation) och Digital Tower Assistance (RESKILL/DITA) (nivå 5 automation). Paradoxalt nog visade fallen med högre nivåer av automation mer lätthet i tillvägagångssättet än nivå 0 automation för multi-remote torn. I det sistnämnda fallet hade operativa experter svårt att hitta automatiseringen, hur man tillämpar teorin och hitta något som kan betraktas som en "förväntan". Fallens motsägelsefulla resultat illustrerar effekterna av implicit versus explicit kunskap. Rekommendationerna inkluderar att förädla riktlinjen för att hjälpa bedömare och operativa experter att identifiera förväntningar som kan leda till diskrepanser med automatiseringsbeteendet.

DELIVERABLES

NO	DELIVERABLE DESCRIPTION	REFERENCE IN REPORT
1	List of methods and evaluation criteria for safety assessment	Section 2.2, 2.3, 4.1
2	Planning workshops and interviews and carrying them out	Section 3.2 and 3.3
3	List of methods containing expert opinions and assessment of the method's suitability for safety assessment of New Technology	Section 3.3
4	Preliminary guidelines for safety assessment of New Technology	Section 4.2 and 4.3
5	Verification of guidelines using three automation cases	Section 5
4	Recommendations for future implementation and development of methodology for safety assessment of New Technology	Section 6

1 MOTIVATION AND THEORETICAL BACKGROUND

Flight deck operations enjoy the benefits of automation by increased safety margins and cost-efficiency. Air traffic management (ATM) and related air traffic control (ATC) has so far not followed this emerging trend to the same extent, which is increasingly expected to change with the deployment of AI-supported decision-making tools. The introduction of automation comes at the cost of a phenomenon in which the operator may expect something different from what automation may provide, known as “automation surprise” (AS). It is assumed so far that AS is specifically symptomatic for higher levels of automation as it might be found on flight deck, such as decision selection and action implementation.

But even nowadays, new techniques are making the work environment of air traffic controllers (ATCO) more complex, involving interactions and dependencies among multiple actors, systems and interfaces that increasingly include incorporating automation. ATM-related automation is a product of our advanced digital technology that supports the operator with features that will increase information aimed at improving situational awareness and attention management; the system will provide early detection of conflicts and advice on allocating resources. ATM thus follows in the footsteps of the flight deck.

Automation enables the optimization of various aspects and services of ATM to utilize all available capacity in the airspace and runways while maintaining safety margins. Automation solutions also influence conceptual designs, such as generic en route and approach control services, and influence concept design. An example of a need for higher automation is the implementation of multiple remote towers. New towers, combining new procedures with new technology, challenge established routines and alter the role-play between the dynamics of operator-system interaction. Another example is AI support in en route. The trend is moving towards operations, generic authorizations, and more dynamic use of airspace and operator service utilization. To be successful in this area, increased automation with the help of machine learning and artificial intelligence is required. However, the trend towards a higher degree of automation is accompanied by automation-specific risks. The identification of risks in a highly automated environment and the implementation of countermeasures require advanced safety assessment methods that support the process with appropriate safety models and the corresponding framework to equip the safety assessor with the right tools. A measure of maturity is required, and it is important to find out whether existing methods meet this challenge.

While implementing changes is inevitably associated with costs and associated risks, the underlying assumption is that the equilibrium between induced risks and enhanced efficiency and safety propels the change itself. Advanced and highly optimized solutions stemming from paradigm-shifting innovations necessitate a realistic and robust evaluation of the performance and capacity-safety trade-off in advance. This approach mandates precise methodologies capable of determining or even quantifying the trade-off, delineating reasonable levels of challenges concerning risks. These challenges extend not only to the operator but also to decision-makers and safety assessors. They must factor in risks and potential risk mitigations throughout the conceptualization, migration, and training phases of the new solution. Particular emphasis is placed on actions addressing the management of complexity inherent in paradigm-shifting solutions, considering the myriad interactions and potential failure modes that may arise. A simple relationship arises from this interaction: The more uncertain the safety margins are due to inaccurate methods, the more conservative the trade-off must be. However, the better the risks can be predicted and mitigated, the more certain and accurate the safety margins are. In the trade-off between risks and efficiency, this certainty has the following effect: the more certain the safety margins are, the greater the scope gained for reducing costs and increasing efficiency. Investing in the further development of safety assessment methods therefore clearly pays off.

Undoubtly, one of the most significant sources of uncertainty is considered the AS phenomen. It's basically an inevitable side-effect of the increasing complexity that makes the operator being surprised when his/her expectations deviate significantly from the de-facto behavior of automation. AS is a well-known phenomenon on the flight deck when pilots detect that the flight management system (FMS), autopilot (FCU), or other systems initiate actions that conflict with the pilot's original plan. A phenomenon that has been identified as safety-relevant as it contributes to accidents, such as Turkish Airlines Flight 6491 on 16 January 2017, in which the autopilot intercepted a false glideslope, which resulted in a steep glide angle and false touch-down zone at the end of the runway. Pilots were not able to interpret the indications given by the autopilot and failed to identify the false operational mode due to complexity issues.

Palmer 1995 gives a representative definition [1]:

> *"An Automation Surprise occurs when the automation behaves in a manner that is different from what the operator is expecting*".

As mentioned above, ATC was often considered unaffected by AS phenomenon. This was, as the level of automation is relatively low, and actions are not initiated by automation on its own. However, this is only partly true as the levels of automation in ATC still support information acquisition and analysis as illustrated in Figure 1, which is still considered level 0 automation according to the ATM Masterplan [2]. Even level 0 automation is automation that provides relevant elements in the decision-making of the air traffic controller. To illustrate the extent of the potential severity, incorrect assumptions about flight information, control, weather observation or emergency services can affect pilot's decisions who rely on the ATCO's situational awareness.

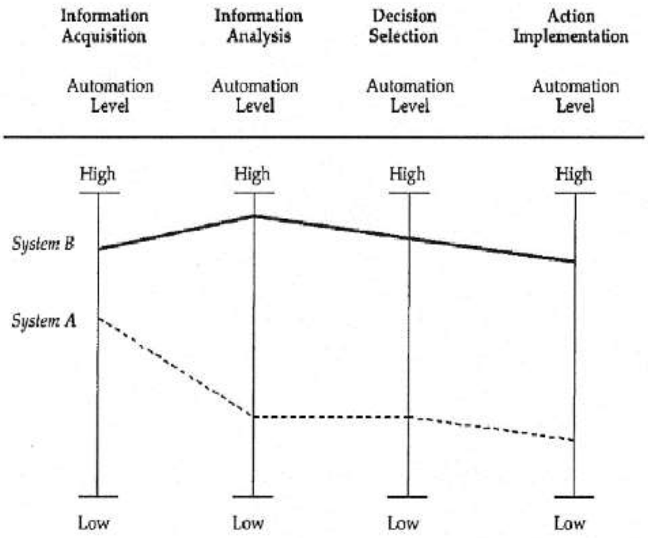

Figure 1 Levels of automation for independent functions of information acquisition, information analysis, decision selection, and action implementation. Examples of systems with different levels of automation across functional dimensions are also shown. Example by Parasuraman 2000 [3].

An illustrative example of an ATC-related automation surprise plays a crucial role in the disappearance of flight Malaysia Airlines MH370 on 8 March 2014 [4]. The secondary radar contact was lost after the transponder of MH370 was deliberately deactivated at navigation point IGARI at 1720, with a cleared level of 35,000 ft. This caused a loss of position signal in all secondary radar systems related to FIR Ho Chi Minh (HCM), Singapore, Bangkok, and Kuala Lumpur (KL). On the seek for MH370, Malaysian Airways (MAS) OPS Centre assumed that "their flight tracker (officially referred to as 'flight-following system') was able to exchange signals with the flight. Based on this assumption, KL ACC informed HCM ACC that MH370 was still flying and that the aircraft was continuing to send position reports to the airline. KL ACC relayed the latitude and longitude to HCM ACC as advised by Malaysian Airlines Operations. At 1930, the MAS Operations Centre informed KL ACC that the flight-following system was based on flight projection and unreliable for aircraft positioning.

Consequently, until 1930, over two hours after the deactivation of the secondary radar transponder, KL ACC and HCM ACC relied on the false assumption that MH370 continued its flight, assumed to be situated over Cambodia. Although the MAS flight tracker was never intended to be used as information for safety-relevant tasks, it successfully confused the assumed situation of ATCC in this emergency situation. To generalize this case: Despite the flight-follow system representing a relatively low level of automation (information analysis), the decision-makers and ATCO set too high expectations for the system capabilities' reliability and are ultimately surprised by the de facto behavior. This fits the definition of an automation surprise.

This tragedy demonstrates that the phenomenon of AS deserves efforts to set the groundwork, delivering essentials and fundaments of how to safety assess the phenomenon. Current research certainly gives hope. The progress made in the field of human-automation interaction is significant, which could drive the development of safety assessments that rely on state-of-the-art models and associated knowledge to reach the next level. Understanding the complex socio-technical interactions has evolved with models that show how to record, describe, and model behavior at interaction points that reflect the interaction between parallel processes, as defined by the Joint Cognitive Framework proposed by Lundberg 2015 [5]. Validation methods were improved with the ability to quantify operators' risk compensation effect, as shown in the project MERASSA [6]. New measurement technology shows the ability to track the impact of stress and fatigue. The research comes with a number of results, related models, and techniques that make it possible to adapt these for use in the safety assessment process of new technologies and thus improve the quality and explanatory power.

With this in mind, the primary objective is to develop a framework that incorporates the phenomenon of AS into a Safety I[1] context. This paves the way to develop a guideline that supports safety assessors and safety decision-makers in the prospective identification of AS-related risks. Against this background, this project is to be seen as groundwork, or a preliminary study, which takes the first steps to manage the risks associated with AS. To achieve these goals, this project follows a step-by-step approach:

1. A literature review investigates the state-of-the-art of automation surprise, including models and concepts of explanation. As well, the investigation includes available safety assessment methods that assesses concerning the integration and support of automation surprise. This shall clarify the development of requirements for future assessment methods that specifically address the increasing complexity of the interaction between the operator, other actors, automation, and procedures. Here, it is essential to understand the limitations and consequences of current methods that should lead to the necessary features of the next generation of methods.
2. A safety database incident investigation is performed, providing an overview of possible occurences of automation surprise in the ATCC Malmö. As well, an interview study and a survey study is conducted to further shape the knowledge available about automation surprise.
3. A guideline and classification scheme is conceptually developed to support safety assess the risks related to automation surprise at higher level of automation. This shall provide safety assessors and other stakeholders with an overview of current problem analysis and thus optimize decision-making.
4. The guidline is then applied to three cases. This shall demonstrate the concept's capability in comparison to a conventional approach. The three showcases represent examples of solutions under development that meet the criteria for paradigm-shifting solutions: Multi Remote Tower, AI-supported decision making in En-Route and Digital Air Traffic Assistance (DITA).

The final report is structured according to these points, which in principle follow the services promised in the project description, which are listed in the table in the "Deliverables" chapter.

Two publications were authored and submitted in the scope of the project, which are listed in section 8. The citation of these publications differ from all others by using APA citation style only (Name and year) without the additional citation using IEEE style (number in a square bracket). The literature review in section 2 of this document is largely a repetition of the ICAS conference publication Meyer et al. 2022. The conference article contains further details on the methods. As does the EAAP conference publication Fritz et al 2024 with the review study in section 3.2.

2 LITERATURE REVIEW AUTOMATION SURPRISE

2.1 Phenomenon

Firstly, literature is reviewed that supports the understanding and problem analysis of the AS phenomenon. By this, findings addressing our research questions above are provided. The literature review focuses on the theories and concepts available in the respective research area. AS research area

[1] *"Safety-I is defined as a state where as few things as possible go wrong. A Safety-I approach presumes that things go wrong because of identifiable failures or malfunctions of specific components: technology, procedures, the human workers and the organisations in which they are embedded. Humans—acting alone or collectively—are therefore viewed predominantly as a liability or hazard, principally because they are the most variable of these components."* [7]

of the review shall be synthesized qualitatively according to our questions. This approach is, hence, consistent with a meta-synthesis. The overview does not claim to be exhaustive. Still, it is preliminary and representative based on a selection from acknowledged databases and the corresponding number of citations, as well as on the judgment of the authors who have experience in the field of safety assessment, accident investigation, and human factors.

