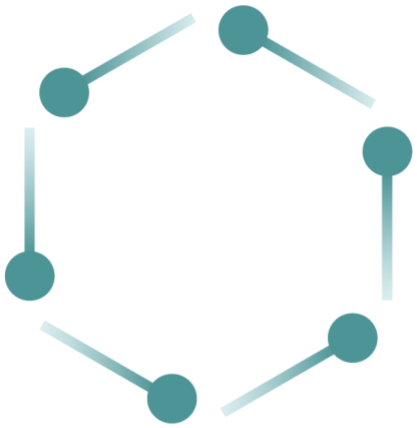

D2.2 – Human-Machine Collaboration in Operations and Performance Monitoring

Document number	D2.2
Document title	Human-Machine Collaboration in Operations and Performance Monitoring
Version	1.0
Work package	WP2
Edition date	30.06.2025
Responsible unit	ONERA
Dissemination level	PU
Project acronym	SafeTeam
Grant	101069877
Call	Safe, Resilient Transport and Smart Mobility services for passengers and goods (HORIZON-CL5-2021-D6-01)
Topic	HORIZON-CL5-2021-D6-01-13: Safe automation and human factors in aviation – intelligent integration and assistance

This project has been funded by the European Union under Grant Agreement 101069877

SafeTeam Consortium

	Innaxis (INX)
	Agencia Estatal de Seguridad Aérea (AESA)
	Technische Universität München (TUM)
	DataBeacon
	ONERA
	Rise Research Institutes of Sweden AB (RISE)
	PEGASUS HAVA TASIMACILIGI ANONIM SIRKETI (PEGASUS)
	CAA INTERNATIONAL LIMITED (CAAi)

Document change record

Version	Date	Status
0.1	28/05/2025	Initial draft
1.0	30/06/2025	Final submission

Abstract

The SafeTeam project aims to investigate the power of digital assistants and how new technique can improve safety in the aviation domain and incorporating human factors theory to ensure that safety measures are followed throughout the implementation process. In particular, SafeTeam project has developed a framework with the purpose of helping individuals who lack expertise in human factors to consider such aspects and improve human-autonomy collaboration. Part of this framework considers how to quantify the impact of the digital assistant proposed on critical human dimensions.

The work presented in this deliverable describes the process of selecting relevant metrics for practitioners interested in quantifying and characterizing the impact of the introduction of artificial agents on human operators. To guide this choice, this deliverable proposes a selection of metrics for each dimension of interest. It also specifies, for each metric, the expertise required to administer this metric, the time required to administer and/or compute this metric, the material necessary to administer this metric and the moment of the moment when it should be administered. Commentaries and limitations regarding each metric are added to help evaluators with their choice.

Table of Contents

Abstract *3*

1 Introduction *5*

1.1 Background 5

1.2 Theoretical foundations 5

1.3 Outline 6

2 Process Description *8*

2.1 Case description 8

3 Metrics Selection Process *10*

3.1 Evaluation methodology: Experimental approach 10

3.2 Metric selection: Required properties 10

3.3 Categories of measurements 12

3.4 Data collection methods 12

3.5 Selection Criteria and Dimensions of Interest 18

4 Selected metrics *24*

4.1 Team performance 24

4.2 Mental Models 26

4.3 Trust 28

4.4 Fluency 31

4.5 Agency/Controllability 32

4.6 Communication and coordination 33

4.7 Explanation satisfaction 35

4.8 Vigilance, Attention allocation 36

4.9 Workload and cognitive load 37

4.10 Usability and acceptability 38

5 Conclusions *40*

6 References *41*

1 Introduction

1.1 Background

The aviation industry is undergoing a significant transformation, with the integration of intelligent digital assistants and increasing levels of automation becoming central to future operational concepts. While these innovations offer significant potential to enhance safety, efficiency, and resilience, they also introduce new challenges, particularly concerning human performance, trust, and human-machine collaboration (Halawi, Miller, & Holley, 2024; Kirwan, 2023). A key issue in this context is the lack of accessible human factors (HF) evaluation methods and criteria that can be reliably applied to assess and guide the integration of these technologies, especially by stakeholders who may not be HF experts.

This gap is particularly problematic given that human error remains a predominant contributor to aviation incidents and accidents (Mathavara & Ramachandran, 2022). As automation becomes more capable, there is a risk that poorly designed interactions can exacerbate human performance issues such as complacency, workload mismanagement, and the well-documented Out-Of-The-Loop (OOTL) phenomenon (Endsley & Kiris, 1995; Gawron, 2019). Without clear guidance or usable evaluation frameworks, designers, operators, and regulators alike may struggle to ensure that digital assistants truly support—rather than undermine—the human operators they are meant to help.

The SafeTeam project—and specifically this work package (WP2)—aims to address this challenge. In Deliverable 2.1 (D2.1), a simplified design guide was introduced to support the integration of human factors into system design, even for non-experts. D2.1 presented a structured, iterative approach for modelling human-autonomy teaming (HAT) and incorporating human factors from the early design stages. This report, Deliverable 2.2 (D2.2), builds on that foundation by focusing on how to evaluate human performance and human-machine collaboration in the context of these systems.

The objective of this report is to present an accessible and practical framework for defining and selecting metrics and procedures to access the human side of digital assistant integration. The goal is to enable a wide range of stakeholders—including designers, developers, assessors, and end users—to evaluate how automation affects human performance and to ensure safe and effective collaboration between humans and machines. This framework is particularly designed to be applicable across different operational case studies, thereby contributing to the generalisability and scalability of the SafeTeam approach across the aviation domain.

1.2 Theoretical foundations

The SafeTeam framework is grounded in core concepts from HF, Cognitive Systems Engineering (CSE), and Human-Autonomy Interaction (HAI), with a particular focus on supporting non-expert users in understanding and evaluating the impact of automation. At its core is the recognition that automation in complex socio-technical systems, such as aviation, must be designed not only for technical performance but also for human compatibility, situation awareness, and cognitive resilience.

Human Factors and Ergonomics (HF/E) aim to optimize the interaction between people and systems by designing technologies, tasks, and environments that align with human capabilities and limitations (Salvendy, 2012). Within aviation, HF is critical to ensure that digital assistants enhance—rather than compromise—performance, safety, and trust (Kirwan, 2023). As automation becomes more autonomous and less transparent, the need to actively design for explainability, control, and mutual understanding becomes paramount (Hoffman et al., 2023).

Human-Autonomy Teaming (HAT) expands traditional Human-Autonomy Interaction (HAI) by framing automation as a cooperative agent rather than a tool. This perspective highlights the importance of shared goals, mutual observability and directability, and adaptive coordination strategies. In high-stakes environments like flight operations or air traffic control, effective teaming between human and automated or autonomous agents must consider trust calibration, workload management, and the preservation of operator agency.

To address these challenges, the SafeTeam approach introduces a structured, iterative methodology that draws on several theoretical foundations:

- Levels of Automation (LOA), such as the Sheridan and Verplank scale (Sheridan & Verplank, 1978), which help classify the degree of autonomy in system functions and inform appropriate task allocation between human and machine.
- Function allocation principles, particularly the updated HABA-MABA (Humans Are Better At / Machines Are Better At) approach, which encourages evaluating tasks based on strengths and limitations of each party in dynamic, context-dependent ways (Dekker & Woods, 2002).
- OOTL performance problem, which describes the cognitive risks associated with operators being disengaged from the system due to excessive or opaque automation (Endsley & Kiris, 1995).
- Transparency and situation awareness frameworks, which emphasize the importance of understanding system intent and state for maintaining effective human performance (Endsley, 1996).

These foundations were operationalized in D2.1, where a process was developed for non-experts to assess human-autonomy collaboration in the design phase. D2.2 extends this by focusing on the evaluation phase; specifically, how to select and use metrics to monitor performance, diagnose risks like OOTL, and support iterative improvement. To keep this report focused and accessible, a full recapitulation of the theoretical underpinnings of the design framework (and the models and processes of the design framework itself) is avoided here, but readers are encouraged to refer to D2.1 for a detailed exposition.

1.3 Outline

This document provides a practical guide for evaluating human-machine collaboration in systems that incorporate intelligent digital assistants. Building on the HF design principles established in D2.1, this report focuses on how to assess and monitor human performance, particularly in relation to collaboration, safety, and resilience.

The structure of the document is as follows:

- **Section 2: Process Description**

  Before evaluation can begin, the system under assessment must be clearly defined. This includes understanding who the end users are, how the system is intended to be used, what problems it aims to solve, and what the potential implications are for safety, efficiency, and user satisfaction. For those who followed the design framework in D2.1, much of this information will already be documented. For others, this section outlines the minimum information required to proceed with a meaningful evaluation.

- **Section 3: Metrics Selection Process**

  This section guides practitioners through the process of selecting appropriate evaluation metrics. It includes practical considerations such as feasibility, constraints, and the influence of system characteristics (e.g., level of autonomy, type of interaction). Various parameters influencing the choice of metrics and selection criteria are presented to drive the practitioners to ask the right questions and consider the essential aspect for their study when considering Human-system interactions assessment.

- **Section 4: Selected metrics**

  This section lists many metrics available in the literature to evaluate human-system interaction in the dimensions addressed by the SafeTeam framework. Metrics are grouped into categories and suggestions are provided for combining methods through triangulation. Guidance is also offered on data collection techniques, (e.g., questionnaires, eye-tracking, observation) and on how to match evaluation methods to the specifics of the case at hand. All the assessment parameters presented so far (practitioner's expertise, time requirement, material necessary, etc.) are considered and associated to each metric individually in order to provide a synthesis for practitioners to support their evaluation and select the appropriate metrics.