The first research question is intended to be answered based on literature reviews that satisfy the criterion of developing concepts for explaining AS. This may involve higher-level causal classes, terms, taxonomies, and models. Findings that suit providing an answer are finally derived from experienced cases of AS (inductive). For simplicity, the search focuses on cockpit crew experience and cockpit automation, as this research area is at the leading edge and provides the most experience. Qualitative studies of interest may also aim to develop concepts and discussions, which in turn may be based on literature reviews and logical reasoning.

Regarding the source of the cases, experience-based data from operators involved, such as surveys or interviews, are subjective but externally valid and provide details about the working context of the particular situation in which AS occurred. The literature review may include investigation reports from safety occurrences involving accidents and incidents. These cases rely on, e.g., flight data recorders, cockpit voice recorders, and pilots' testimonies. Factors contributing to AS that represent an intersection of the findings from the literature reviewed shall be identified. A classification system was applied to structure the factors and the discussion based upon it. The chosen classification system SOAM from Eurocontrol [8] is a well-known system usually used in accident and incident investigations and is well suited for our approach

The search of the term "automation surprise" in the title, keywords, and abstract resulted in 86 hits on the Scopus search platform. The additional narrowing with the keyword "safety" in all text led to 68 hits. Further filtering was applied according to the targeted profile in the method section, reducing hits to five. Additionally, six studies that matched the target criteria were found through citations and were therefore added to the selected list.

Sherry and Mauro 2014 [9] investigated 19 loss-of-control aviation accidents, focusing on understanding the sequence of events that coincided with inadequate intervention actions by the flight deck crew. They found that the decision-making logic was not adequately supported by the automation cues available in the cockpit, which are necessary for an appropriate response. The cues of concern are aircraft structure and airfoils, aircraft sensors, control surfaces, propulsion systems, and automation. Rare failure events that are related to the cues were not detected in the accidents due to the "hidden" nature of fail-safe sensor logic, "silent" and/or "masked" automation responses. Additionally, it was mentioned that cues are absent that support the pilots to anticipate speed envelope violations, in recognizing a speed envelope violation due to noise in the airspeed signal, in recognizing the airspeed envelope violation due to non-linearity and latency in the thrust response near the idle thrust setting.

Dehais & Peysakhovich 2015 [10] provoked the analyzed response of the flight deck crew to automation surprise in a flight simulation using eye-tracking. Indeed, "automation surprise" led to an excessive but inefficient visual search that prevented pilots from extracting the relevant information (i.e., the speed indicator). Whereas conflict solving was "straightforward" (i.e., reducing the selected speed with the dedicated FCU knob), most pilots were stuck and failed to deal with the situation immediately. Many participants made typical "fixation errors" as they persisted in disengaging and reengaging the autopilot (i.e., lateral/vertical guidance) or dialing in vain the altitude or vertical speed knobs on the FCU. Moreover, the analysis of ocular events revealed that the volunteers exhibited higher visual search (more short fixations and saccades) to the detriment of information processing (fewer fixations) during conflict compared to baseline. The eye movement analysis revealed that such conflicts impair attentional abilities, leading to an excessive visual search and inability to extract relevant information.

Rankin, Woltjer & Field identified, with the help of 20 pilots interviewed in 2019 [11] , 9 categories on the basis of 48 cases of automation surprise experienced in cockpits. The causes investigated were an absence of salient cues, causing confusion of switches. Pilots found it challenging to detect passive and insidious disturbances that build up slowly over time, making the autopilot suddenly disconnect. It was also found time-critical to deal with conflicting and inconsistent data from multiple failures.

Boer and Hurts investigated in 2017 [12] using a survey of flight deck crews, finding that AS occurs three times per pilot a year without severe consequences. They found AS events not to be the result of cognitive failures but rather the consequence of the current complexity of the cockpit system and interface design choices that possibly exceed the bounds of human comprehension. The effect of experience and operational intensity indicates that the initial training curriculum for pilots is insufficient to avoid AS events. Given that system complexity and interface design choices are a major factor in so many (non-consequential) AS events, this seems to override individual cognitive errors and differences in knowledge and training.

In 2017, Boer and Decker [13] examined the theories of AS obtained to date. They compared two models that explain the occurrence of AS, identifying a common pattern of occurrence. First, automated systems act independently without immediately preceding instructions, input, or commands from human(s). Second, there are gaps in users' mental models of automation. And third, feedback about automation activities and future behavior is weak.

Parasuraman and Mazey reviewed in 2010 [14] the different phenomenons of automation complacency and automation bias, prerequisites to the arousal of AS. Automation complacency arises primarily in the attention allocation strategy of keeping track of parallel tasks in a mixed environment of automated and manual work. Attention is preferably shifting to manual work at the expense of monitoring automation. Automation bias aims at omission and commission error when decision aid is wrong, evoked by the operator's attitude or assumption of relying on flawless automation.

Endsley notes in 2016 [15] that automation can negatively affect situational awareness. A major factor in automation-related errors is the operator's lack of awareness of the state of automation. The lack of awareness is indicated by a slow detection of problems with extra time needed to understand relevant system parameters and settings. The cause might be a loss of vigilance and increasing complacency, instead taking up a position of monitoring and passively receiving information rather than processing and anticipating in the scope of appropriate situation awareness.

Sarter & Woods explained in 1997 [16] that automation surprises are indications that a crew has misunderstood, miscommunicated with, misused, or mismanaged the automated systems. They distinguish two different types of AS:

- Automation does not execute actions that were expected and
- Automation changes inputs or executes differently than the operator told it to do.

Sarter, Woods, and Billings 1997 [17] argued that the gap between user-centered intentions and technology-centered development raises the likelihood of AS arousal. AS likely occurs when

- designers oversimplify the pressures and task demands from the users' perspective,
- assuming that people can and will call to mind all relevant knowledge,
- overconfidence that they have taken into account all meaningful circumstances and scenarios,
- assuming that machines never err,
- making assumptions about how technology impacts human performance without checking for empirical support or despite contrary evidence,
- defining design decisions in terms of what it takes to get the technology to work,
- sacrificing user-oriented aspects first when trade-offs arise and
- focus on building the system first, then integrating the results with users.

Further, AS is not simply the result of over-automation or human error. Instead, they represent a failure to design a coordinated team effort across human and machine agents as one cooperative system.

Woods & Sarter summarizes in 2000 [18] that the potential for automation surprise is most significant in the following cases: (1) automated systems act on their own without immediate preceding directions from their human partner, (2) gaps in the user's mental models of how their machine partners work in different situations, and (3) week feedback about the activities and future behavior of the agent relative to the state of the world.

Decker provides a list of circumstances in 2002 [19] under which AS is likely to occur. The automation may be undergoing a mode change from someone who programmed it a while ago or follows a pre-programmed logic. There is insufficient feedback about its behavior; the automation communicates the status to the user. Event-driven circumstances can create situations where the automation dictates to the user how quickly to think, decide, and act. It may be difficult for the user to assess what input is required for the automation to do what the user wants.

2.2 Safety Assessment Methods

The preliminary literature search resulted in 16 methods that are briefly presented below. The methods provide a representative cross-section of the most common safety assessment methods and do not claim to be exhaustive. The chronological distribution of the presented methods should reflect the method development process over the years, starting in 1949 with the Failure Mode and Effect Analysis method. Earlier methods, such as Heinrich's Accident Triangle from 1931, have not been considered in this preliminary review because it is assumed that the results have been transferred to later generations of methods and thus are implicit there.

Failure Mode and Effects Analysis (FMEA) was established in 1949 [20][21]. FMEA performs the analysis inductively (from the bottom up). This is in reverse order compared to the fault tree analysis (FTA) model described below. FMEA is based on components or subsystems for which each type of failure is analyzed in terms of its potential effect on the system. The model requires specialist knowledge of the audited system. It is a structured method of finding weaknesses in the system. The technology is detailed and carries the risk of missing overall disruptions. The method focuses on components and does not explicitly address automation surprise (AS) or situational awareness (SA).

Closely related to FMEA is the FMECA (Failure Mode, Effects, and Criticality Analysis) model from 1949 [20][21]. Similar to qualitative error trees, a risk matrix can be used to assess the different types of errors. The failure mode, impact, and severity analysis introduces a column for severity, also known as severity. To fully use the model, specialist knowledge of the system under review is required. The method focuses on components and doesn't explicitly address automation, surprise, or situational awareness.

Human Reliability Analysis (HRA) was initiated in 1952 and originated in the nuclear power industry [20][21][22]. HRA is an umbrella term for methods for assessing factors that can affect human reliability in probabilistic risk analysis in a socio-technical system. There are many different methods with varying degrees of complexity. They are all based on the same underlying principle (the Swiss cheese model). HRA takes a structured approach to identifying potential human error events (HFEs) and systematically estimating the likelihood of error using data, models, or expert judgments. The method does not provide a clear picture of SA but indirectly by using performance shaping factors (PSF) for human activities. The method is based on expert knowledge and does not explicitly address AS. HRA is associated with probabilistic risk assessment (PRA) [20][21], based on fault tree and event tree analysis, and was established in 1965. The model is used in the nuclear power industry, but also a lot in oil and gas. Like HRA, the model addresses this indirectly because automation surprise can be treated as an unwanted event as a starting point. The method doesn't explicitly address automation surprise or situational awareness.

The fault tree analysis (FTA) model was developed in 1962 [20][21]. The model identifies collaborative events that can lead to malicious events. The method is mainly used if the end consequence is severe. However, the malicious events are challenging to model accurately. The model easily overlooks combinations and can thus give an inaccurate picture. Fault tree analysis can be complex and challenging to get an overview of. It is difficult to find relevant data if fault trees include quantification. If the FTA is done correctly, it is possible to identify combinations and weak links in the system. The model can also show how common an event might be. The model handles indirect automation surprise because the event can be treated as an unwanted event as a starting point. However, AS is not explicitly mentioned in the methodology. The same goes for SA.

Preliminary hazard analysis (PHA) or rough analysis was taken in 1969 [20][21]. The model detects malicious events with a focus on high-level events. The model is reported as risk levels and is often performed as an initial analysis to identify preliminary hazards. The model provides limited detection of causes. The method addresses AS indirectly because AS can be treated as an undesirable event as a starting point. However, AS is not explicitly mentioned in the methodology. The same goes for SA.

Hazard and Operability (HAZOP) was established in 1974 [20][21]. HAZOP is a systematic, team-based approach to assessing process risks and potential operational issues. HAZOP identifies hazards and possible operational issues in a system or process that are proportionate to the available level of detail and generates a set of actions to eliminate or minimize them. The model requires specialist knowledge of the systems being analyzed. This method is best suited for processes and operational processes. It is closely related to the FMEA method, but the failure modes have been defined and made uniform. AS can be treated based on specialist knowledge but is not directly supported by the method. The same goes for SA.

Hazard Identification (HAZIDE) was introduced in 1993 [20][21]. The model is a modification of the HAZOP model, specifically to identify human error. HAZID is a systematic, team-based approach to identifying hazards and their potential consequences. HAZID is used at different stages of a project or the life cycle of a system, including the operational phase. It is often used to identify safety, health, and environmental risks early in a project to help develop safer design options and to help guide future risk mitigation activities. HAZID identifies hazards in a system or process proportionate to the available detail level and generates recommendations and actions to help eliminate or minimize identified hazards. The results of a HAZID study should be documented as a HAZID report, and measures should be followed up and completed. AS can be corrected based on specialist knowledge and not directly from the method. SA is not explicitly addressed.

The event tree analysis (ETA) model was introduced in 1974 [20][21], and the model identifies how a peak event can escalate to possible eventual events. Event trees can be complex and challenging to get an overview of. Indicates possible and probable closing events. With simple means, a rough quantification can be made to support the probability assessment in the risk assessment. The model addresses AS indirectly because AS can be seen as the starting point of an unwanted event. However, it is not explicitly mentioned in the methodology, which also applies to SA.

The Bowtie model was introduced in 1979 [20][21]. The Bowtie model is based on fault tree and event tree combined in a single method. The Bowtie method analyzes a hazard or a critical event through cause and consequence analysis. The left side of the fly is formed by a fault tree, which models how combinations of primary events cause the danger. The right side of the fly is based on an event tree, which models the consequences of the threat. Fault tree analysis (FTA) and event trees (ETA) are based on linear cause-and-effect paths. The model can address AS indirectly by using AS as the starting point for an undesirable event. However, AS is not explicitly mentioned in the methodology. Similarly, SA is not used in the model.