- **Appendices**

  To keep the main report concise and accessible, examples and more detailed illustrations of the process are included in the appendices. These include adapted SafeTeam case studies and visual tools such as flowcharts or decision aids.

Each section is intended to be usable on its own, but together they provide a complete guide; from scoping the evaluation to selecting and applying metrics in operational or simulated settings.

2 Process Description

2.1 Case description

To effectively evaluate human-machine collaboration, it is important to first define the system under assessment in sufficient detail. Evaluation is only meaningful when it is grounded in a clear understanding of the system's purpose, how it is used, and by whom. This section outlines the minimum information needed to support metric selection and performance evaluation.

Core Elements for Evaluation

Practitioners should begin by clarifying the following aspects of their system:

- **End Users**

  Who are the people interacting with the system? What are their roles, responsibilities, training levels, and relevant constraints? Understanding the user population is important for choosing appropriate metrics and interpreting results. See ISO (2019) for further reading.

- **Use Context**

  In what operational setting is the system deployed or intended to be deployed? What are the environmental, procedural, and organizational conditions that frame its use (e.g., cockpit operations, ATC tower, training simulator)? Are there critical moments of use (e.g., high workload, emergency scenarios)? See ISO (2019) for further reading.

- **Envisioned Human-System Interaction**

  How does the system support or collaborate with the user? What is the level of automation, the degree of user control or supervision, and the nature of feedback and information exchange? This includes understanding whether the system operates in real time, offer recommendations, or autonomously takes action. See Stanton et al. (2013) and Kirwan and Ainsworth (1992) for relevant methods.

- **Design Goals and Intended Benefits**

  What problems or gaps is the system designed to address? Are the goals primarily safety-related (e.g., reducing error), performance-related (e.g., increasing throughput), or experiential (e.g., improving trust or user satisfaction)?

- **Potential Implications for Safety, Efficiency, Satisfaction**

  What are the anticipated risks or benefits of introducing the system? This includes changes in workload, situation awareness, human error, communication patterns, or training needs. These anticipated effects will help focus the evaluation on what matters most. See Stanton et al. (2013) and Kirwan and Ainsworth (1992) for a selection of methods.

Starting Point: With or Without D2.1

For readers who have designed their system using the SafeTeam design framework from D2.1, much of this information will already be available. The output from D2.1, such as the Hierarchical Task Analysis (HTA), agent roles, and identified design considerations, can be directly reused as input to the evaluation process.

For those entering at the evaluation stage without having followed D2.1, we strongly recommend documenting the above points before selecting metrics. This documentation does not need to be exhaustive, but it should provide a shared understanding across stakeholders (designers, evaluators, end users) of what the system is (structural), what it is supposed to do (functional), and what constitutes success.

Summary: Minimum Required Information

At a minimum, the following elements should be available before metric selection begins:

- A description of the system and its operational context
- A definition of the user population and use scenarios
- A high-level model of the human-system interaction (e.g., tasks, control flow, feedback loops)
- A list of targeted outcomes or concerns (e.g., reduced workload, increased situation awareness or trust)
- Known constraints (e.g., time, access to users, technical limitations)

This shared understanding serves as the foundation for identifying what to measure, why, and how.

3 Metrics Selection Process

3.1 Evaluation methodology: Experimental approach

The aim of the methodology is to explore the impact of proposed design choices on human-system cooperation. This implies being able to quantify this impact rigorously. However, this quantification comes up against a major challenge: the variability of human behavior. This behavioral variability results from the fact that each behavior depends on different characteristics:

- Characteristics specific to the situation,
- Characteristics specific to the subject (history, expertise, biological characteristics).

The experimental method is traditionally used to put the human operator in a controlled situation and to analyze his activity through different levels of analysis (subjective, behavioral, or physiological). This approach is based on a probabilistic principle which, with reference to a database, makes it possible to account for the probability that certain events, under certain conditions, may be correlated. This gives rise to a principle in which an attempt is made to generalize from a small sample (Campbell & Stanley, 1963; Millsap & Maydeu-Olivares, 2009).

Experimentation therefore consists of creating (or isolating) sources of variation and checking that the variations induced are statistically significant compared with simple chance. These sources of variation that we are trying to test are what we call independent variables, i.e. variables that we are going to manipulate voluntarily and that are supposed to have an impact on the phenomenon to be measured. The induced variations that we are seeking to measure are dependent variables. These dependent variables are selected to best represent the factors we wish to assess (e.g., cognitive load, fatigue, performance, etc.). These two elements (situations in which the subjects are placed and measurements) constitute an experimental paradigm (Campbell & Stanley, 1963; Millsap & Maydeu-Olivares, 2009).

In order to ensure that the effects observed are not the result of chance or of factors other than those being manipulated, the experimental approach ensures (1) that the conditions being compared differ only in the variable being manipulated, and (2) that the participants carry out a sufficient number of repetitions. Finally, to control the influence of all the other variables likely to influence behaviour (fatigue, learning), it is important to account for the order of the trials and the counterbalancing between conditions: this is known as the experimental design (Campbell & Stanley, 1963; Millsap & Maydeu-Olivares, 2009).

All the choices made (independent variables, dependent variables, experimental paradigm, experimental design) constitute the *experimental protocol*. The main objective of this document is to define the dependent variables, i.e. the measures, that will be collected to characterise the impact of the proposed manipulations on human AI cooperation.

3.2 Metric selection: Required properties

Several criteria exist and have been proposed as guidelines for selecting and developing techniques (Wierwille and Eggemeier 1993; NASA 2014; Matthews et al. 2015; Longo 2015; Heard, Harriott, and Adams 2018).

The first set of criteria establishes most directly the validity of the measure. That is, the measure should reflect changes in the dimension measured (sensitivity) but no change in other constructs that

are not directly reflective to this dimension (selectivity). Moreover, the measure should be able to distinguish the sources responsible for the change (diagnosticity). While each single measure will not be able to respond to all these criteria, it is rather the combination of measures that should fulfil these properties. As there is no universal ideal combination of measure for all tasks and all contexts, it is necessary to define for which level the measure is relevant (range) and whether the measure is dedicated to specific tasks or can be used in different tasks (transferability).

- **Sensitivity**: Capacity to detect changes in the cognitive state investigated. The sensitivity of the measure describes the extent to which the measure changes when this cognitive state changes.
- **Selectivity or reliability:** Reflects the sensitivity only to differences in the cognitive state explored, not changes in other variables. It describes how consistently the measure will change when and only when this cognitive state changes.
- **Diagnosticity**: Capacity to differentiate distinct sources resulting in change of the cognitive state explored.
- **Bandwidth or range:** Relates to the region where the metric is reliable for the cognitive state explored. For example, when considering workload, range corresponds to the level of workload (underload, fitting load, overload) within which the measure reliably reflects workload changes.
- **Transferability:** Consistent assessment of the cognitive state explored both within and across tests that allow it to be used in different applications.
- **Temporal resolution**: Relates to the time window necessary to compute the metric. The smaller the global required time, the better the temporal resolution.

The second set of factors is essential for ecological contexts and focuses on elements that may lead to turn down specific measurements. It emphasizes possible interference with the required task (intrusiveness), but also unsuitable technical requirements associated with the measurement technology, and the question of operator acceptance:

- **Intrusiveness**: Lack of interference with task performance; moreover, the measure should not be a source of change in the dimension observed, or in any other dimension.
- **Implementation requirements**: Practical constraints associated with instrumentation, software, and training.
- **Acceptability**: Operator perception of the validity and usefulness of the procedure.

If sensitivity is always a major consideration in the choice of a measurement technique, other criteria may be particularly relevant, especially when the experiment is not a laboratory task. Notably, intrusiveness, implementation requirements and operator acceptability appear as a first concern in SafeTeam methodology as the methodology is intended to be used by non-expert practitioners in ecological environment. For example, intrusiveness of the measuring system (the hardware for physiological measures as well as the frequency and length of interruptions for subjective ones) may lead to the Hawthorne effect – or a halo effect of social desirability (Krumpal, 2013) – rendering the results untrue in a real-life scenario. Decreasing the intrusiveness of the experimental set-up would make it less likely to observe such behavior. Acceptability may be of great significance in a real operational environment, while the implementation requirements (and also the complexity of the analysis) can clearly have an impact on the ability of non-HF experts to use this method.

Based on these various constraints, a number of useful metrics will be proposed in the remainder of this document.

3.3 Categories of measurements

Evaluation data can be collected from multiple perspectives or viewpoints to enable broad and rich insights into the qualities and integration effects of a product. These perspectives include:

- **Task performance:** Quantifiable measures such as time-on-task, error rates, number of actions, and end states, gathered through manual recording or system logs.
- **Participant self-assessments:** Subjective ratings collected via validated or customised scales capturing constructs like workload, situation awareness, perceived performance or difficulty, and usability.
- **Behavioural observation:** External assessment of participant behaviours during interaction, conducted in real-time by observers or retrospectively via video analysis.

Orthogonal to these perspectives is the distinction between *objective* and *subjective* data types. Though often positioned as opposites, in practice they serve complementary roles in evaluation. Objective measures aim to capture observable, quantifiable phenomena, such as the time taken to complete a task, the number of actions performed, or error rates. These data are valued for their repeatability and reduced susceptibility to personal biases. However, they can overlook important internal experiences like perceived effort, frustration, and strategic decision making that are not easily observable (Jahedi & Méndez, 2014).

Subjective measures, on the other hand, captures the participant's internal state or personal evaluation, for example through self-reported workload, situation awareness, or satisfaction. Although subjective ratings are inherently influenced by individual perceptions, emotions, and memory biases, they provide insights into the user experience that objective measures alone cannot reveal (Muckler & Seven, 1992).