In 1997, the "Barrier Analysis or Layers of Protection Analysis" (LOPA) model was introduced [20][21][23] . LOPA is a scenario-based risk analysis and can be said to be a simplified form of an error tree and event tree. Therefore, the model is linked to the Bowtie model. LOPA is based on quantifying frequencies, risk reduction factors, and probabilities of relevant deviations, barriers, conditions, and escalation factors. The method is used to assess and report on the adequacy of the barrier measures taken – a structured study of the existing barriers and the opportunity to assess the reliability of the current application. The model makes it difficult to find relevant data on initiating event frequencies and the probability of component failure. This can lead to both over- and under-evaluation of scenarios. Either AS or SA is used in the model.

The 2000 Safety Assessment Method (SAM) [24][31] is based on the following phases: Functional Risk Analysis (FHA), which identifies hazards and assesses their effects and severity; Preliminary System Safety Assessment (PSSA), which includes fault tree analysis, event tree analysis, common cause analysis; System Safety Assessment (SSA) which provides documentation of evidence, data collection, testing, and validation. SAM is a framework that contains methods and techniques for developing safety assessments of changes in functional systems for air navigation service providers (ANSPs). SAM presents a general overview of the safety assessment of air navigation systems from a technical perspective. The model focuses on engineering issues and not primarily on changes in the functional system. The model is based on both fault trees and event trees. AS can be handled indirectly as a starting point for an undesirable event. The same goes for SA.

Root cause impact analysis (CCA) encompasses applying the cause-impact diagram method to static systems [20][21]. The model was introduced in 2002 and aimed to model the sequence of events that can develop in a system due to combinations of fundamental events in diagrammatic form. Root cause impact analysis combines bottom-up and top-down techniques for binary decision diagrams (BDDs) and fault trees. The result is that potential accident scenarios are developed (and therefore not necessarily used in safety assessments). Neither AS nor SA are mentioned in the model.

Systems Theoretic Process Analysis (STPA) was first described in 2012 [20][21][25]. STPA is a qualitative risk analysis technique that assumes that accidents occur not only because of component failure but also because component behavior restrictions are not adequately enforced. It is used to identify cases of inadequate control that may lead to hazards, identify safety-related limits necessary to ensure acceptable risk and gain insight into how these restrictions can be breached. This information can control, eliminate, and mitigate system design and operational risks. STPA can be applied to existing designs or proactively to help guide design and system development. The STPA is based on the Systems-Theoretic Accident Model and Processes (STAMP). The model overview includes control measures provided to influence a controlled process, feedback that can be used to monitor the process, the process model (beliefs) formed based on feedback and other information, and the control algorithm that determines appropriate control measures given current thoughts. AS is mentioned mainly because the model is based on automatic control. SA is not explicitly mentioned in the model but can be identified as a hazard to the system.

The 2019 Specific Operations Risk Assessment (SORA) [21][26] is a step-by-step risk assessment process that aims to analyze the risks posed by certain unmanned aircraft operations and determine the necessary mitigating measures and robustness levels. The model is explicitly aimed at unmanned aerial vehicles and is based on traditional risk assessment methods – a qualitative method where AS can be handled indirectly. SA is not addressed in the model.

High-fidelity risk modeling (HFRM) from 2022 [26] . While SORA is a qualitative risk modeling with high reliability, HFRM is a quantitative estimate of the business's expected mortality (ERC). Neither AS nor SA were directly concerned. Neither AS nor SA are directly addressed. However, AS and SA can be acted upon as unexpected events to model the system behavior.

Regulation (EU) 2017/373 (2020) [27] is not a safety assessment methodology. The regulation aims to identify changes in the functional system from a safety perspective. Therefore, it may not always be possible or desirable to specify safety criteria in quantitative risk values in the safety assessment of functional systems. Instead, the safety criteria can be defined in terms of other risk-related measures. These metrics are called proxies, which are indirectly the measure of risk. 373 provides the option to use risk analysis (for example, traditional security methods such as SAM) in the case of vulnerabilities or the use of vulnerabilities in the case of proxies. The regulation distinguishes between air traffic service providers (required to carry out safety assessments) and non-air traffic service providers (necessary to prepare safety assessments for support). Neither AS nor SA are explicitly mentioned in the model. However, it can be fixed as failure modes that can initiate an unwanted event.

The Functional Resonance Analysis (FRAM) method [28] was introduced in 2012. FRAM is a systems-based method developed to understand complex socio-technical systems. FRAM focuses on learning from security events or undesirable conditions and can be used to understand how things work well in a system. This is included by identifying gaps between work imagined (WAI) and work performed (WAD). FRAM is used to model the features needed for the success of daily performance and can then be used to explain specific events by showing how functions can be linked together and how the variation in everyday performance can sometimes lead to unexpected and unscaled results, either good or bad. FRAM is based on the four basic principles: equivalence of successes and failures, approximate adjustments, emergence, and functional resonance.

2.3 Gap Analysis Safety Assessment Methods

Section 2.2 lists 17 assessment methods, with 15 not referring to any explicit help to identify automation surprise or situational awareness issues. These 15 rely on high-level model assumptions that require the safety analysts to define events independently. Using event criteria that involve human error and implications of situational awareness is the analyst's choice. In addition, methods give no procedural guidelines that add support to conducting empirical studies such as historical data or human-in-the-loop simulations. In the majority, the judgment on the probability of events and effects on operations is based on expert knowledge. Leveson's STPA method is an exception, which relies on a control model framework that defines operator and automation in a continuous loop. The analyst uses the framework to model interaction elements that pass the interface between the operator and automation. Concerning supplementary frameworks, FRAM can also model events associated with automation surprise and situational awareness.

The Human Reliability Analysis represents the franchising of methods that rely on predefined error classes applied to human perception, decisions, and actions.

Previous and current safety assessment methods have generally a Safety-I perspective, where risk analysis has been performed based on historical data. However, Safety-I assessment methods neither offer explicit support to identify emergent states in the human-automation collaboration nor to mitigate causes of AS. There were two exceptions that picked up concepts of how to assess human error probability. First are the methods related to human reliability analysis. They classify and estimate the probability of human error and consider the human as a machine component, following the example of a Probability Risk Assessment (PRA). Aspects of AS are partly covered by "omission of actions" or "error of commission". Secondly, STPA offers the possibility to consider human-automation collaboration and to find emerging states in the control loop.

FRAM is a supplementary framework that can be used in the scope of a safety assessment. It is capable of developing an understanding of how a sociotechnical system works. FRAM can be utilized to model any kind of performance or activity and can therefore be used to develop a model of a system's functions as a basis for analysis. Consequently, it should be possible to use FRAM to model the effects of automation surprise.

According to regulation (EU) 373/2017, the consequences of changes in a functional system should be expressed in terms of harmful effects of the change and the hazards associated with safety risks. This means that automation surprise can be handled as a failure mode, a starting point for an unwanted event with a harmful effect. Regulation (EU) 373/2017 demands that hazard identification aim to complete coverage of any condition, event, or circumstance related to the change, which could induce a harmful effect individually or in combination. Hence, according to (EU) 373/2017, the hazard identification process potentially can solely identify automation surprises with a path to a harmful effect.

The methods share common assumptions concerning risk modeling. Those that we see clearly here are linear relationships and binary-distributed event occurrences. Linear models have the advantage of being easily understood and applied by safety analysts. They base on the sequential principle of A causing B causing C and so on. This principle has its roots in the domino model according to Heinrich

[29] and allows the analyst to suggest causal relationships in the form of event tree and fault tree models into which safety analysts can embed any event of interest.

The application of linear relation comes along with two assumptions.

Firstly, there is a well-defined hazard event, which occurs or not (binary distributed). Further, a hazard event is specified using criteria and conditions of occurrence for the purpose to design worst credible case scenarios. The divergence between automation behavior and operator expectation may represent such a hazard that could impact safety negatively. Discrepancies that may arise when using AS phenomena as a hazard may include the following:

- AS is not a clear measurable event with sharp bounds because it is not necessarily related to a certain action. Rather it arises from an invalid expectation of automation behavior, which is simply tied to cognitive processes and ties up the operator's capability to anticipate. The concept of situation awareness might be a good approach to explain AS by a divergence of the anticipation of the situation.
- A closely related side effect of this sequential principle of linear models concerns the nature of hazardous events, which are considered binary distributed events: “On” or “Off”. This is at odds with the states and processes of the operator's situational awareness, which are inherently viewed as non-binary but continuous: “More” or “Less”.
- As AS might trigger uncountable variants of reactions of the operator, a broad range of possible follow-up scenarios need to be considered in order to assess the consequence of AS.

The AS phenomenon is a generic event, based on retrospective investigation. As such, it has not undergone concretization and contextualization at the application of future implementation, nor does it describe an event with sharp boundaries. It is therefore more the task of the safety analyst to concretize AS into the application context and define what exactly the surprise might be.

Second, there are also preceding and subsequent events connected to the hazard event and depend on conditions or transfer probabilities. Effects and causes shall be assessed, monitored (or observed) and mitigated if tolerable limits are exceeded. Complexity, on the other hand, acts in a network of dependencies that does not feature linearity. This sequential principle is in conflict with the nature of dynamic systems and complex socio-technical systems, showing a large number of interdependencies. Most high-reliability organizations involve a surge in complexity and, consequently, requirements on performance variability. These complex systems are elaborate and contain many details, and the principles of some of the system functions can partly be unknown. These systems are interdependent on other actors in the system and at the system boundaries, and the system changes before the description of the system can be completed.

Linearity of event occurrence does not capture humans as a multidimensional complex of states, processes, and systems knowledge. The inevitable consequence is that relations between states are ignored, which thus contributes to enlarging the gaps in the barriers. If we take the safety assessment at the remote tower as an example, the impact of, for example, a black screen on the visualization depends on the traffic situation and the ATCO's situational awareness. The traffic situation is then dependent on the traffic but also the ATCO clearances, which is based on procedures and ATCO training and experience. Also, the outcome of the situation is dependent on the ATCO and the situation awareness when the failure occurred. This sequence of events is difficult to describe as cause and effect. One could say that there are more conditioning events than basic failure events. In these types of systems, the key to successful performance lies in the ability of the human operators to compensate for incomplete procedures and instructions and adjust their performance accordingly. Hence, adjusting performance is necessary to match the ever-changing system demands, resources, and constraints.

3 EXPLORING AUTOMATION SURPRISE

3.1 A safety database investigation in ATCC Malmö

A selection of incident reports/error reports from Malmö ATCC have been reviewed. While the current ATM system is lower on automation than e.g. a cockpit, many if the tools and aids for ATCOs function automatically to some extent. As a result, AS events (or "unexpected system behavior") do occur. Underlying factors can be system errors, incorrect operator input, and incorrectly set parameters, among other things.

Reporting of such incidents contributes to system corrections and/or information efforts to describe the system functionality – all to maintain and increase aviation safety.

The ATM system contains several parameters and data to effectively identify and present such conflicts and risks to the ATCO. These parameters construct a profile for each flight, e.g., a 4D presentation of the flight path through the sector. These parameters can, simplified, be divided into two categories:

Static parameters, e.g., dataset constraints or aircraft performance databases:

- Aircraft performance (vertical speed in different altitudes)
- Dataset constraints (flight profile adjusted to avoid certain airspace volumes)
- Flight plan data (route, requested flight level, horizontal speed during climb/descent)

Non-static parameters:

- Ground speed, horizontal speed in cruise
- Aircraft position, surveillance data
- Manual input (vertical/horizontal change in aircraft profile)

What can cause an AS event is when one or more of the above parameters lead to unexpected system behavior for the ATCO. For example:

- Incorrect conflict warning (absent, misleading, or superfluous)
- Incorrect sector sequence
- Incorrect/varying profile calculation (see figures below)
- Missing flight plan information

As previously mentioned, the ATM system does not initiate or execute actions but presents information to ATCO. However, several of the tools used by ATCOs work automatically "in the background". Some completely autonomously, others with a combination of human input and fixed parameters such as flight plan information.

Regardless of this description, the ATM system works well and serves an essential purpose in supporting our ATCOs. The system is constantly developed and evolving, as are the ATCOs, with new functionality and understanding of how the system works.

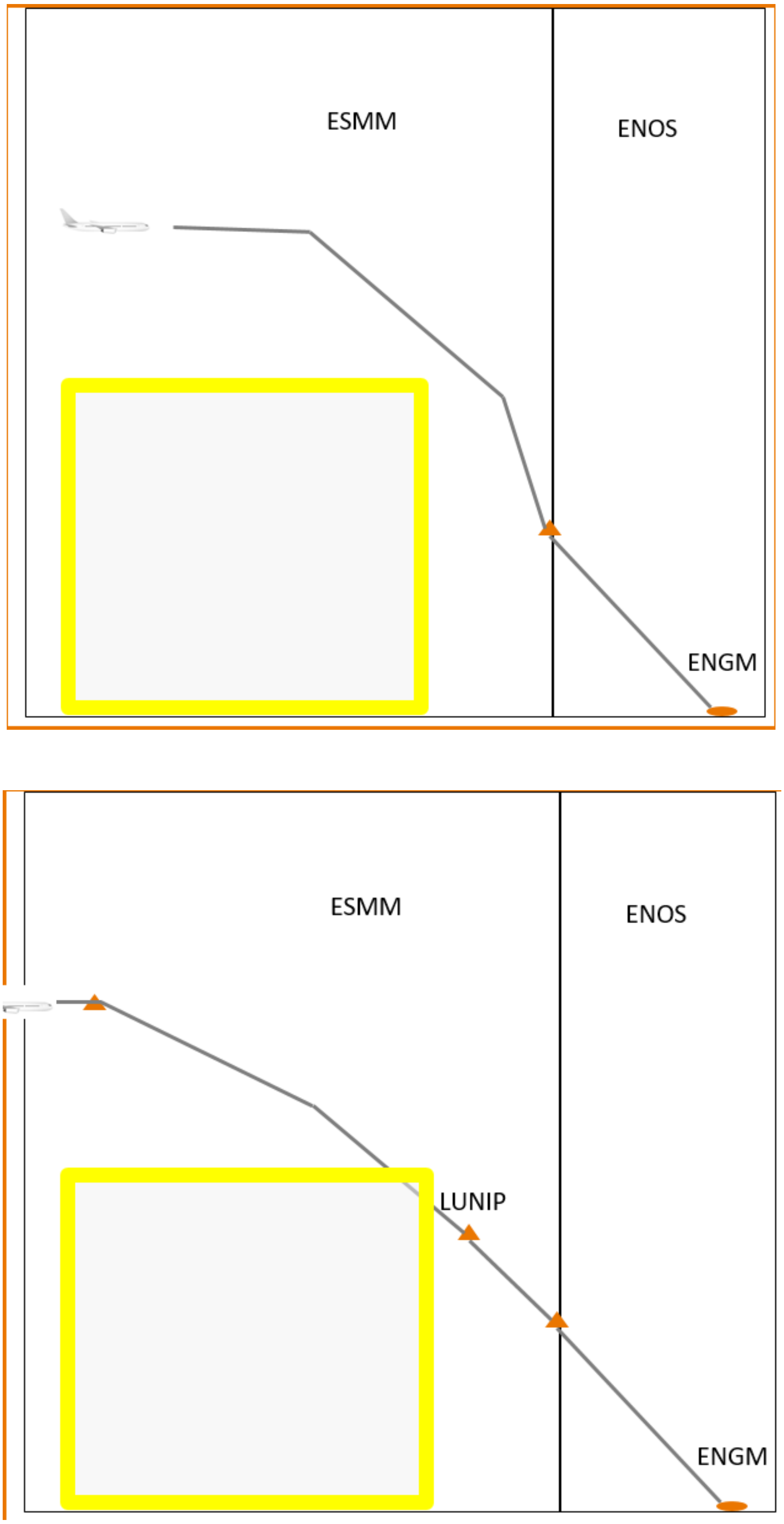

Figure 2: show two system-calculated vertical profiles, one of which considers the yellow military sector–the other does not. This difference can occur for the same flight if specific parameters/inputs change the profile accordingly.

3.2 A survey study in ATC Enroute

A survey study was conducted at ATCC Malmö and Stockholm to identify the actual relevance of AS in an operational ATC environment. This is of particular interest as the working environment of the en-route controller is today supported by systems that correspond to low levels of automation, in most cases information acquisition and analysis. AS was proved to be of safety relevance in the flight-deck work environment by de Boer and Hurts 2017 [12], environments with support of low-level automation were never investigated. The survey focuses on two questions:

1. Is AS observable in En Route operations today?
2. If yes, what is their appearance, and what is the difference from those observed on the flight deck?

The survey was supported by n=48 air traffic controllers with active licenses who were on duty at the LFV control centers in Malmö and Stockholm. Controllers operate with the Top Sky System manufactured by Thales at both control centers, providing control services as required by Swedish air regulation law and using the same operating manuals and procedures.

The question sheet was designed in the scope of a Master of Science Thesis authored by Fritz [30]. To ensure comparability with the flight deck, the questions were kept similar to those designed in de Boer and Hurts 2017 [12]. A copy of the questionnaire sheet is left in the online database, further described in section 9.

The result of the analysis shows that there are surprises observed during the time in the work position that match the definition of automation surprise. Figure 3 shows an exemplary analysis that illustrates the participant's response to the observed time on shift as a relative frequency over 24 hours. Two major peaks are observable between local time, 9 am and 12 am, and 6 pm and 8 pm.

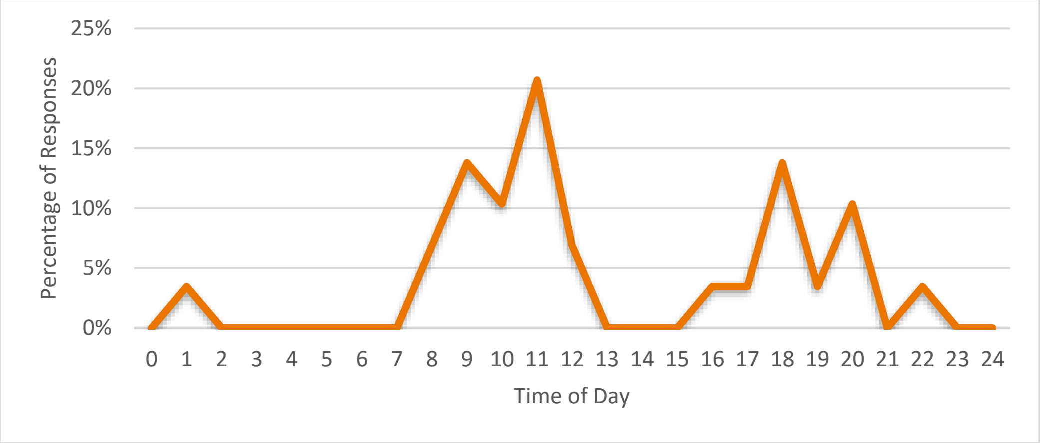

Figure 3: Relative temporal frequency of observed automation surprise over 24 hours.

Figure 4 shows the distribution of the contributive factors during the surprise occurrence. The figure shows the results from the de Boer and Hurts flight deck for comparison. “System Malfunction” (42%) is identified as a major factor, followed by “System Knowledge” (23%) and “Unclear Display” (23%). Surprisingly, “Fatigue” was not considered a contributing factor, which sets a contrast to the flight deck.

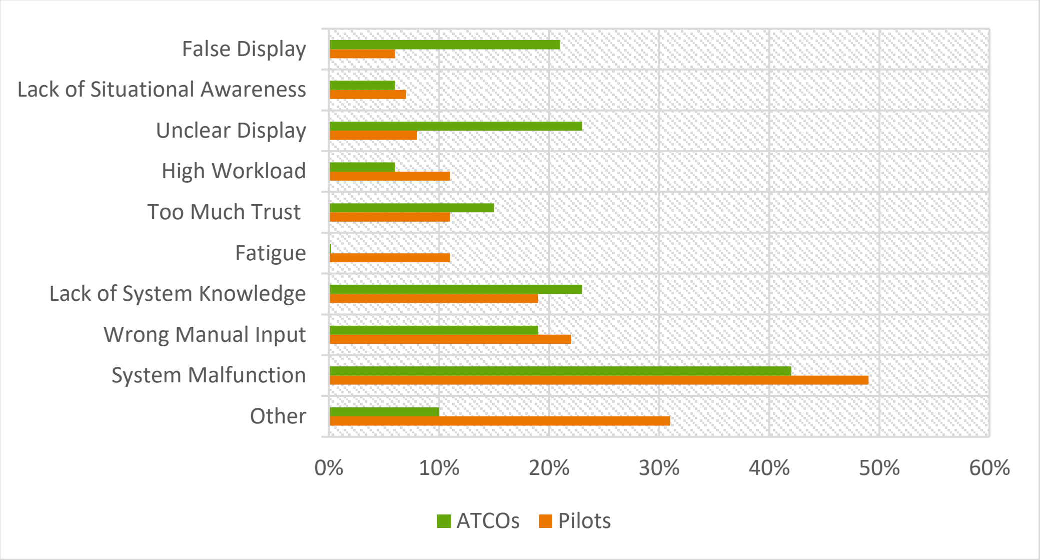

Figure 4: Reported Contributing Factors of Last AS-Event

The conclusion is that the automation surprise is a significant event in today's LFV En Route operations. This answers question 1). However, the responses show that there is no case where the occurrence of AS has contributed to a loss of separation or was considered a severe risk. This means that the operators have managed to contain the effects of the surprise event by early detection. This performance is a key element to prevent further escalation beyond the barrier of the working position.

The most significant difference to the flight deck is the higher response in the categories “False Display” and “Unclear Display”. In contrast, workload issues and fatigue were considered minor factors. Further analysis and explanations are provided in the Master of Science thesis by Fritz 2024 [12] and the related conference article by Fritz et al. 2024, referred to in section 9.

3.3 An Interview Study

With the scope to explore the status quo knowledge and create a good picture of the actual relevance, an interview study was conducted. This interview study was conducted on a smaller scale, involving scientists known in aviation psychology and safety with a well-acknowledged reputation in human performance. The recruitment of an appropriate set of interview partners with this level of competence is based on the individual's regular engagements at conferences such as SESAR Innovation Days, ATM Seminars, and the European Association of Aviation Psychology.

Table 1: Interview partners with competencies in aviation psychology and safety.

NO	INTERVIEW	DATE
1	Prof. Dietrich Manzey, TU Berlin	2022-11-09
2	Prof. Hartmut Fricke, TU Dresden	2022-11-10
3	Prof. Harald Kolrep, HMWK Berlin	2022-11-10
4	Prof. Henk Blom, TU Delft	2022-11-15
5	Brian Hilburn, MITRE	2023-09-19

The following questions were developed by the SB2030 team, aiming at the causal relationships in terms of contributing factors (simplified as "causes"), possible consequences, and mitigation measures.

SECTION	QUESTION
1	General
1.1	What is automation? (Examples and Definitions)
1.2	What are the arguments for automation? ATCOs may say that the current situation is "sufficiently good". Why change anything?
1.3	What are the risks of automation?
1.4	Definition of Automation Surprise: Automation surprises occur when "crews are surprised by actions taken (or not taken) by the automated system". Do you agree?
2	Causes
2.1	What are the causes of automation surprises? Free dialogue
2.2	How is workload related to automation surprise?
2.3	How is automation level related to automation surprise?
3	Consequences
3.1	What are the contextual conditions that make automation surprise cause severe consequences?
3.2	What is the importance of detecting deviant automation behavior in terms of severity?
3.2 a)	Non-detected (undiscovered surprise)
3.2 b)	Self detected vs extern-detected
4.	Mitigations
4.1	What mitigation possibilities do you propose?
4.2	What training possibilities do you propose?
4.3	What about the selection of pilots and ATCOs?

In all, 17 methods were reviewed, 15 of which do not provide explicit help in identifying automation surprises or situational awareness issues. These 15 methods rely on high-level model assumptions that require the security analyst to define events independently. Event criteria involving human error and situational awareness consequences are the analyst's choice. In addition, the methods do not provide any procedural guidelines supporting empirical studies such as historical data or human-in-the-loop simulations. In most cases, assessing the probability of events and effects on the business is based on expert knowledge.

Leveson's STPA method is an exception based on a control theory framework that defines operators and automation in a continuous loop. The analyst uses the framework to model interaction elements that send the operator's and the automation's interface.

Regarding complementary frameworks, FRAM can also model automation, surprise, and situational awareness events. Human Reliability Analysis represents the franchising of methods that rely on predefined error classes applied to human perception, decisions, and actions.

4 A NEW APPROACH TO SAFETY ASSESSMENT

4.1 Requirements

Based on literature reviews and explorative studies into the AS phenomenon described above, we have developed safety assessment requirements that may support the identification of AS-related risks:

1) Develop a model that describes the AS phenomenon with cause-effect links in a chain of events.
2) Integrate AS into established Safety I models, treating AS as a hazard with potentially severe outcomes.
3) Consider the human operator in an interaction loop involving automation and the environment.
4) Facilitate the propagation of human error to automation and the environment and operations.
5) Assist the safety assessor in identifying potential emerging situations related to the AS phenomenon in new systems.