Importantly, the distinction is not always clear-cut; an observation by a third party (like an expert rating a user's performance) may be considered "objective" in form but still carries interpretive subjectivity. Likewise, physiological measures (e.g., heart rate variability, eye-tracking, or EEG) are technically objective but may require subjective interpretation regarding their meaning in context (Muckler & Seven, 1992).

Combining these multiple sources and perspectives enable triangulation of findings, offering a more comprehensive and reliable evaluation (Denzin, 2017). While objective performance metrics provide hard data on outcomes and efficiencies, subjective assessments and behavioural observations add important context about user experience, strategy, and affective states (Jahedi & Méndez, 2014; Muckler & Seven, 1992). By integrating performance data, self-assessment data, and behavioural observation data (across both objective and subjective dimensions) evaluators can mitigate the limitations inherent in any single data source, such as self-report biases, observer interpretive variance, and measurement artifacts, and gain a holistic and nuanced view of human-system interaction.

3.4 Data collection methods

Having identified the contextual constraints and categories of measurement, the next step is to choose suitable methods for retrieving the relevant data. Each method corresponds to different types of information, e.g., objective vs. subjective, first-person vs. third-person perspectives, and offer distinct strengths and limitations. The selection of methods should balance practical feasibility (as discussed in sections 3.1 and 3.3) with the need for validity, repeatability, and interpretability (as discussed in section 3.2).

Where possible, we recommend a triangulation approach: using multiple methods to measure the same construct, or collecting data from different stakeholder perspectives (e.g., end-user, system, observer) (Denzin, 2017). This not only increases robustness but also helps reconcile discrepancies between subjective experience and observed or recorded behaviour.

Below, we review key categories of data collection methods and describe how they can be applied in the context of HAT evaluation.

3.4.1 Physiological and biometric measurements

Psychophysiology attempts to interpret psychological processes through their effect on the body state. Among the many advantages of this approach, physiological measures produce continuous and objective measurements of the operator state (Lohani et al., 2019).

On a general note, the ever-evolving technologies have allowed to record more and more easily operators' state in everyday-life-like or ecological settings. The aeronautics domain is not spared by this evolution and is actually a lead actor in ecological psychophysiological evaluation of the state of operators. The necessity to move from lab tasks to ecological setting is essential, and was already mentioned more than 20 years ago. On the one hand, the Hawthorne effect (Mayo 2010) has proven to be detrimental as it tends to improve participants' performances and motivation. On the other hand, the various stimulation but also environmental noise and variable experimental conditions tend to lower performance, as well as our ability to record exploitable data. Many factors such as task difficulty, engagement, perceptual stimulation and attention variation are very different in lab tasks compared to real experiments. Nevertheless, the miniaturization of systems and improvement of data-processing algorithms has allowed to record psychophysiology data out of the lab. A quick search on "real flight psychophysiological measures" in Google Scholar reveals more than 29000 publications with more than half of them (17300) in the last 10 years. This motivation is not limited to MWL evaluation in aeronautics but also to emotion evaluation (Healey et al. 2010), auditory perception (Debener et al. 2012), inattentional deafness (Somon et al. 2022), and general cognition (see Lohani et al., 2019 for a review) for various applications. Nowadays, physiological markers are used as real-time indicators of users' cognitive and emotional states, such as mental workload, situation awareness, arousal, and stress. Common techniques include:

- **Eye-tracking**

  Tracks gaze direction, number of fixation and their duration, eye blink rate and blink duration, saccades, and pupil diameter. These metrics are useful for assessing visual attention distribution, interface usability, and situation awareness (e.g., failure to fixate on critical displays), and trust (Hergeth et al., 2016).

- **Electroencephalography (EEG)**

  Measures electrical activity in the brain via scalp electrodes. This can be useful for inferring cognitive load, fatigue levels, or task engagement or vigilance (Berka et al., 2007).

- **Heart rate and Heart Rate Variability (HRV)**

  Captured through elecrocardiography (ECG) or wearable sensors. HRV is often used as a proxy for mental effort or stress, where low HRV suggests higher workload or stress. This can be useful for comparing baseline to task performance periods (Lohani et al., 2019).

- **Respiratory activity**

  Duration of inspiration (Ti), duration of expiration (Te), total cycle time (Ttot), and tidal volume (VT) that represents the volume that is displaced by one breath. Stress and mental effort are associated with an increase in respiratory rate and a decrease in respiratory volume (NATO 2004) whereas the respiratory volume is generally increased when the demands are very high (Harding 1987).

- **Functional Near-InfraRed Spectroscopy (fNIRS)**

  Measures the hemodynamic brain activity with optical near-infrared light to estimate the cortical variations of oxy- and deoxyhemoglobin (respectively HbO2 and HbR) concentration in response to neuronal activity (Izzetoglu et al. 2005). As illustration, studies have demonstrated an increase of HbO2 concentration under high MWL conditions, as well as increased oxygenation (as computed with the [HbO2]/[HbR] ratio and [HbO2] versus [HbR] difference) in the prefrontal cortex (Mandrick et al. 2016).

- **Galvanic Skin Response (GSR) / Electrodermal Activity (EDA)**

  Reflects changes in skin conductance due to sweating. This is typically correlated with elevated arousal or stress (Boucsein, 2012).

- **Infrared thermography (IRT)**

  Detect the distribution of the temperature of a scene through infrared radiations analysis. At the human physiological level, IRT is used on facial imagery by assessing the emitted electromagnetic radiations reflected by the surface of the skin. Correlated to change in emotions (Dzedzickis, Kaklauskas, and Bucinskas 2020; Zenju et al. 2004), alertness (Sakamoto et al. 2006), arousal (Diaz-Piedra, Gomez-Milan, and Di Stasi 2019) and mental workload (Mizuno, Mito, and Itakura 2020).

However, physiological measurements come with important limitations. First, they typically require specialized equipment and calibration. Second, they may be obtrusive or impractical in operational environments. Finally, as mentioned in section 3.3, interpretation can be non-specific, e.g., increased arousal could mean stress, engagement, or surprise. Therefore, physiological measurements are rarely useful or insightful on their own but benefit from triangulation to contextualize interpretation.

3.4.2 Behavioural observation and system interaction logging

This class of methods captures what users *do* during interaction with a system, typically from an external perspective. It includes real-time observation and retrospective analysis of recorded behaviour (Kuniavsky, 2003; Rubin et al., 2008). Common methods include:

- **Direct observation (in person or via video stream)**

  Human observers document user actions, hesitations, interactions, and behaviours using structured coding schemes, free notes, or active participation. This is valuable for understanding task strategies, identifying usability issues, and assessing communication in team settings (Cooke, 1994; Howitt, 2013; Jordan & Henderson, 1995; Wixon et al., 1990).

- **Simulation or real-world video review**

  Sessions can be reviewed by evaluators or users themselves, enabling debrief and retrospective insights and the discovery of latent issues not apparent (or possible to examine) in real-time (e.g., Alhadreti & Mayhew, 2018; Mahatody et al., 2010).

- **System and interaction logs**

  Automatically capture time-stamped data on task performance (e.g., task duration, error rates, system states), use of automation (e.g., frequency, overrides, timing), or response to system prompts or failures (e.g., Hilbert & Redmiles, 2000).

General limitations of these family of methods include the need for trained coders to ensure consistency in live and video observation and note-taking. Real-time observation may miss subtle behaviours, while video allows replays and cross-checking. System logs provide objective data but lack context or reasoning.

3.4.3 Self-report instruments

Self-report methods offer a direct way to access the internal, subjective experience of users; what they felt, thought, or perceived during or after interaction with the system. While such data are inherently introspective and potentially biased, they are essential for assessing constructs what are otherwise difficult to observe or measure objectively, such as trust, usability, perceived workload, satisfaction, or clarity in system behaviour (Muckler & Seven, 1992; Jahedi & Méndez, 2014). Key approaches include:

- **Structured questionnaires**

  Structured questionnaires provide a predefined set of items and response options to elicit user input. These can be administered immediately after a task, at the end of a session, or periodically over time (e.g., in longitudinal evaluations). They are efficient, scalable, and allow for quantitative comparison across participants or systems.

  - **Standardized tools**

    These are validated instruments developed and tested across multiple studies or domains. They allow for benchmarking and statistical comparisons, often using fixed formats like Likert scales (Cooke, 1994). Examples include the NASA Task Load Index (NASA-TLX) (Hart, 2006; Hart & Staveland, 1988) to measure perceived workload (over dimensions like mental demand, physical demand, or frustration) and the System Usability Scale (SUS) (Lewis, 2018) to provide a quick and general measure of perceived usability. It could be unidimensional ratings, hierarchical ratings or multidimensional ratings.

    Standardized tools are especially useful in mature projects or comparative evaluations, and benefit from established scoring and interpretation guidelines. Rating scales are the most practical and generally applicable measures.

  - **Custom questionnaires**

    When no standard instrument fits the evaluation needs, tailored questionnaires can be developed. These are particularly useful for capturing system-specific concerns (e.g., perceived helpfulness of a flight assistant), following up on observed issues or

themes raised in interviews, or eliciting opinions on new features or changes during iterative prototyping.

Custom instruments should still be designed with care; using clearly worded items, balancing phrasing (to avoid acquiescence bias), and appropriate response scales (Lietz, 2010; Oppenheim, 2000; Sauro & Lewis, 2011).

- **Free-text feedback and open-ended questions**

  These allow users to explain their experience in their own words to highlight contextual or emotional nuance (e.g., "it felt like the assistant took control at the wrong time"), emerging issues not anticipated by designers, and potential mismatches between system behaviour and user expectations. While free-text responses are more difficult to quantify, they are valuable in exploratory or early-stage evaluations and can point to areas needing redesign (Cooke, 1994).

- **Interviews**

  Interviews provide a richer, dialogic approach to understanding user experience. They are typically used after simulation runs or system use and can be conducted in person, remotely, or during post-session debriefs.

  Structured interviews follow a fixed set of questions and are ideal for formal studies requiring comparison across users. *Semi-structured* interviews begin with a guide of key topics but allow flexibility to explore unexpected areas of concern or user insight. *Unstructured* interviews are conversational and exploratory, suitable during formative research or usability testing (Cooke, 1994; Howitt, 2013).

  Interviews are especially useful for uncovering mental models, misunderstandings, or value judgments users may not express in structured forms.

- **Focus groups**

  Focus groups gather multiple stakeholders (e.g., pilots, controllers, instructors) to discuss their experiences, concerns, or expectations. These sessions are useful for identifying areas of consensus or disagreement, exploring social and organizational dynamics (e.g., training acceptance, trust issues), and validating early design directions or evaluation criteria (Howitt, 2013; Krueger & Casey, 2015).

  Group dynamics can stimulate new ideas, but moderation is important to ensure balanced participation and avoid groupthink or dominance effects (Howitt, 2013).

The use of self-report methods warrants several considerations. For instance, timing the administration of self-report instruments is important; immediate post-task responses reduce recall bias. Interpretation of self-reported data also benefits from triangulation, where subjective data are strengthened when supported by behavioural evidence or expert analysis. Another potential issue is fatigue, where over-surveying can impact data quality. Self-report methods should be used sparingly or combined with short-form tools where needed.

Subjective measures have practical advantages (ease of implementation, overall non-intrusiveness) and are considered the easiest method for cognitive state measurement. However, several drawbacks still remain. The first main criticism of this type of measures is that they are subject to both subjective distortion and social desirability bias (Fürstenau & Radüntz, 2022; Radüntz, 2017), leading to possibly biased values. Second, they have generally poor temporal resolution. While increasing the sampling

rate for operator feedback via the questionnaire could lead to improved temporal resolution, this may actually have the negative impact to interrupt the task being performed. In addition, most subjective measures tend to suffer from memory lapses as the measure is not made during the event but after. Finally, it must be noted that these subjective measures generally have high between-rater variability due to their possible emphasizing of personal subjective biases.

Considerations for all self-report methods overviewed in this section include selecting the appropriate timing of administration (e.g., immediately post-task to minimize recall bias), limiting survey fatigue through concise instruments, ensuring consistent interpretation of scale anchors across participants, and triangulating subjective responses with behavioural or physiological data to improve interpretability and robustness.

3.4.4 Expert and peer-based evaluation

Expert and peer-based evaluations rely on the informed judgment of individuals with domain knowledge, either in human factors, systems engineering, operations, or training. These methods are particularly valuable when objective metrics are difficult to define, or when qualitative interpretation of performance is needed. They also provide useful input in early-stage design (e.g., identifying potential issues) or as part of a validation exercise where external judgment is required (e.g., certification, safety case preparation) (Klein et al., 2004; Mahatody et al., 2010; Nielsen & Molich, 1990). Common approaches include:

- **Heuristic evaluations**

  Heuristic evaluation involves expert reviewers systematically inspecting a user interface or system workflow against established human factors or usability principles (Nielsen & Molich, 1990). For example, "Does the system provide clear and timely feedback?", "Can the user understand and control automation behaviour?", or "Is the interface consistent and predictable?".

  This method is low-cost and rapid, particularly useful during prototyping or for highlighting known usability traps (e.g., hidden states, unclear handovers). Common heuristic sets include usability heuristics (general interface design) (Nielsen & Molich, 1990), aviation-specific checklists (e.g., for cockpit HMI layout), or HAI principles like transparency, observability, controllability.

- **Cognitive Walkthroughs and Structured Expert Analyses**

  This class of methods involves expert evaluators systematically stepping through specific user tasks or scenarios to assess whether a system supports effective and intuitive interaction. Unlike heuristic evaluations, which focus on general interface principles, cognitive walkthroughs and task analyses emphasize task-specific reasoning by simulating the user's problem-solving process at each step. Experts examine whether users will know what to do, whether they can do it, and whether they can interpret the system's response (Cooke, 1994: Mahatody et al., 2010).

  Cognitive walkthroughs are particularly well-suited for early prototypes and for identifying issues in learnability, action sequencing, and feedback clarity. They typically require more preparation, including defined user goals and action sequences, and benefit from multidisciplinary teams familiar with the domain and the user population (Mahatody et al., 2010).

Other structured expert methods, such as GOMS (Goals, Operators, Methods, and Selection rules) (John & Kieras, 1996) analysis or pluralistic walkthroughs (Hollingsed & Novick, 2007), extend this approach by introducing predictive models of user behaviour or collaborative multi-stakeholder evaluations. These methods are resource-intensive but offer rich diagnostic value, particularly when empirical user testing is constrained.

- **Instructor or peer assessment**

  Often used in training or operational settings (e.g., Evidence-Based Training), instructors or peers assess a participant's performance using defined competency frameworks (e.g., Crew Resource Management, workload management, decision making), behavioural markers or grading rubrics, or annotated video or live observation tools (see section 3.4.2) (O'Connor et al., 2008: Salas et al., 1999). This method is valuable for both formative feedback and summative evaluation. In HAT contexts, peer video can also surface concerns about trust, clarity of roles, or perceived fairness of automation decisions.

General considerations for this family of methods include the importance of calibration, since multiple raters must agree on criteria and interpretations to avoid inconsistency. Therefore, these methods work best when combined with structured observation or system logs for triangulation. Since these methods are subject to interpersonal or organizational biases, data anonymization can help ensure validity of the results.

3.5 Selection Criteria and Dimensions of Interest

In order to select the metrics of interest, we first need to identify the phenomena and states we wish to assess. SafeTeam project proposes to draw on existing literature to identify these dimensions of interest, particularly regarding the constraints and difficulties generated by the introduction of more or less autonomous virtual assistants.

3.5.1 OOTL associated issues

When a new automation solution is introduced into a system, or when there is an increase in the autonomy of automated systems, developers often assume that adding "automation" is a simple substitution of a machine activity for human activity (substitution myth, see Woods & Tinapple, 1999). However, the fascination regarding the possibilities afforded by technology often obscures the fact that automation also produced new loads and difficulties for the humans responsible for operating, troubleshooting, and managing high-consequence systems. Whatever the merits of any automation technology are, automation does not merely supplant human activity but also transforms the nature of human work.

Empirical data on the relationship of people and technology suggest that traditional automation has many negative performance and safety consequences associated with it stemming from the human out-of-the-loop (OOTL) performance problem (see Endsley & Kiris, 1995; Kaber & Endsley, 1997). Particularly, automation is frequently accompanied by a decrease in operator performance, such as a reduced sensitivity to important signals (Billings, 1991; Wiener, 1988), excessive or insufficient trust in system ability (Parasuraman et al., 1993), and loss of operator situation awareness (Carmody & Gluckman, 1993; Endsley, 1996; Endsley & Kiris, 1995). As a major consequence, the OOTL performance problem leaves operators of automated systems unable to take over manual operations in the case of automation failure. Particularly, the OOTL performance problem causes a set of difficulties including a longer latency to determine what has failed, to decide if an intervention is necessary and to find the adequate course of action (Billings, 1991).

Different issues are associated to this OOTL phenomenon:

- **Situation Awareness issue**

The lack of operator involvement in supervisory modes and passive information processing contributes to critical human cognitive errors, specifically the loss of operator situation awareness (SA), to which many safety incidents have been attributed. Particularly, OOTL phenomenon is characterized by both a failure to detect and to understand the problem and by difficulties to find appropriate solutions.

Failure to detect – It is now clear that humans are less aware of changes in the environmental or system state when those changes are under the control of another agent (automation or human; Endsley & Kiris, 1995; Parasuraman & Riley, 1997; Wickens, 1994; Metzger & Parasuraman, 2001). Several works indicate a lack of operator awareness of automation failures and a decrease in detection of critical system state changes when involving in automation supervision (for a review see Endsley & Kiris, 1995), and numerous incidents have been attributed to these difficulties in perception when operating in an automated mode.

Failure to understand - In addition to delays in detecting that a problem has occurred necessitating intervention, operators may meet difficulties to develop sufficient understanding of the situation and to overcome the problem. "Automation surprises" are a direct instantiation of these difficulties in automation understanding and take-over situations (see Sarter, Woods & Billings, 1997). Automation surprise is said to occur when the automation behaves in a manner different than its operator expects (see Palmer, 1999). When interacting with automated systems, human operators will develop a mental model of the system's behavior and use it to anticipate how the machine will behave in the near future. However, with increase in system complexity (for example, the multiplication of the number of possible "modes"), it is sometimes difficult for the human operator to track the activities of their automated partners. The result can be situations where the operator is surprised by the behavior of the automation asking questions like, what is it doing now, why did it do that, or what is it going to do next (Wiener, 1989). These "automation surprises" are particularly well documented (e.g., Degani & Heymann, 2000; Palmer, 1995; Sarter & Woods, 1994, 1995; Moll van Charante et al., 1993) and have been listed as one of the major causes of incidents (see for example FAA, 1995).

- **Vigilance issue**

When the inclusion of automation appears critical to increase safety and efficiency, such highly automated environments will require maintaining high levels of vigilance for a long period of time. Interestingly, research on vigilance has shown that humans are poorly suited for monitoring roles (Davies & Parasuraman, 1982; Parasuraman, 1987; Wiener, 1987). Indeed, we observe a decrease of human operator vigilance in case of interaction with highly automated systems (see for example O'Hanlon, 1981; Wiener, 1988; Strauch, 2002). Nowadays, there is some consensus that vigilance decrement is one the major index of OOTL phenomenon and insufficient monitoring and checking of automated functions is one important behavioral aspect of the OOTL performance problem, (i.e. information on the status of the automated functions is sampled less often than necessary) (see for example Billings, 1991; Kaber & Endsely, 1997).