4.2 A Safety I-framework for Automation Surprise

A principle delivery is the development and justification of a new approach to safety assessment, focusing explicitly on the automation surprise phenomenon. The purpose is to set the basic assumptions that support safety assessors in their task to prospectively identify the potential for automation surprise. The pivotal challenge lies in integrating the AS phenomenon into models used in Safety I. This presumes the AS phenomenon to be deployed as an event model with causal links to pre-states and consequences.

The following subsections describe our approach to translating the phenomenon into a safety-I rationalized concept, using the event "discrepancies of expectation" as a translating element that bridges the gap between both conceptional assumptions of AS and Safety I. The steps are necessary to integrate AS into a regular safety assessment process, involving event models such as the "hazard". An appropriate frame on which most of todays Safety-I rely on is Reason's model, "Trajectory of Accident Opportunity". It defines and classifies AS as part of a human "error detection and correction". The definition of a taxonomy gives AS, to the first, the "discrepancy of expectation" a central role and attributes a formal set of suits describing the external appearance using qualitative traits. Even the seemingly similar appearance of AS can be distinguished using this taxonomy.

4.2.1 Adopting AS as an Unsafe Act

As mentioned earlier, AS is defined as

> *"An Automation Surprise occurs when the automation behaves in a manner that is different from what the operator is expecting"* [1].

As defined here, the expectation and the factual automation behavior constitute a contradiction or, more simply, a discrepancy. We see the discrepancy as a central trait of AS that is not necessarily the surprise itself, as the discrepancy may stay undetected or have occurred some time in advance. Therefore, it seems logic to differentiate between the discrepancy and the surprise. Following this logic, we define two key event

1) The "discrepancy of expectation", defines the event in which the operator's expectation significantly deviates from the factual automation behavior;
2) The surprise event is a consequence of the discrepancy mentioned right before. The first event is a prerequisite to the latter.

Based on this assumption, a risk model of automation surprises will be presented based on James Reasons' "Trajectory of accident opportunity", also known as the Swiss Cheese Model of Accident

Causation (see Figure 5). The "unsafe act"[2] is represented by a "discrepancy of expectation," considered a pivotal prerequisite to the arousal of surprise. The discrepancy may originate in the organization, training, system understanding/knowledge, or system development, which will be addressed in the following subsections. In terms of the safety assessment, we consider this event a 'hazard', defined by Eurocontrol SAM 2006 [31] as

A hazard is anything that might negatively influence safety.

Further, the surprise is a possible but unnecessary consequence of the discrepancy. This qualifies the surprise to be a response to another event considered relevant in the chain of events, the detection of discrepancy. Surprisingly, these model assumptions turn the AS into a part of a safety barrier that allows the operator to become aware and take measures to prevent further escalation. As such, the surprise becomes a principal role in an error detection and correction mechanism, as described by Reason 1990[3].

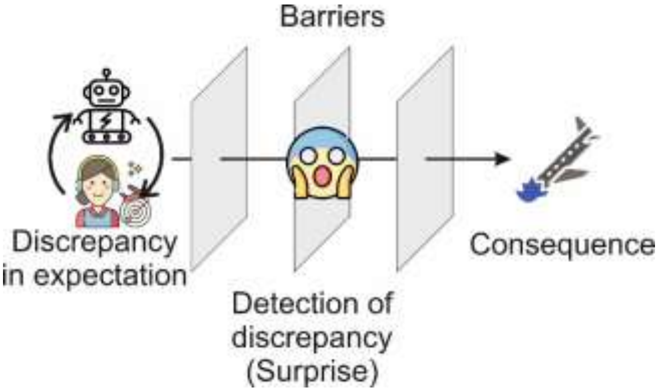
Barriers
Discrepancy in expectation
Detection of discrepancy (Surprise)
Consequence

Figure 5: Automation Surprise-Safety Model adaption based on Reasons Model "Trajectory of Accident Opportunity" [32]

4.2.2 Safety-relevance of Critical Cues Perception and Mental Models

The expectation is ".….the feeling or belief that something will or should happen" (Cambridge Academic Content Dictionary). This definition sets the presumption that the operator bases the expectation on the projection of a future state. Generally, the expectation may be considered a multi-factorial product generated by a human based on age, experience, attitude, prejudice, and other attributes that shape the individual expectation.

As part of the situation awareness loop, Endsley defines the projection of future states and the expectation to be a product of a well-developed mental model, which is the outcome of experience gained with a particular system:

> *"When an individual has a well-developed mental model for the behavior of particular systems or domains, the model will provide (a) for the dynamic direction of attention to critical cues, (b) expectations regarding future states of the environment (including what to expect as well as what not to expect) based on the projection mechanisms of the model,…"[15]*

This reveals the mental model of systems/automation to be of crucial relevance, drawing attention to those cues that deliver the information necessary to develop the right expectation of automation behavior. To refer to Endsley's terminology, "cues" that may be relevant for safety-critical decision-

[2] The unsafe acts of aircrew can be loosely classified into two categories: errors and violations [32]

[3] This error detection and correction mechanism according to Reason sees the operator into a constant progressing loop of actions and reactions by which the operator may offer the chance to detect errors and threats and correct for them [32].

making are here "critical cues". As cues, in general, may reveal time-dynamic and subject to expiration, critical cues remain inside a permanent and regular synchronization process, aligning the mental model and related assumption with reality. The frequency of reverification may depend on the half-life period of the information associated with the cue and its dependency on the situation dynamics. Establishing a work pattern that considers these time dynamics by verifying assumed cues timely is in the operator's hands. This ensures the decision-making relies on a sufficiently matured mental model.

With this background, the definition of a cue is as follows:

A cue is an elementary unit of information used for decision-making by the operator

The definition of the critical cue extends this scope by its safety relation

A critical cue is an elementary unit of information used for **safety-critical** decision-making by the operator

An example of critical cue perception is given in Figure 6 from project ETRA [33], which indicates a simplified visualization of the scan path of an air traffic controller during inbound traffic from initial contact to taxi clearance.

4.2.3 Risk of Unverified Assumptions

Cues may have been misperceived, incompletely understood, or not perceived. This case is summarized under "unverified assumption", in which a cue turns out to be insufficiently accurate for the decision process to be taken. The cue is considered invalid or "corrupted". Mental models based on invalid cues pave the way to an incomplete/corrupted expectation of the future projection of the environment. In the case of a critical cue, the discrepancy in expectation is given room to unfold.

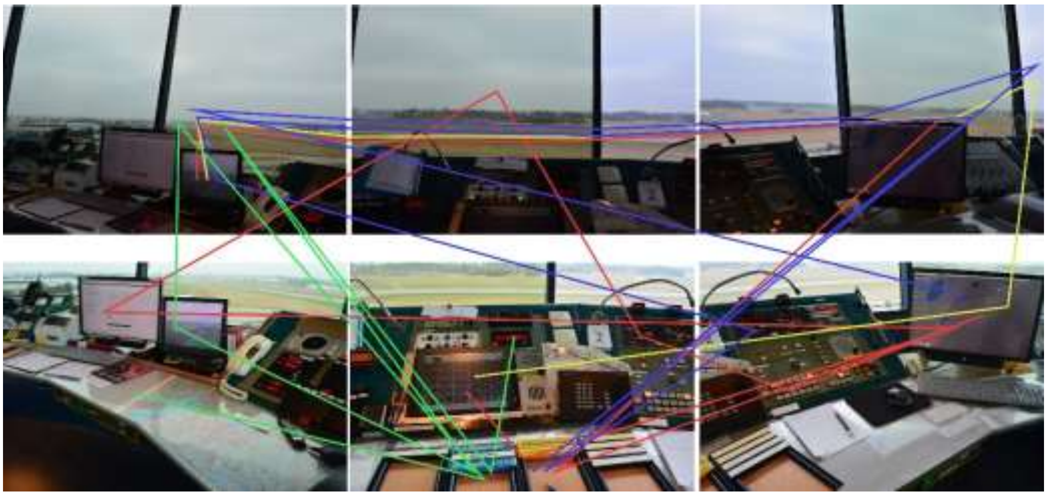

Figure 6: Air Traffic Controller Visual Scan Pattern, SAAB Conventional Tower, Landing KLM 1181 via Bedos FL100, 28 August 2020

Example: Asiana Airlines Flight 214
At July 6, 2013, Asiana Airlines Flight 214, the Boeing 777-200ER operating the flight, stalled and crashed on final approach into San Francisco International Airport in the United States. A crucial factor in the course of accident was the unintended deactivation of the autothrotte due to a mode confusion. Auto thurst (autothrottle) is an example among others, which may be expected to be activated during cruise and landing, constituting a cue as part of the pilots expectation concerning the thrust setting. The indication is visualized in Figure 7 as the example of the A320 primary flight display. The visual indication of the auto-thrust being de-factor activated may verify the Pilots expectation. This assumes

an intentional checkup of the indicator LED as part of a regular work scheme or by chance finding when striving over the indicator LEDs of the work panel for "all is well".

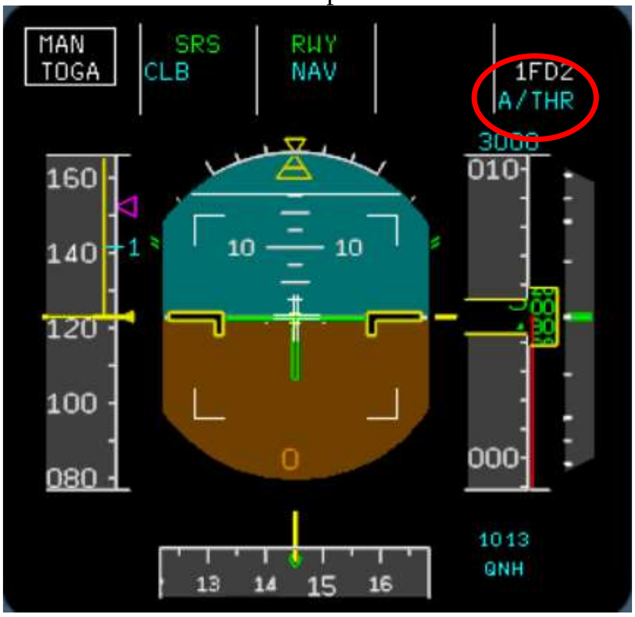

Figure 7: Auto-Thrust indicator on the A320 primary flight display

4.2.4 Classification of the Discrepancy of Expectation

The discrepancy of expectation event is assumed to be allocated to characteristic attributes, or qualitative traits, summarized in the AS taxonomy model depicted in Figure 8.

4.2.4.1 Origin

A principal attribute is the **origin of a discrepancy**, which constitutes the discrepancy as a critical event in a chain of events. A common approach is dividing the discrepancy's origin into functional and non-functional. The latter assumes the human expectation to be invalid, expecting something the automation may not provide, although it may operate as specified. The first considers the automation behavior deviates from its pre-specified behavior. An example might be a technical malfunction or a failed update process that applied an invalid software/hardware release that was mismatched with the education and training the operator received. These two cases are depicted in the AS taxonomy to the far right in Figure 8.

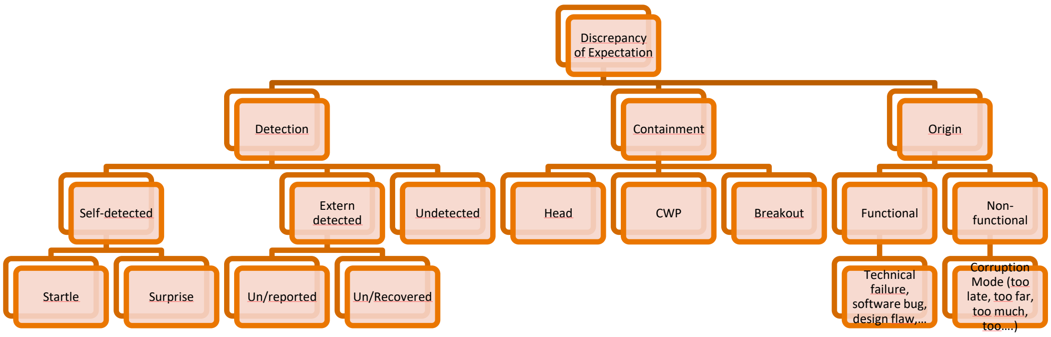

Figure 8: Discrepancy of Expectation-Attribute Taxonomy Model

Functional discrepancy assumes the automation to fail or provide corrupted services, as defined by the "functional hazard". The assessment of the functional discrepancy is covered by the assessment methodology related to technical equipment as described by Eurocontrol's Safety Assessment Methodology (SAM). This case is not covered in this project's scope as methodological knowledge on technical failures has matured sufficiently over the decades.