Interestingly, alteration of vigilance process could also generate, or increase, out of the loop situation, particularly Mind Wandering (MW) episodes.MW is the human mind's propensity to generate thoughts unrelated to the task at hand (Christoff, 2012; Stawarczyk et al., 2011). Studies point MW as a possible cause of many driving accidents (Galera et al., 2012), plane crashes (Casner and Schooler, 2013) and medical errors (van Charante et al., 1993). Recently, a link between automation and MW has been proposed (Casner & Schooler, 2014; Gouraud, Delorme, & Berberian, 2018). Recently, a link between automation and MW has been proposed. Casner and Schooler (2014) conducted a study where pilots were instructed to handle the approach – flight phase before landing – in a simulator by

following beacons at altitudes given by the ATCo. Probes inquired about their state of mind at predetermined times while pilots had to report their position to the ATCo. They reported that when using higher levels of automation, pilots were more prone to MW when they had no interaction with the system and when the previous call had been made. Time saved by automation, which should normally be used to plan the flight, was instead fulfilled by task-unrelated thoughts.

- **Complacency / Trust issue**

Complacency defines the cognitive orientation toward high reliability automation, particularly prior to the first time it has failed in the user's experience (Rovira, McGarry & Parasuraman, 2007). Complacency is perceived as a strategy to optimize performances operators working with systems that fail once every ten million hours of use tend to underestimate the possibility of automation errors and over-trust the system (Parasuraman & Wickens, 2008). Because they have the feeling that the system does not require them to work efficiently, they instinctively lower cognitive resources allocated to monitoring (Morrison, Cohen, & Gluckman, 1993). This overreliance on automation represents an important aspect of misuse that can result from several form of human error, including decision biases and failure of monitoring (Wiener, 1988; Parasuraman, Molloy, & Singh, 1993; Parasuraman & Riley, 1997; Singh, Molloy, & Parasuraman, 1993).

This phenomenon is directly linked to the concept of trust. The concept of trust in automation describes to what extent an operator relies on an automatic control system. The role of trust in human-automation interaction has been the focus of much research over the past decade (e.g., Dzindolet, Peterson, Pomranky, Pierce, & Beck, 2003; Madhavan & Wiegmann, 2007; for a comprehensive review, see Lee & See, 2004). It has been proved that high levels of trust in automation that is not perfectly reliable lead to overreliance and failure to monitor the "raw" information sources providing input to automation – so-called complacency. Conversely, if automation is underestimated it may limit its use, restricting its benefits (Lee & See, 2004). In certain conditions, system opacity appears to limit the trust in automation and generate such disuse.

3.5.2 Beyond the OOTL phenomenon

This is a new relation between the human and the machine, as an automatic machine may be said to be intelligent. The new form of interaction differs dramatically from the traditional interaction of the human with the tools and devices that possess no intelligence, in which all sensing and control were done by the human operator. Adding or expanding the machine's role changes the cooperative architecture, changing the human's role, often in profound ways (Sarter, Woods, & Billings, 1997). The introduction of automation into complex systems has led to a redistribution of operational control between human operators and computerized automated systems. Moreover, as pointed out by Flemisch et al. (2012), in addition to control, authority, ability and responsibility are also modified according to the level of automation within the human–machine system.

- **Sense of control and acceptability issue**

Often neglected, the psychosocial aspects of automation may prove to be the most important of all, because they influence the basic attitudes of the operator toward his task, and we would presume, his motivation, adaptability, and responsiveness. The significance of these questions lies not in the spectra of massive unemployment due to assembly line automation, but in the effects of automation on the changing role of human operators.

Improving acceptance of new technology and systems by human operators is an important area of concern to equipment suppliers (see Horberry, Stevens, & Regan, 2014). To be acceptable, new technology must be reliable, efficient and useful. Although performance and preference are often positively correlated (Nielsen & Levy 1994), high levels of performance do not guarantee user acceptability. Further, it appears that users indeed tend to reject systems that enhance their

performance in favor of systems that are less efficient but more acceptable. For instance, Inagaki et al. (2007) showed that drivers preferred collision warnings than automated control, which tends to be misunderstood; even when automated control provided led to better performances. As pointed out by Shneiderman and Plaisant, (2004), users "strongly desire the sense that they are in charge of the system and that the system responds to their actions". Increase in automation has the potential to seriously threaten this sense of control.

Recently, the concept of agency has been applied to the HCI domain (McEneaney, 2013; Berberian et al., 2012; 2013; Obhi & Hall, 2011; Limerick, Coyle & Moore, 2014). The term 'sense of agency', or sense of control, is the subjective awareness of initiating, executing, and controlling one's own volitional actions in the world (Jeannerod, 2003). This form of self-awareness is important not only for motor control but also for social interactions, the ascription of causal responsibility and serves as a key motivational force for human behaviour. Unfortunately, it has been repeatedly shown that the progress in automation technology can alter the development of this sense of agency (Berberian, 2019). What makes our understanding of agency especially relevant is the fact that a decrease in agency could generate critical concern regarding both automation acceptability and operator behavior. As pointed out by Baron (1988), "the major human factors concern of pilots in regard to introduction of automation is that, in some circumstances, operations with such aids may leave the critical question, who is in control now, the human or the machine?". This ambiguity about who is in control could impact user acceptance, but also user engagement in the task.

- **Coordination issue**

This new form of interaction also introduces new coordination demands and the emergence of new classes of issues due to failures in the human-machine relationship. Automated tools are increasingly being modelled as 'partners' rather than as tools (Klein et al. 2004). These partners should support or assist the human in performing functions that may either be difficult or even impossible for the operator to perform without the assistance of a 'knowledgeable team-mate'. This entails new coordination demands for the operator – they must ensure that their own actions and those of the automated agent are synchronized and consistent. Designing to support this type of coordination is a post-condition of more capable, more autonomous automated systems. Critically, it appears necessary to design a system able to give feedback about its state and the course of its action to support cooperation. Unfortunately, as previously discussed with the concept of "Automation Surprise", such cooperation is difficult to obtain. The result can be automation which leaves its human partners perplexed, asking Wiener's (1989) now familiar questions: what is it doing? Why is it doing that? What is it going to do next?

These new coordination demands generate the emergence of new classes of issues due to failures in the human-machine coordination. Amongst others, system opacity appears as a first concern. one of the foundations of any type of cooperative work is a shared representation of the problem situation (e.g. Grosz, 1981; McCarthy et al., 1991). In human-human cooperative work, a common finding is that people continually work to build and maintain a "common ground" of understanding to support coordination of their problem-solving efforts (e.g. Patterson et al., 1999). We can break the concept of a shared representation into two basic (although interdependent) parts: (1) a shared representation of the problem state, and (2) representations of the activities of other agents. The first part, shared representation of the problem situation, means that the agents need to maintain a common understanding of the nature of the problem to be solved. The second part, shared representation of other agents' activities, involves access to information about what other agents are working on, which solution strategies they are pursuing, why they chose a particular strategy, the status of their efforts (e.g. are they having difficulties? Why? How long will they be occupied?), and their intentions about what to do next. When we consider automated team members, this information no longer comes for free – we have to actively design representations to generate the shared understandings which are

needed to support cooperative work. Nowadays, this work for "OPENING UP THE BLACK BOX" remains unsatisfactory.

In this context, the main problem with automation is not the presence of automation, but rather its inappropriate design (Norman, 1990). Where designers really need guidance today is how to support the coordination between people and automation, not only in foreseeable standard situations, but also during novel, unexpected circumstances. Understanding the actions of the automated system is central for human operators. However, as previously discussed with the concept of "Automation Surprise", such understanding is difficult to obtain. The lack of system predictability is certainly a central point in understanding OOTL phenomenon and associated difficulties of takeover (Christoffersen & Woods, 2000; Dekker & Woods, 2002; Klein, Woods, Bradshaw, Hoffman, & Feltovich, 2004; Norman, 1990). With the progress of technology, current man-made complex systems tend to develop cascades and runaway chains of automatic reactions that decrease, or even eliminate predictability and cause outsized and unpredicted events (Taleb, 2012). This is what we may call "system opacity": the difficulty for a human operator to have a clear idea of the system's intentions and to predict the sequence of events that will occur. In that sense, the main problem with automation is not the automation per se, but rather its inappropriate design within the human-computer interaction (Norman, 1990). For example, previous studies have showed that ATCo performance can be compromised when ATCos do not have ready access to aircraft intent information (Castaño & Parasuraman, 1999; Galster, Duley, Masalonis, & Parasuraman, 2001). This situation is likely to generate difficulties in anticipating/understanding the actions of my artificial partner, thereby generating difficulties in terms of coordination, acceptability and feeling of control. In this sense, the intelligibility of artificial systems (i.e., producing clear, predictable and understandable behavior) is a major challenge for the systems engineering community.

3.5.3 Selected dimensions

Considering the different issues revealed, we have decided to select a set of measures that make available:

- The quality of cooperation, in particular team performance, Cooperativeness / Coordination, Fluency, Shared Situation Awareness,
- The attitude toward the artificial partner, in particular, Trust, Usability, Acceptability, Controllability,
- The state of the operator regarding OOTL phenomenon, in particular Vigilance/Mind Wandering, Feeling of Agency, Situation Awareness, Complacency.

In addition to the relevance for the dimensions of interest, we have considered ease of use as a central element in our selection of metrics, as the methodology must be accessible to people who are not experts in HF.

4 Selected metrics

Based on an extensive literature review relating to studies about human-machine collaboration in operations (see Appendix A) and regarding the dimensions of interest previously identified, we have selected a set of metrics allowing to assess: Human-Machine Collaboration (HMC), Attitude towards the artificial partner (Attitude), and the Operator State (State). For each metric, various characteristics are provided:

- The name and definition of the metric;
- The dimension addressed by this metric: either HMC, Attitude or State;
- The type of metric as described in Section 3: either self-report, behavioural or physiological, but also either subjective or objective;
- The expertise required to administer this metric: on three levels namely novice, intermediate or expert;
- The time required to administer and compute this metric: on three levels namely low, medium, or high;
- The material necessary to administer this metric;
- The moment when it should be administered: either before the task, during the task, after each trial, or after the task;
- Commentaries and limitations regarding this metric.

4.1 Team performance

<table>
<tr><td colspan="4"><u>General domain addressed:</u> Team Performance
Team performance is in this case specifically related to taskwork, meaning how the human-artificial agent team performs at completing the assigned tasks.</td></tr>
<tr><td colspan="4"><u>Name:</u> Response Time
Response Time (RT) is the "time duration between the presentation of stimulus to a human and the human's response".</td></tr>
<tr><td><u>Dimension addressed:</u>

State
Attitude</td><td><u>Type of metric:</u>

Behavioral
Objective</td><td><u>Expertise required:</u>
Novice</td><td><u>Time requirements:</u>

Low</td></tr>
<tr><td><u>Material necessary:</u>

Time-accurate response recording system.</td><td colspan="3" rowspan="2"><u>Commentary/limitations:</u>

It requires a specific action from the operator, associated to an event.
It can be interpreted only in time-pressured evaluations.
It can be computed at the trial level or averaged separately for experimental conditions to be compared.</td></tr>
<tr><td><u>When:</u>
After each trial</td></tr>
</table>

Name: **Total task time**

Total Task Time (TTT) refers to the total elapsed time between the beginning of a task (a specific work, a trial, etc.) and its end. It includes all activities (breaks, interruptions, etc.) within that time frame. It has to be dissociated from the time on Task (ToT) which refers to the active part of the TTT.

Dimension addressed:	Type of metric:	Expertise required:	Time requirements:
HMC Attitude	Behavioral Objective	Novice	Low

Material necessary:	Commentary/limitations:
Timer	It requires to define what the "task" is. It needs start and stop elements which are associated to a specific meaning and requirement.
When: After the task	

Name: **Accuracy [Hit/Error rate]**

The accuracy is defined in signal detection tasks as "the proportion of trials in which a signal is present and the participant correctly responds that it is." It is defined in %. The hit rate [error rate] is computed as the ratio between the number of correctly identified [missed] events over the total number of presented events * 100

Dimension addressed:	Type of metric:	Expertise required:	Time requirements:
State	Behavioral Objective	Novice	Low

Material necessary:	Commentary/limitations:
None	This measure refers to a wide number of concepts and has low specificity towards them. The conclusions drawn from it can be limited. It can be computed to compare experimental conditions. It also requires the dichotomic identification of a correct and incorrect answer. When combined with other measures in the signal detection theory framework it allows to compute additional metrics such as the sensitivity (*d'*) or bias (*c*) towards a specific response.
When: After the task	

Name: **Task Completion Rate**

Task completion rate corresponds to the ratio between the number of completed trials and the total number of trials * 100.

Dimension addressed:	Type of metric:	Expertise required:	Time requirements:
Attitude	Behavioral Objective	Novice	Low

Material necessary:	Commentary/limitations:

	None When: After the task	This measure refers to a wide number of concepts and has low specificity towards them. The conclusions drawn from it can be limited. The task completion rate also requires defining thresholds satisfactory for success and failure according to the task and context.

4.2 Mental Models

General domain addressed: Mental Models

The term "Mental models" refers to an individual's internal representation or understanding of how something works or how different elements relate to each other within a system. It's a cognitive framework that people construct to help them interpret and interact with the world around them.

Name: Automation Awareness

Automation awareness refers to the quality of the representation that an individual has of its artificial partner mental model. It is measured after each trial using six statements, each rated with a 5-point Likert scale of strongly agree to strongly disagree.

Dimension addressed:	Type of metric:	Expertise required:	Time requirements:
State	Self-report Subjective	Novice	Medium

Material necessary: Paper/pencil	Commentary/limitations: An average value of automation awareness can be obtained by averaging the values of all items.
When: After each trial	Questions need to be adapted regarding the action performed by the Artificial Agent (AA) in each use case.

Name: Mental model Formal Framework

This metric's objective is to define a framework to detain a set of questions relevant to a specific system, supported by 3 definitions regarding: *subject overlap, compatibility*, and *agreement*.

A model M is a mental model that is shared to the extent θ *by agents* A_1 *and* A_2 *with respect to a set of questions Q iff there is a mental model* M_1 *of* A_1 *and* M_2 *of* A_2, *both with respect to Q, such that:*

$$1. SO(M, M_1, Q) = 1, and SO(M, M_2, Q) = 1$$

$$2. A(M, M_1, Q) \geq \theta, and A(M, M_2, Q) \geq \theta$$

Dimension addressed:	Type of metric:	Expertise required:	Time requirements:
State	Self-report Subjective	Expert	High

	Material necessary: Paper/pencil When: After each task	Commentary/limitations: This doesn't directly define the questions to be proposed to participants, it requires to adapt this set of questions to the situation and tasks. It is thus highly adaptable to various scenarios. The framework also allows to define *Shared Mental Models* (SMM) in a specific situation. Still, a complete questionnaire cannot be proposed, but a methodology to detain a suiting set of questions. The model is a mental model in the mind of an agent, sharedness is defined with respect to a relevant set of questions.

	Name: Situation Awareness General Assessment Technique The SAGAT is addressed during full simulations where the experimenter freezes the environment and probes the participant either orally, via pencil/paper or on a tablet to4valuate their current perception of the situation before the simulation resumes. Queries concern either the environment or the system state and relate to Level 1 (perception), 2 (comprehension) or 3 (projection) of the SA model.			
	Dimension addressed: State	Type of metric: Probing Subjective	Expertise required: Expert	Time requirements: High
	Material necessary: Simulation environment, Paper/pencil or tablet When: After each task	Commentary/limitations: This evaluation requires subject matter experts to identify relevant elements in the environment or the system state. Scoring needs to be addressed regarding a specific time-point and situation. Scoring some queries may be provided by subject matter experts with perfect knowledge of the situation at the time of the freeze. SAGAT scores are normally expressed as percent correct for each query, based on operationally relevant tolerance bands. Many researchers have varied from this recommended approach, instead combining the scores on all SAGAT queries into a combined overall score, or into three combined scores that represent Level 1, 2, and 3 SA. This evaluation requires a highly controllable simulation.		

	Name: Task reflection For this retrospection task, participants describing their reasoning after conducting the task by i) replaying the events, ii) identifying decision points and reflecting on them; and iii) self-explaining their own understanding of the task.			
	Dimension addressed: State	Type of metric: Self-report Subjective	Expertise required: Expert	Time requirements: High
	Material necessary: Camera recording of situation or simulation	Commentary/limitations: An empirically derived expression of the content or the ebbs and flows that compose a user's mental model must contribute to the evaluation of mental model goodness (i.e., correctness, comprehensiveness, coherence, and usefulness). This evaluation needs to be performed in complementarity with other situation awareness and mental model		

	environment, audio recording system When: After each task	evaluation tasks. This evaluation requires subject matter experts to be able to address the precise events and issues related to the task performed by the participant. It is also time consuming and requires a highly controllable simulation.

4.3 Trust

General domain addressed: Trust

Trust has been associated to many definitions and models according to the domain of application. Trust influence how willing a user is to rely on a machine agent, system, or automation to perform a task, based on the user's perception of the machine's ability, integrity, and predictability. Judgement to which the user can rely on the automated system to achieve his or her goals under conditions of uncertainty.

Name: Human Computer Trust Questionnaire

This questionnaire consists of 5 constructs (perceived reliability, perceived technical competence, perceived understandability, faith, and personal attachment) with 5 corresponding items. The participants answer on a Likert scale providing their level of agreement with each item.

Dimension addressed:	Type of metric:	Expertise required:	Time requirements:
Attitude	Self-report Subjective	Intermediate	Medium

Material necessary:	Commentary/limitations:
Paper/pencil When: After each task	The average of the 25 items represents overall trust. It can be computed for experimental conditions separately to compare them. Quite long to administrate regarding other trust questionnaire.

Name: Trust scale

This scale is based on 6 items for which participants have to provide their agreement on a 5-point Likert scale going from 1. strongly disagree to 5. strongly agree. It evaluates, regarding a specific situation the current trust state.

Dimension addressed:	Type of metric:	Expertise required:	Time requirements:
Attitude	Self-report Subjective	Novice	Low

Material necessary:	Commentary/limitations:
Paper/pencil	This scale is very easy and quick to implement. It can be used to compare various versions of a similar agent. There are several modified versions of

<table>
<tr><td></td><td>When:
After each task</td><td colspan="3">this scale with, for example, 4 items for collaborative Human-Human-Agent triads questioning both Trust in the agent and Trust in the teammate.
It can be computed for experimental conditions separately to compare them.</td></tr>
<tr><td></td><td colspan="4">Name: Propensity to trust machine scale
This scale is based on 6 items for which participants have to provide their agreement on a 5-point Likert scale going from 1. strongly disagree to 5. strongly agree. It provides a general evaluation of trust as a trait for each participant.</td></tr>
<tr><td></td><td>Dimension addressed:
Attitude</td><td>Type of metric:

Self-report
Subjective</td><td>Expertise required:

Novice</td><td>Time requirements:
Low</td></tr>
<tr><td></td><td>Material necessary:
Paper/pencil

When:
Before the task</td><td colspan="3">Commentary/limitations:
This scale allows to consider inter-participants variations. Propensity to trust and trust (see trust scale) are two separated, but related, constructs.</td></tr>
<tr><td></td><td colspan="4">Name: Dynamic reporting of trust
This measure corresponds to one simple question allowing to evaluate trust towards the artificial agent on a 0 (I don't trust the artificial agent at all) to 100 (I trust the artificial agent completely) scale. It allows to measure trust either after a trial or as probing (dynamic).</td></tr>
<tr><td></td><td>Dimension addressed:
Attitude</td><td>Type of metric:

Self-report
Subjective</td><td>Expertise required:

Novice</td><td>Time requirements:
Low</td></tr>
<tr><td></td><td>Material necessary:
None

When:
After each trial or during the task</td><td colspan="3">Commentary/limitations:
This measure has been used as quick and frequent (even continuous) measure. If performed frequently, it requires an interruption of the task even if it is quick and minimal. It near real-time measure possibility allows to be used for the consideration of the temporal specificity and evolution of trust.