4.2.4.2 Detection

The second attribute of the discrepancy event is the **detection event,** in which the operator identifies the discrepancy as a threat. This detection event is part of the operator's error detection and recovery mechanism. The detection may be distinguished between self-detected and external detection. Self-detected detection covers the case in which the operator in charge of the specific automated task of the subject is the same as the one subjected to the discrepancy. The automation surprise then describes the event of detecting that the expectation deviated significantly from the de-factor behavior of automation. The detection event may even involve a physiologic response in which the operator is subject to an automation startle. This is to be distinguished from an external detection in which a person other than the one being subjected to discrepancy succeeds in identifying the discrepancy. This role could involve a colleague or an incident/accident investigator who has the chance to succeed in the detection as availability or accessibility to all information necessary is given. To give a concrete example, a second pilot or a checker on the flight deck may detect issues with a poor setup of the FMS, which has not been detected previously by the one setting it up. The comrade check is a procedure deeply rooted in the aviation industry safety culture for the cooperative detection of human error. A third and assumingly not unusual case is the discrepancy remaining undiscovered. A case in which the discrepancy remains unrecovered as accomplishing any recovery presumes the prior detection by any operator involved. No reliable figures on the number of unreported cases are known, but it is assumed that some run dead like a rolling stone at the flattening end of the slope with an unknown escalation. A description that best matches this attribute is 'latent condition', with a minority of recurring discrepancies finding one-day conditions to unfold.

4.2.4.3 Containment

The third attribute considered significant is 'containment', which describes the degree of escalation to which the effects of the discrepancy unfold. The containment assumes the detection and successful recovery of discrepancy is accomplished at some point during events. The first degree of containment assumed the impact of being bounded mentally by adjusting the expectation while the automation was found to operate as intended. No physical actions are needed to recover, whereas the surprise and the required adjustments to the expectations may be found stressful. One further escalation step may reveal inputs made by the operator to the automation that are invalid due to a prior discrepancy taking effect. The main point is that no further actions were taken than those at the working position and the systems involved. Recovery actions are limited to adjustments to the automation setup. This contrasts the last escalation step, called 'breakout', in which an operational effect is significant regarding measurable actions to aircraft control and related flight path and speed changes that were against the operator's expectation. The recovery requires corrective actions to re-establish the intended state of operations with the required safety margins.

4.2.5 Non-functional Discrepancy in Automation

Focusing on Non-functional discrepancy, the description matches with an event referred to as "functionally unimaginable hazards" as defined by Eurocontrol SAM 2006 [31] and illustrated in Figure 9:

> "There are hazards that are hard to identify by means of the functional approach. Such hazards are called "functionally unimaginable" or shortly "unimaginable" hazards."

Examples given by SAM are "situational awareness problems of pilots" or "air traffic controllers might become overly reliant on a well-functioning alerting system". The definition and the examples identify the human-induced expectation discrepancy as an "unimaginable" hazard in the scope of a Safety-I assessment. This sets the expectation and its creation process as pivotal in the assessment.

As introduced in section 4.2.2, the operator's expectation builds upon a specific amount of critical cues. We define these critical cues to contain safety-relevant information, considered time-dynamic with a certain time under which the information is valid (see Meyer et al. 2022 [34] for further explanation of cues in the decision-making of ATCOs). Critical cues can be affected by a so-called "corruption mode", which compromises the cue by a certain feature called "mode". This affects the overall expectation in consequence. The explanation and the related concept of "corrupted critical cues" are provided in section 4.2.3. We define the non-functional hazard as follows:

Non-functional discrepancy = Critical Cue +Corruption mode.

Some examples:

- Cues are: "Warning message", "position indicator", "speed indicator", "mode indicator", "recommendation issued by automation",…
- Corruption modes are "too early", "too many", "too late", "too small", "too poor", "too silent", and "too similar/undistinguishable",…

More examples are provided in section 4.3 below.

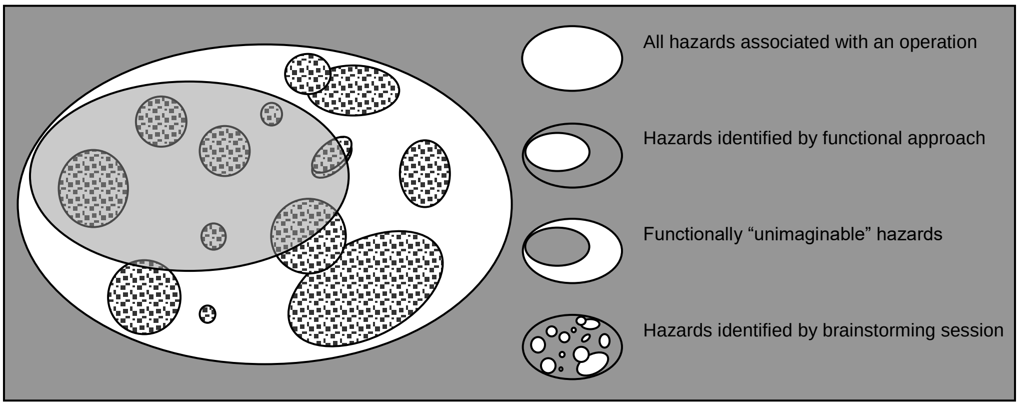
All hazards associated with an operation
Hazards identified by functional approach
Functionally “unimaginable” hazards
Hazards identified by brainstorming session

Figure 9: Identification of Hazards according to Safety Assessment Methodology of Eurocontrol [31].

4.2.6 Cues and Corruption Modes Considering Levels of Automation

Characteristics of critical cues may vary depending on the automation level. The model of automation levels used in the ATM Masterplan is simplified into four elementary classes

(1) Low automation levels of **information acquisition and analysis** may involve cues specifically intended to enhance situation awareness by information that is otherwise more time-consuming or impossible to retrieve without computer aid. Such automation provides extra comprehension of the overall situation. At such low-level automation, critical cues involve operational, relevant information that **optimizes situational awareness** for increased safety and effectiveness. The general assumption is that the information provided at this level can be seen as a mere extension of that information, which is presented anyway. Such extensions still allow work to be carried out in an old-fashioned way, as it could be done without such automation. Therefore, the information provision is to be seen as optional.
However, the information provided by automation may conflict with the operator's genuine work pattern, which involves task management, control of attentional focus, and timing of the sequencing of activities. Integrating information, especially critical cues, into the prevailing work pattern is to be seen as an adaption process to optimize the behavior concerning the benefits of such automated information provision. A possible downside of the benefit is a complacency effect due to the operator's overconfidence in the automation over time. In general, automation complacency is when the operator views the automation as superior and does not question the automation's assumption of how the information was generated. Operators affected by the complacency effect tend to show opportunistic or reactive behavior, allowing the information cues to shape the work pattern significantly.

Consequently, corruption modes rely primarily on the operator's overconfidence in the information automation provides. This concerns the impact on decision-making when operators establish a reactive behavior of the information supplied. Automation has, in this case, a behavior-shaping or sensation-seeking effect. Another point is that model assumptions, under which the provided information was determined, deviate significantly from how the operator would have generated the information them self.
Examples are the secondary radar-based separation tool (sep tool) that assumes a projection of the course and speed into the future based on the **linear extrapolation assumption**. Another example is Medium-Term Conflict Detection (MTCD), which is based on the assumption of trajectory predictions. Predicting future traffic situations generally relies on mathematic operations with assumptions about the real-world behavior that may be right for most cases but not under all circumstances.

(2) Levels of automation, in which automation is assigned **to recommend actions,** provide cues that optimize decision-making. A smaller focus is on enhancing the operator's situation awareness instead of giving the operator an action alternative. The operator assumes the automation to rely its recommendation on an appropriate comprehension of the overall traffic situation, the identification of the specific challenge given by the situation, and the evaluation of an appropriate solution in the light of that specific challenge. Critical cues address these assumptions made by the operator, who seemingly knows about the automation processes and logic in the background. The corruption mode that may arise from these critical cues is a conflict between the automation and the operator for any of the above-given process steps. This involves first the comprehension of the specific challenge of the situation and, second, the logic of the planning. Neither is necessarily aligned with the recommended action's quality nor timing. The resulting corruption is that neither the resulting traffic solution nor the timing is compatible with the operator's overall strategy of traffic flows in the controlled airspace.

(3) Levels of automation under which **partial or full autonomy implementation of actions** is given do have similar cues to those suggesting action recommendations mentioned prior. The key

difference here is that single decisions are extended by a logic that controls processes involving a time-dynamic sequence of decisions. The cues refer to the parameters constituting the automation behaviour in an operations process. Corruption modes involve unverified assumptions about the automation's parameter setting and operational modes, such as transitioning the modes without pilots' awareness (mode confusion). A good example is the Flight Management System (FMS), which provides a solution from after the initial climb to the final approach, involving competencies of flight guidance and navigational skills along most flight phases.

4.3 A Guideline to Safety Assessment Tackling Automation Surprise

The guideline developed in this section aims at AS-related hazard identification as part of an overall safety assessment process. This shall consider the AS taxonomy and the assumption in the previous section. By this, the requirements line out in section 4 shall be met.

Besides the theoretical framework, the guideline builds upon Eurocontrol SAM 2006 [31] and de Jong 2007 [35], providing recommendations on conducting safety workshop brainstorming to identify unimaginable hazards. The guidelines given here extend these recommendations by a so-called "expectation-centered" approach that extends the hazard identification by the AS taxonomy introduced above.

Conduction of a safety workshop is assumed with the required competencies and roles specified in Eurocontrol SAM 2006 [31]. The following steps are recommended

1. **Identification of automation and related functions**: The task is to note the automation function. The function is being identified on the application level using indicators that are part of the interface. Examples are: "Speed indicator" in the cockpit, "distance indicator" on the radar system, "visual warning indicator" for the MTCD, or even "acoustic indicator" for TCAS warning or advisory.
2. **Sum up the critical cues of interest per function that interest the operations.** Sum up all information provided by that specific function for decision-making that is relevant for deciding about the frame of the decision, the planning of the best solution, and the timing and monitoring of the solution. Even recovery actions should be covered by potential non-normal situations that might deviate from the expected daily operations. Examples of simple critical cues are given in the following table.

CATEGORY	EXAMPLES
Aircraft	Type, Callsign, Performance, Altitude, Flight Level, Speed, Track, Heading, Intention, Flight Plan, Distances, Closest Point of Approach, Timing,
Weather	QNH, wind, shear winds, visibility, runway condition, precipitation, clouds, etc.
Airport	Runway configuration, active runway, runway length, runway occupation, runway clearance, taxiways, crossings, objects such as cars and pedestrians, and the airport environment with terrain and buildings….
Exceptions	Restriction zones, Emergencies, Construction Sites, Procedural changes,

Critical cues may vary strongly depending on the specific function of the automation. Therefore, it might be necessary to decomplement cues into sub-elements and proceed from there. For example, flight level might be distinguished as cleared, actual, or requested flight level.

3. **Allocate a corruption mode to the cue.**
   A discrepancy between the actual automation setup and the expectation may unfold if a cue remains unverified by not understanding, comprehending, or disregarding information. The designation of a corruption mode serves the purpose of describing the origin of the corruption and why the verification did not succeed. Examples are given above, such as "too small", "too similar", and "too late". Even basic knowledge, training, and education of the automation in use might be involved, assuming an incomplete understanding of the underlying principles, rules, and algorithms.
   Referring to the example of the auto-thrust, it might be a reasonable explanation that the visual indication is "too small" or "too hidden" due to improper design of the visual indication, failing to make the pilot aware of a discrepancy of expectation. In the case of the accident of Asiana Airlines Flight 214, the accident investigators concluded the following: "The flight crew's mismanagement of the airplane's vertical profile during the initial approach led to a period of increased workload that reduced the pilot monitoring's awareness of the pilot flying's actions around the time of the unintended deactivation of automatic airspeed control.". This description supports the assumption that the deactivation of the auto-thurst happened with "too small/unclear" indications, with a pilot flying consequently being unaware of the change in the setup and the cue "auto thrust activated" remaining unverified.
4. **Define the hazard by describing the cue to be relevant in the respective situation and the overall picture of expectation**. Describe then how it failed to match with the automation setup and behaviour using the corruption mode as an explanatory attribute associated with the cue.