It can be averaged for experimental conditions separately to compare them, or interpreted over time.</td></tr>
<tr><td></td><td colspan="4">Name: XAI trust scale
The XAI Trust Scale asks users directly whether they are confident in the XAI system, whether the XAI system is predictable, reliable, efficient, and believable. It is an 8 items list to which participants are asked to respond on a 5-point Likert scale form "I agree strongly" to "I disagree strongly".</td></tr>
<tr><td></td><td>Dimension addressed:
Attitude</td><td>Type of metric:
Self-report
Subjective</td><td>Expertise required:
Novice</td><td>Time requirements:
Low</td></tr>
</table>

	Material necessary: Paper/pencil When: After each task	Commentary/limitations: This scale is focused specifically on the end-user's trust in machine-generated explanations. The scale is initially oriented towards tools but can be adapted to the evaluation of artificial agents. It can be computed for experimental conditions separately to compare systems or conditions.		
	Name: Acceptation-Compliance / rejection rate of artificial agent's suggestion This measure is computed as the percentage of the artificial agent's suggestion accepted or rejected. The compliance and agreement rate corresponds to the number of times the participant follows recommendations given by the system or positively responds to system alarms. They can be computed over a few trials, or a whole task.			
	Dimension addressed: Attitude	Type of metric: Behavioral Objective	Expertise required: Novice	Time requirements: Low
	Material necessary: None When: After each trial or after the task	Commentary/limitations: This scale can be used to assess performance as well as trust. The valence of each recommendation in terms of rejection/acceptance needs to be evaluated carefully. This measure could affect other trust metrics due to lowered trust associated with wrong recommendations detection. Additionally, this measure is very sensitive to other variables such as workload or emotions.		
	Name: Intervention/Task delegation This measure refers to the number of times the human agent intervenes [or oppositely chooses to delegate] in the artificial agent's task. In opposition with "reliance".			
	Dimension addressed: Attitude	Type of metric: Behavioral Objective	Expertise required: Novice	Time requirements: Low
	Material necessary: None When: After the task	Commentary/limitations: This measure can be used to assess performance as well as trust. Additionally, this measure is very sensitive to other variables such as workload or emotions.		
	Name: Ocular metrics Ocular metrics can be divided into two main types of measures: gaze-tracking measures and pupillometry. Both can be the physiological expression of monitoring behaviour towards the system. Specific measures such as monitoring frequency (glances), fixation frequency or fixation duration have been associated to trust.			
	Dimension addressed: Attitude	Type of metric: Physiological Objective	Expertise required: Expert	Time requirements: High

	Material necessary: Eye-tracker When: During the task	Commentary/limitations: Eye tracking measures have low specificity and must be measured in association with subjective and/or behavioural measures. Their quality and precision also have an inverse relationship with the intrusiveness of the eye tracking hardware. Additionally, eye-tracking measures are very sensitive to the luminosity of the environment, which should be stable to perform comparisons.

	Name: Decision/Verification time Decision time refers to the time to make a decision, generally complying to a recommendation. Verification time refers to the act of confirming the accuracy of a teammate's actions or recommendations and may precede compliance, reliance, or intervention. They can both be associated to trust.			
	Dimension addressed: Attitude	Type of metric: Behavioral Objective	Expertise required: Intermediate	Time requirements: Medium
	Material necessary: Timer When: After each trial	Commentary/limitations: This measure is sensitive to extraneous variable (i.e., workload and risk level) that may not capture trust. It can also be biased give the balance with the cost of said verification for the participant.		

4.4 Fluency

General domain addressed: Fluency Fluency in joint action is the quality existent when two agents perform together at high level of coordination and adaptation, in particular when they practice a task repetitively, and are well accustomed to the task and to each other. In simulation, anticipation has been shown to lead to improved task efficiency and fluency, as well as a perceived commitment of a simulated robot to the team and its contribution to the team's fluency and success.

	Name: Extended version of the fluency in Human-Robot interaction scale This scale aims to shed light on different aspects of human-robot interaction to characterize high-quality cooperation between the human and the robotic counterpart. This psychometrically-validated measurement tool allows for repeated testing and improving the understanding of HRI fluency and its perspectives.			
	Dimension addressed: HMC	Type of metric: Self-report Subjective	Expertise required: Intermediate	Time requirements: Medium
	Material necessary: Paper/pencil	Commentary/limitations: This scale can be provided in its extended, or adjusted version. The extended version allows to address several sub-dimensions of fluency		

	When: After the task	through trust, learning, shared goals, etc. These sub-dimensions can also be assessed individually. This scale can be adapted to wider autonomous agents (not only robots) but some items must be modified.

4.5 Agency/Controllability

General domain addressed: Agency/controllability

Controllability refers to "how much a user is "in control" of the process. Controllability reflects to what extent they can control the automation or alter its result to reach their goal, and how easily and rapidly can this control be carried out." Agency refers to the "experience of controlling one's own actions, and, through them, events in the outside world".

Name: Result controllability

Result controllability is measured as a 7-item scale to which participant must provide their agreement on a 7-point Likert scale from "1. Strongly disagree" to "7. Strongly agree". The items address different sub-dimensions of controllability, namely: perceived accuracy, perceived controllability, feeling of control, feeling of accomplishment, feeling of responsibility, satisfaction and enjoyment.

Dimension addressed:	Type of metric:	Expertise required:	Time requirements:
Attitude	Self-report Subjective	Novice	Low

Material necessary: Paper/pencil	Commentary/limitations: This scale is focused primarily on the result and the outcomes. Moreover, it is mostly relevant when considering an artificial agent as a tool rather than a partner. Additionally, this scale can be redundant with usability, acceptability and agent's performance. The results have to be considered individually for each item. They can be computed for experimental conditions separately to compare systems or conditions.
When: After each task	

Name: Autonomous Agent Teammate-Likeness Scale - Perceived agentic capability of the system

The perceived agentic capability scale refers to one of the 6 sub-dimensions of the Autonomous Agent Teammate-Likeness Scale. It is the perception that the intelligent agent as an autonomous agent, has some degree of decision-making latitude and an affordance, ability, and authority for self-control. It is measured as a 7-item scale to which participants provide their agreement on a 5-point Likert scale from "1. Strongly disagree" to "5. Strongly agree".

Dimension addressed:	Type of metric:	Expertise required:	Time requirements:
State	Self-report Subjective	Novice	Low

	Material necessary: Paper/pencil	Commentary/limitations: At the opposite of result controllability scale, it is mostly relevant when considering an artificial agent as a partner rather than a tool. It requires to interact with artificial agent with high level of autonomy.
	When: After the task or after each trial	

	Name: We-agency scale This measure corresponds to one simple question allowing to evaluate to which teammate (him or the artificial partner) the agent attributes the control over an outcome on a scale from 1 (definitely the artificial partner) to 9 (definitely me). It allows to measure agency either after a trial or as probing (dynamic). The question is: "Who produced the outcome of the joint action?"			
	Dimension addressed: State	Type of metric: Self-report Subjective	Expertise required: Novice	Time requirements: Low
	Material necessary: Paper/pencil	Commentary/limitations: The experimenter needs to ensure and be mindful about the use of the term responsible for, especially for populations who are required to be in constant control. The prompt can be adapted if necessary. Additionally it requires to consider a specific joint action and cannot be measured during system monitoring tasks.		
	When: After each trial			

4.6 Communication and coordination

General domain addressed: Communication and coordination Communication concerns the transmission of information, which may be by verbal (oral or written) or nonverbal means. Humans communicate to relate and exchange ideas, knowledge, feelings, and experiences and for many other interpersonal and social purposes. In the framework of human-system interaction, communication can be regarded as the process of at least two entities "sharing" something, suggesting an act of "bringing together". Coordination refers to the capacity of various parts to function together. In human-system interaction, it refers to the process of aligning the actions and interactions between humans and machines to achieve common goals effectively. The key aspects of human-machine coordination include turn-taking, communication, and the management of dependencies between activities.
Name: Human-Autonomy Teaming assessment scale This measure assesses to which extent the artificial agent supports four teaming skills namely: communication, coordination, cooperation and cognition. It is an 8-item scale to which participants have to provide their agreement on a 5-point Likert scale from "5. Strongly agree" to "1. Strongly disagree".