The succeeding steps of defining the operational conditions, operational effects, and consequences are to be performed according to the Eurocontrol guidance material corresponding to the usual assessment process.

4.4 Additional Material to Support the Hazard Identification

Identifying expectations, cues, and possible corruption modes may be supported by a life experience of the automation of interest to operational experts involved in the brainstorming. This life experience is intended to refresh the expert's awareness of the functions in question. The life experience may involve

(4) A demonstration of the automation in the simulator or a mock-up
(5) Extending (1) by a media recording that might be used for applying the retrospective think-aloud technique

During brainstorming, operational experts may refer to specific situations with a traffic context and a specific state of the automation. The identification is thus extended by a situation context where a discrepancy may occur.

As media recording, video camera or screen recordings with voice com are suitable, helping the expert to conclude on the cues available on the environment, traffic situation, and system state as well as the projected future situation that they expect

5 APPLICATION TO THREE CASES

The guideline has been tested using three cases considered innovative and reasonable candidates concerning the arousal of Automation Surprise. The cases cover different levels of automation from 0 to 5, from information acquisition to partly autonomous action implementation. ATM Masterplan defines the level of automation on a scale from level 0 to 6, from information acquisition to full autonomous automation.

5.1 AI-supported decision-making in-route (Level 2 automation)

This case assumes the provision of conflict detection and resolution (CD&R) in enroute control using an AI-supported assistant. The assistant supported the controller with recommendations to solve the conflict effectively. These had to be issued in a timely manner and implemented by the controller.

The related workshop was conducted on 2 December 2022 with two air traffic controllers from enroute. One has participated in experimental data collection, simulating traffic situations and conflict scenarios supported by such an exemplary AI-supported assistant. The resulting hazard log is transcripted and available under the online transcript database section 9.

The conclusion of this case is, in general, positive. Applying the methods recommended by the guideline works fine. The efforts to think and identify cues that are of relevance during decision-making remain to be a challenge.

5.2 Multi Remote Tower (Level 0 automation)

The multi-remote tower concept was developed in the house of LFV with the central idea of providing tower control services to two airports at a time by just one controller. This idea became realistic with the emergence of remote tower technologies, allowing the present high-fidelity video-based visualizations of the airport environment (visual presentation) right next to each other.

The principle step of innovation does not involve more automation than **information acquisition and analysis**, using radar screens, electronic flight strips, and overlay elements presented on the visual presentation, such as wind information and atmospheric pressure QNH. That's why this automation is considered compliant to level 0.

The safety workshops were carried out on the 2 and 17 October, 2023, using two licensed ATCOs working at Remote Tower Center Sundsvall, endorsed for Sundsvall and Örnsköldsvik. The resulting hazard log is transcripted and available under the online transcript database section 9.

The feedback from controllers is that the application of the method is difficult. The main reasons are that (1) the right words must be found to express expectation and (2) difficulty identifying automation. An example system was presented, "Automated Weather Observing System (AWOS)," which might be easier to identify as automation.

Supportive material such as video recordings are considered supportive but require an excellent setting using a field of view that supports the ego-centric perspective.

5.3 Digital Tower Assistance (RESKILL/DITA) (Level 5 automation)

DITA is a rule-based assistance that supports the tower controller in a multi-remote tower work environment (as described in the previous section) by planning and monitoring services for arriving traffic at one airport, considering simple traffic situations only.

The operator acknowledges the provision of control services for an arriving movement right before entry into the Approach Sector (TMA). From there, the operator can set DITA by a plan that provides control services from entry into TMA to landing on the runway at the destination airport. This comprises issuing clearances such as "descend to ", "direct to," and even a landing clearance for a specific runway.

The safety workshop was conducted on 18 December 2023, with

- Two operational experts who participated in the experimental data collection study as air traffic controllers
- Two design experts from LFV and LiU
- A human Factors expert from LFV

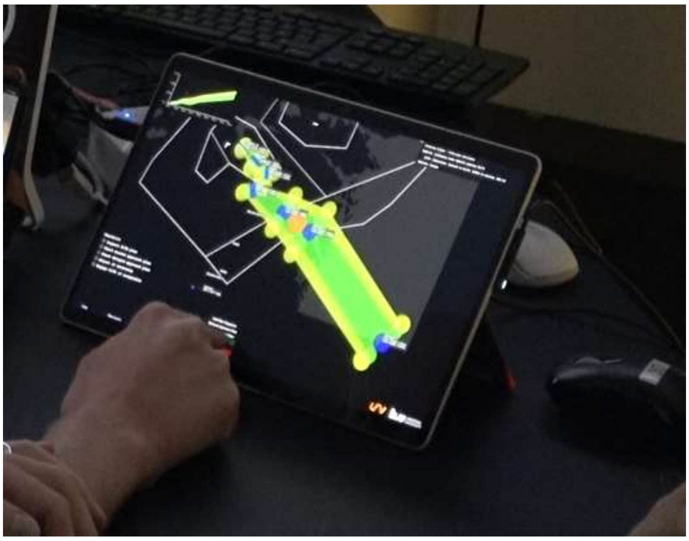

Figure 10: Digital Tower Assistant (RESKILL/DITA) in a Multi Remote Tower Setup.

Before the workshop, there was a live demonstration of DITA in the simulated tower environment, where participating operational experts gained experience. The operational experts stated that this was beneficial for reminding and refreshing the experience made under the experimental data collection in November '21. This contributed to a higher awareness of all automated functions and situations of actual application in the simulated operational scenario.

The resulting hazard log is transcripted and available under the online transcript database section 9. The conclusion from this workshop was very positive concerning the capability of identifying non-functional discrepancies as hazards from the guidelines applied. Many hazards could be identified utilizing this approach during this one day workshop. Collecting and analyzing would have been even more exhaustive with more time.

6 CONCLUSION AND OUTLOOK

The aim of the study was to lay the groundwork for a safety assessment that would support the prospective identification of AS-related risks. This that aligns with future expectations for implementing significant innovations in LFV's area of responsibility. The primarly requires a appropriate understanding of the AS phenomenon. As defined, AS describes situations with a given gap between actual automation behaviour and operators expectations. This is a potent safety risk/hazard, especially when the mismatch is not or late detected by the operator. To assess these risks, the project systematically attempted to integrate it into prevailing safety-I models and processes.

The project began with a literature review on automation surprise, exploring different perspectives of explanations and established assessment methods. The review analyzed state-of-the-art and pioneering (early generation) safety assessment methods to evaluate their effectiveness in integrating the AS phenomenon into their assessment processes. The conclusion it's up to the assessor to describe the hazard properly by studying phenomenons and their safety-relevant effect that matches the hypothetic threatening event. Existing safety assessment methods, which define a generic hazard event only as part of the charin of events, but do not support the assessor in the characteristics of the AS occurrence. There's no established method to model events resulting from human-automation interaction in Safety-I models like event trees or Reasons “Swiss cheese”-model. Exceptionally, a pioneering method emerged called "STPA," (see section 2.2) which integrates the human-automation interaction loop into the assumed safety models first. The findings are provided in the article Meyer et al 2022 at the ICAS conference in Stockholm 2022.

In conjunction with the literature review, evidence supporting the existence of AS was gathered through empiric data, such as interviews, a survey study, and an incident database analysis. Each of these activities focused on distinct areas where AS might arise and could be observed, contributing to a comprehensive analysis and characterization of the phenomenon.

The interviews were conducted to collect experiences and opinions from aviation and safety experts, which might not be easily obtained from scientific publications alone. The responses painted a relatively consistent picture of the phenomenon, converging on the statement: "AS is a conflict of strategies between the operator and automation." Paradoxically, AS events could occur for similar reasons when surprised by a colleague (rather than automation) in situations involving close collaboration and coordination (e.g., captain and first officer on the flight deck). From this perspective, automation can be compared to a human to a colleague exhibiting unusual or unexpected working patterns. Another key insight was that AS is considered a "discrepancy in expectation." there may not be a surprise if this discrepancy goes unnoticed by the operator. Many discrepancies probably never manifest significantly, as the situation may have enough margins to tolerate their effects. Moreover, this questions the subjective frequency of AS events, as it is possible that the majority goes unnoticed. This qualifies the discrepancy of expectation as a latent condition for failures.

The survey study aimed to inquire about ATCOs' experiences with AS events while working with the Thales Top Sky system (Level 0 automation – information acquisition and analysis) at ATCC Malmö. Based on Boer and Hurts 2017 [12], the survey included 20 questions about the latest personal observations of surprises, considering prevailing conditions such as the specific system involved and related traffic conditions. The results indicated that surprises exist even with Level 0 automation. However, the surveyed events corresponded to "small" surprises, considered insignificant regarding safety-relevant severity. This suggests a high degree of containment capability by ATCOs, demonstrating robustness against discrepancies in expectation through early detection and reporting.

The incident/error report database study identified surprises related to poor or confusing interface designs, leading to misunderstandings. These observations were reported and improved to prevent similar occurrences in the future. This further indicates that a potential for AS exists but is well-controlled and contained. According to the database study, a crucial safety barrier is that automation does not act autonomously but supports ATCOs in understanding and decision-making.

Combining the literature review with the above mentioned studies provides a comprehensive understanding of the AS phenomenon. Based on the available data and analysis results, requirements could be specified for a method capable of identifying risks involving automation surprise.

In the search for a solution, an alternative term, "Discrepancy of Expectation," was proposed, which is strongly inspired by the findings gathered during the interview study and the definition of AS given by Palmer:

> *"An Automation Surprise occurs when the automation behaves in a manner that is different from what the operator is expecting"* [1]

This concept reorients the AS phenomenon by focusing on the "operator's expectation" as a central element, which ties in with the research framework based on Endsley's "situational awareness" model. The "situation awareness"-context gives automation surprise a more mature conceptual framework, which is considered a stable foundation from which further integration into Safety-I models appears easier. From then on, the focus was on returning to the basic idea of "expectations of automation" and looking for explanations of how discrepancies can arise from an expectation. Based on the new context, the AS itself was given a new role as an indicator used by the operator to detect discrepancies. Such an indicator has the function of assisting the operator in error detection and correction, as described by Reason [32]. This may be considered a very relevant safety barrier that allows the operator to correct errors and lapses in situational awareness. The integration into Safety-I could then succeed by applying Reasons' "Trajectory of accident opportunity" (also refered as Swiss Cheese Model) and setting up these two new model events as an "unsafe act" and "safety barrier". AS was finally operationalized as model events in the Swiss cheese model and the integration succeeded.

Upon shifting towards an "expectation-centered" approach, it became more straightforward to elaborate on why an operator's expectations might deviate from the actual automation behavior. As "situation awareness" became a central element, the spotlight went from expectation to "critical cues". These cues are fundamental to the operator's mental model, covering the environment, traffic situation, and automation functionality. The risk of false expectations arises from incomplete or corrupted understanding due to the mental model's erroneous "corrupted" critical cues. There are two reasons for corrupted critical cues. One reason might be the time dynamics they exhibit and the resulting halftime of validity of the mental model, which can be only as accurate as the cues it relies on. This forces the operator to constantly synchronize or update the information visually or acoustically in awareness of the variability of the cue and its halftime. It is assumed that the operator finds himself in a permanent synchronization process for finding a trade-off between the accuracy of the cues and the efficiency of keeping track of several tasks and activities active at a time. This is challenging as attention is a limited resource that the operator has to share between tasks and activities to maintain a sufficiently accurate mental model for decision-making. The risk of incomplete or corrupted cues is unavoidable because of the limitations the attention resource underlies. The state of automation or the related mode of operation's false expectation of automation behaviour may result from the mentioned limitation. The other reason for corrupted critical cues is unverified assumptions of the automation work principles that affect the operator's expectation of how information is processed and assembled considering the conditions and parameters.

The development of the guideline is based on the theory of corrupted critical cues as a primary reason for expectation discrepancies, offering a step-by-step approach for assessors to identify automation, related functions, and associated risks arising from such discrepancies.

The practical application and evaluation of the example of three cases did presume much knowledge and methodological skills. This is why the safety workshops were conducted and led by safety experts with experience in safety assessment methods. The practical deliverables of the safety workshops were three hazard logs. The safety workshops' feedback concerning the guideline's feasibility draws a mixed picture from "successful" to "difficult" to apply. Paradoxically, the cases with the higher levels of automation showed more ease in the approach than the level 0 automation of multi-remote towers. In the latter, operational experts found it difficult to find the automation, how to apply the theory, and find something that can be considered an "expectation". The explanation could be simple but plausible. The multi-remote tower was based on automation that operational experts relied on for many years and became a familiar part of daily work and work life in the control center. Most systems of the example case have been operating for years, although in single remote tower mode and not multi-mode. The situation is then that operators may not question these systems regarding their expectations as they have become perfectly familiar with them for many years. Operators may become unaware of all expectations and information cues assumed deductively reasoned; the participants might have applied the guideline more quickly if the system had been as unfamiliar as it was right before the operational launch. This fact suggests the expert status of the participants whose mindset has an implicit knowledge base. The explanation for the "difficulties" with the multi-remote tower was, therefore, the same reason for the "ease" observed in the AI-supported decision-making and DITA cases, with the small but significant difference that these cases benefited from the novelty of its cases.

The case's contradiction illustrates the effects of implicit versus explicit knowledge. Another alternative explanation is based on the assumption that level 0 automation makes expectations harder to identify, and related to this are the critical cues. In favor of this assumption, it might be that expectations of information acquisition and analysis are less tangible but more subtle. Displays or devices providing information are not considered automation for many operational experts; instead, it is taken for granted. Nevertheless, the suspicion is that too much prior knowledge is required to ensure practical implementation of the guideline. Therefore, the difficulties in identifying expectations in level 0 automation should be subject to further efforts to evaluate and improve the guidelines.

The project's overall positive conclusion provides an understanding of automation surprise and a framework successfully integrating it into Safety-I model assumptions. A safety assessment guideline focused on automation surprise was developed, tested in three cases, and involved numerous safety and operational experts. One of the central findings is that the theory developed here is not the challenging part, rather it’s the practical application that sets the limit for an identification of AS risks in future.

Recommendations include refining the guideline to assist assessors and operational experts in identifying expectations that may lead to discrepancies in automation behavior. The project's abstract nature and theoretical background may require simplification for broader accessibility. Efforts should also focus on making implicit knowledge more explicit, considering that operators quickly familiarize themselves with systems. Additionally, the guideline should be evaluated and improved for its practical implementation, especially in level 0 automation scenarios where identifying expectations is challenging. Further, human-in-the-loop simulation or Retrospective Think Aloud Technique should extend empirical support to any safety workshop by making the participating operational aware of the details of human-machine interfacing. The guideline should be improved using more cases, such as from the project HAIKU, which aims to design the ATC work position using AI-support in the future.

7 VIDEO PUBLICATION FINAL DISSEMINATION EVENT ON 10 NOVEMBER 2023 IN ATCC MALMÖ

[09:30 – 10:00] Presentation of the Safety Assessment 2030 (SB2030) project, including objectives, problem statement, background on the Automation Surprise (AS) phenomenon, project structure, and deliverables. Speakers: Lothar Meyer and Christian Bjursten.

Finaldissemination_SB2030_2023-11-10_part1.mp4

[10:00 - 10:45] "Automation and Situation Awareness in Complex Socio-Technical Systems" with Sebastian Pannasch, Professor at Technische Universität Dresden.

Finaldissemination_SB2030_2023-11-10_part2.mp4

[11:00 – 11:25] "Automation Surprise in Enroute - Presentation of a Survey Study at ATCC Malmö" by Matthis Fritz, Master of Science at Technische Universität Dresden.

Finaldissemination_SB2030_2023-11-10_part3.mp4

[11:25 – 12:00] "A New Approach to Safety Assessment and Addressing Automation Surprise" with speakers Lothar Meyer and Christian Bjursten.

Finaldissemination_SB2030_2023-11-10_part4.mp4

8 CONFERENCE PUBLICATIONS

There are two conference publications available, authored in the scope of this project:

L. Meyer, C. B. Carlsson, Å. Svensson, M. Peukert, L. Danielson, and B. Josefsson, "Stressing safety assessment methods by higher levels of automation," in 33rd Congress of the International Council of the Aeronautical Sciences (ICAS2022), 2022.

Available under ICAS online archive:

https://www.icas.org/ICAS_ARCHIVE/ICAS2022/data/preview/ICAS2022_0903.htm

M. Fritz, L. Meyer, M. Peukert, E. Martinsen, “Investigating Surprise in Operations using Air Traffic Controller’s Experience” in 35th European Association of Aviation Psychology (EAAP) Conference, 23-26 Sep 2024, Athens GR, submitted.

9 ONLINE DATABASE

- All transcripts of interviews
- Hazard logs resulting from safety workshops
- Presentation slides of the final dissemination on the 10th of November
- Questionnaires Sheet from the survey study

Documents are available as a download.

The database is accessible to the public under the URL:

https://web02.droponline.se/shares/folder/LNMYdoS2xPR/

(the URL may change in Q1 2024 due to cancelled contract between LFV and droponline provider. A new host needs to be found for the online database and the URL to be updated)

REFERENCES

[1] Everett Palmer. "Oops, it didn't arm case study of two automation surprises". In: Proceedings of the Eighth International Symposium on Aviation Psychology. Ohio State University Columbus, Ohio. 1995, pp. 227–232

[2] SESAR3. Joint Undertaking, European ATM master plan – Digitalising Europe’s aviation infrastructure – Executive view – 2020 edition.Publications Office, 2020. doi: doi/10.2829/695700

[3] R. Parasuraman, T. B. Sheridan, and C. D. Wickens, “A model for types and levels of human interaction with automation,” IEEE transactions on systems, man, and cybernetics. Part A, Systems and humans : a publication of the IEEE Systems, Man, and Cybernetics Society, vol. 30 3, pp. 286–97, 2000. [Online]. Available: https://api.semanticscholar.org/CorpusID:8664574

[4] The Malaysian ICAO Annex 13 Safety Investigation Team for MH, “Safety Investigation Report - Malaysia Airlines Boeing B777-200ER (9M-MRO) 08 March 2014,” Jul. 2018.

[5] J. Lundberg, “Situation awareness systems, states and processes: A holistic framework,” Theoretical Issues in Ergonomics Science, vol. 16, no. 5, pp. 447–473, 2015.

[6] L. Meyer, M. Peukert, B. Josefsson, and J. Lundberg, “Validation of an empiric method for safety assessment of multi remote tower,” in 13th USA/Europe Air Traffic Management Research and Development Seminar (ATM2019), 2019

[7] E. Hollnagel, R. L. Wears, and J. Braithwaite, “From safety-i to safety-ii: A white paper,” The resilient health care net: published simultaneously by the University of Southern Denmark, University of Florida, USA, and Macquarie University, Australia, 2015.

[8] Eurocontrol, Systemic Occurrence Analysis Methodology (SOAM) Quick Reference Guide issued by Eurocontrol (2015), Brussels, BE.

[9] L. Sherry and R. Mauro, “Controlled flight into stall (cfis): Functional complexity failures and automation surprises,” in 2014 Integrated Communications, Navigation and Surveillance Conference (ICNS) Conference Proceedings, 2014, pp. D1-1-D1–11. doi: 10.1109/ICNSurv.2014.6819980.

[10] Frederic Dehais et al. ""Automation Surprise" in Aviation: Real-Time Solutions". In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. CHI '15. Seoul, Republic of Korea: Association for Computing Machinery, 2015, pp. 2525–2534. ISBN: 9781450331456. DOI: 10.1145/2702123.2702521. URL: https://doi.org/10.1145/2702123.2702521.

[11] Amy Rankin, Rogier Woltjer, and Joris Field. "Sensemaking following surprise in the cockpit—a re-framing problem In: Cognition, Technology & Work 18.4 (2016), pp. 623–642.

[12] Robert de Boer and Karel Hurts. "Automation Surprise: Results of a Field Survey of Dutch Pilots". In: Aviation Psychology and Applied Human Factors 7 (Apr. 2017), pp. 28–41. DOI: 10.1027/2192-0923/a000113.

[13] Robert De Boer and Sidney Dekker. “Models of Automation Surprise: Results of a Field Survey in Aviation”. In: Safety 3.3 (2017). ISSN: 2313-576X. DOI: 10.3390/safety3030020. URL: https://www.mdpi.com/2313-576X/3/3/20.

[14] Raja Parasuraman and Dietrich H. Manzey. "Complacency and Bias in Human Use of Automation: An Attentional Integration". In: Human Factors 52.3 (2010). PMID: 21077562, pp. 381–410. DOI: 10.1177/0018720810376055. eprint: https://doi.org/10.1177/0018720810376055

[15] Mika Endsley. "Situation Awareness in Aviation Systems". In: Handbook of aviation human factors: Second edition. Ed. by J.A. Wise, V.D. Hopkin, and D.J. Garland. 2nd. CRC Press, 2016. Chap. 12, pp. 268–289.

[16] Nadine B. Sarter and David D. Woods. "Team Play with a Powerful and Independent Agent: Operational Experiences and Automation Surprises on the Airbus A-320". In: Human Factors 39.4 (1997). PMID: 11536850, pp. 553–569. DOI: 10.1518/001872097778667997. eprint: https://doi.org/10.1518/001872097778667997.

[17] Nadine B Sarter, David D Woods, Charles E Billings, et al. “Automation Surprises”. In: Handbook of human factors and ergonomics 2 (1997), pp. 1926–1943.

[18] David D Woods and Nadine B Sarter. “Learning from automation surprises and going sour accidents”. In: Cognitive engineering in the aviation domain (2000), pp. 327–353.

[19] Sidney Dekker. The field guide to human error investigations. Routledge, 2017.

[20] Marvin Rausand and Stein Haugen. Risk assessment: theory, methods, and applications. John Wiley & Sons, Ltd, 2020. ISBN: 9781118281116. DOI: https : / / doi . org / 10 . 1002 /9781118281116. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118281116. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118281116.

[21] M.H.C. Everdij and H.A.P. Blom. Safety Methods Database. Version 1.2. Available at. Netherlands Aerospace Centre NLR, 2020. URL: http://www.nlr.nl/documents/flyers/SATdb.pdf.

[22] B. Kirwan. A Guide to Practical Human Reliability Assessment (). 1st ed. CRC Press, 1994. URL: https://doi.org/10.1201/9781315136349.

[23] Thomas Fylking and Christian Bjursten Carlsson. Skyddsbarriäranalys (LOPA): Vägledning för val av numeriska data. 1st ed. IPS, 2016. URL: http://libris.kb.se/bib/19704582.

[24] Eurocontrol. Safety Assessment Methodology - A framework of methods and techniques to develop safety assessments of changes to functional systems. Version 2.1. Eurocontrol. 2006. URL: https://www.eurocontrol.int/tool/safety-assessment-methodology.

[25] Nancy G Leveson. Engineering a safer world: Systems thinking applied to safety. 2016. URL: http://library.oapen.org/handle/20.500.12657/26043.

[26] Leonid Sedov et al. “Qualitative and Quantitative Risk Assessment of Urban Airspace Operations”. In: 11th SESAR Innovation Days, SESAR Joint Undertaking. SESAR Joint Undertaking, 2021.

[27] European Commission. “Commission Implementing Regulation (EU) 2017/373 of 1 March 2017 laying down common requirements for providers of air traffic management/air navigation

services and other air traffic management network functions and their oversight, repealing Regulation (EC) No 482/2008, Implementing Regulations (EU) No 1034/2011, (EU) No 1035/2011 and (EU) 2016/1377 and amending Regulation (EU) No 677/2011 (Text with EEA relevance.)" In: Official Journal of the European Union (2017).

[28] Erik Hollnagel. FRAM: The Functional Resonance Analysis Method: Modelling Complex Socio-Technical Systems. Oct. 2012. ISBN: 978-1-4094-4551-7. DOI: 10.1201/9781315255071.

[29] H. W. Heinrich et al., "Industrial accident prevention. a scientific approach.," Industrial Accident Prevention. A Scientific Approach., no. Second Edition, 1941.

[30] M. Fritz, Exploring Automation Surprise and Involving Factors Among Air Traffic Control. Dresden, Germany: Technische Universität Dresden.

[31] EUROCONTROL Safety Assessment Methodology Task Force, "EATMP Safety Assessment Methodology" (2006), Brussels, BE

[32] Reason, James. Human error. Cambridge university press, 1990.

[33] L .Meyer, "Eye Tracking for Risk Assessment, " Technical Report, Trafikverket, 2022.

[34] L. Meyer, K. J. Klang, S. Boonsong, et al., "Mapping the decision-making process of conflict detection and resolution in en-route control: An eye-tracking based approach," in Proceedings of 12th SESAR Innovation Days, SESAR JU, Budapest, Hungary, 2022

[35] H. H. De Jong, H. A. Blom, and S. H. Stroeve, "How to identify unimaginable hazards," Proc. 25th ISSC, Baltimore, Maryland, 2007.