	Dimension addressed: HMC	Type of metric: Self-report Subjective	Expertise required: Novice	Time requirements: Medium
	Material necessary: Paper/pencil	Commentary/limitations: This measure can be redundant with controllability and usability scales.		
	When: After each task			
	Name: Autonomous Agent Teammate-Likeness Scale - Richness of communication The richness of communication scale refers to one of the 6 sub-dimensions of the Autonomous Agent Teammate-Likeness Scale. It is the perception that the intelligent agent communicates in a way that is relatively complex, sophisticated, clear, highly informative, and interactive. It is measured as a 6-item scale to which participants provide their agreement on a 5-point Likert scale from "Strongly disagree" to "Strongly agree".			
	Dimension addressed: HMC	Type of metric: Self-report Subjective	Expertise required: Novice	Time requirements: Low
	Material necessary: Paper/pencil	Commentary/limitations: Particularly relevant for consideration about the design of human system interaction. Could be both used when considering the artificial agent as a partner or a tool.		
	When: After each task			
	Name: Autonomous Agent Teammate-Likeness Scale - Synchronized mental model The synchronized mental model scale refers to one of the 6 sub-dimensions of the Autonomous Agent Teammate-Likeness Scale. It is the perception that the intelligent agent behaves in a predictable manner and responds as expected such that respective actions and reactions are synchronized, seamless, and natural. It is measured as a 5-item scale to which participants provide their agreement on a 5-point Likert scale from "1. Strongly disagree" to "5. Strongly agree".			
	Dimension addressed: HMC	Type of metric: Self-report Subjective	Expertise required: Novice	Time requirements: Low
	Material necessary: Paper/pencil	Commentary/limitations: This measure can be redundant with mental model questionnaires and assessments. It is highly relevant when performing joint task with artificial partner but remains interesting for supervisory task.		
	When: After each task			

4.7 Explanation satisfaction

<table>
<tr><td colspan="5"><u>General domain addressed:</u> Explanation satisfaction
Satisfaction is a contextualized, a posteriori judgment of explanations. It is measured according to key attributes, such as understandability, feeling of satisfaction, sufficiency of detail, completeness, usefulness, accuracy or trustworthiness.</td></tr>
<tr><td></td><td colspan="4"><u>Name:</u> Explanation Satisfaction Scale
It is measured as a 7-item scale to which participants provide their agreement on a 5-point Likert scale from "Strongly disagree" to "Strongly agree".</td></tr>
<tr><td></td><td><u>Dimension addressed:</u>
HMC</td><td><u>Type of metric:</u>
Self-report
Subjective</td><td><u>Expertise required:</u>
Novice</td><td><u>Time requirements:</u>
Low</td></tr>
<tr><td></td><td><u>Material necessary:</u>
Paper/pencil</td><td colspan="3" rowspan="2"><u>Commentary/limitations:</u>
This measure can be redundant with trust and communication scales.</td></tr>
<tr><td></td><td><u>When:</u>
After each trial</td></tr>
<tr><td></td><td colspan="4"><u>Name:</u> System Causability Scale (SCS)
SCS is measured as a 10-item scale to which participants provide their agreement on a 5-point Likert scale from "1. Strongly disagree" to "5. Strongly agree". It measures the quality of the explanations provided by the system, their timing and their granularity.</td></tr>
<tr><td></td><td><u>Dimension addressed:</u>
HMC</td><td><u>Type of metric:</u>
Self-report
Subjective</td><td><u>Expertise required:</u>
Novice</td><td><u>Time requirements:</u>
Low</td></tr>
<tr><td></td><td><u>Material necessary:</u>
Paper/pencil</td><td colspan="3" rowspan="2"><u>Commentary/limitations:</u>
An overall value is computed as the sum of the 10 items divided by 50. This measure can be redundant with shared mental models, controllability and acceptability scales. The result can be computed for experimental conditions separately to compare systems or conditions.</td></tr>
<tr><td></td><td><u>When:</u>
After the task</td></tr>
</table>

4.8 Vigilance, Attention allocation

General domain addressed: Vigilance, Attention allocation

Attention refers to a state in which cognitive resources are focused on certain aspects of the environment rather than on others and the central nervous system is in a state of readiness to respond to stimuli. Research in this field has been devoted to discerning which factors influence attention and to identify the underlying neural mechanisms. Past experience or conscious perception, as well as qualities of stimuli in the environment, such as intensity, movement, repetition, contrast, and novelty can impact attention. On the other hand, vigilance refers to a state of extreme awareness and watchfulness directed by a person toward the environment, often toward potential threats. In various contexts, vigilance tasks demand maximum physiological and psychological attention and readiness to react, characterized by an ability to attend and respond to stimulus changes for uninterrupted periods of time.

Name: Ocular metrics – Time on tool

Ocular metrics can be divided into two main types of measures: gaze-tracking measures and pupillometry. Both can be the physiological expression of monitoring behaviour towards the system. The raw eye movement is captured by the eye-tracking device with areas of interest defined beforehand. Here, the analysis of raw eye movements or scan-paths focuses on fixations (maintaining visual gaze within a specific area of the screen or regions of interest) and saccades (rapid eye movements between fixations). Fixations are defined as relatively stable positions of the eye, for at least 100ms, allowing information encoding. This requires a classification algorithm to identify fixations (and implicitly the saccades between them) from the raw scan-paths. A longer fixation time could be interpreted as a marker of the complexity or the importance of a piece of information.

Dimension addressed:	Type of metric:	Expertise required:	Time requirements:
State	Physiological Objective	Intermediate	High

Material necessary:	Commentary/limitations:
Eye-tracking system	The specificity of ocular metrics can be very limited and requires other measures (performance, subjective) to be associated with. It is very sensitive to the environment and various task parameters. Analysis could be highly time consuming.
When: During or after the task	

4.9 Workload and cognitive load

General domain addressed: Workload and cognitive load

The notion of workload is related to the cost for an operator to achieve a task in a given environment. Workload can be defined as the effort invested by the human operator into task performance; workload arises from the interaction between a particular task and the performer. Workload refers to a hypothetical construct that represents the cost incurred by a human operator to achieve a particular level of performance. Cognitive load, on the other hand refers to the effort being used in the working memory.

Name: Instantaneous Self-Assessment of workload (ISA)

ISA involves participants self--rating their workload as a function of spare mental capacity during a task (normally every two to five minutes) on a scale of 1 (low) to 5 (high). The frequency and timing of the workload ratings should be determined beforehand by the analyst. It is crucial that the provision of a workload rating is as unintrusive to the participant's primary task performance as possible.

Dimension addressed:	Type of metric:	Expertise required:	Time requirements:
State	Self-report Subjective	Novice	Low

Material necessary:	Commentary/limitations:
Paper/pencil	In order for the results to be valid, the participants should have the same understanding of each level of the workload scale i.e. what level of perceived workload constitutes a rating of 5 on the ISA workload scale and what level constitutes a rating of 1. ISA is a very simplistic technique, offering only a limited assessment of operator workload. To ensure comprehensiveness, ISA is often used in conjunction with other subjective techniques.
When: During the task	

Name: NASA-Task Load indeX (NASA-TLX)

NASA Task Load Index (TLX) method assesses workload on a 6 dimensions 7-point scales. It is a subjective, multidimensional assessment tool used to rate perceived workload in order to assess a task, system, or other aspects of performance. NASA TLX should be used at the end of the experiment / block of trials of the considered condition. Users require two separate forms. The first form is a table of definitions for their reference throughout the process (NASA-TLX Reference Sheet Definitions). The second form contains the actual survey items (NASA Task Load Index Rating Scales).

Dimension addressed:	Type of metric:	Expertise required:	Time requirements:
State	Self-report Subjective	Intermediate	High

Material necessary:	Commentary/limitations:
Paper/pencil	Two versions (weighted or unweighted) can be administered to take into account the individual participants sensitivity to the various workload sources as defined by the 6 subdimensions of the NASA-TLX. It is very important that the definitions of the 6 subdimensions are interpreted with the same meaning by all participants. In the non-weighted version, the scores from every item are summed and then the sum is divided by 6
When: After the task	

		to obtain an overall score between 0 - 100. The weighted value is computed as an overall weighted workload score for each respondent is computed by multiplying each rating (between 0 and 100) by the weight (between 0 and 5) given to the factor by that respondent. The sum of the weighting ratings for each task is then divided by 15 (the sum of the weights). The added value of this measure is the possibility to break down workload into various sources, but some dimension could be difficult to interpret.

4.10 Usability and acceptability

General domain addressed: Usability and acceptability

Acceptability refers to the extent to which a system is perceived by users as appropriate, useful, usable, trustworthy and desirable. It influences their willingness to adopt and continue using the system on long term. Several dimensions such as trust, safety and ease of use are known to be fundamental for user acceptance. On the other hand, usability refers to "the extent to which a system can be used by specified users to achieve specified goals effectively, efficiently and with satisfaction in a specified context of use." Usability has multiple components and is traditionally associated with five attributes: Learnability, Efficiency, Memorability, Errors, Satisfaction.

Name: System Usability Scale (SUS)

The SUS yields a single number representing a composite measure of the overall usability of the system being studied. It takes into account two factors of a system: its usability and its learnability. It is a 10-item scale in which participants provide their agreement on a 5-point Likert scale going from "1. Strongly disagree" to "5. Strongly agree". This scale should be used prior to any debriefing with the participant

Dimension addressed:	Type of metric:	Expertise required:	Time requirements:
Attitude	Self-report Subjective	Novice	Low

Material necessary: Paper/pencil	Commentary/limitations: All items are not presented with the same valence. To calculate the score, one should add the score of each item. For items 1, 3, 5, 7 and 9, the individual score is the grade received minus 1. For items 2, 4, 6, 8 and 10, the contribution is 5 minus the grade received. The sum of all scores is then multiplied by 2.5, and this is how the total value of SUS is obtained. After the scoring and the calculation of the score, it is possible to classify the evaluated system.
When: After the task	

Name: Acceptability scale

The acceptability scale is a tool for studying acceptance of new technological equipment. It is simple and consists of 9-item where users provide their evaluation on 5-point rating scales. These items load on two scales, a scale denoting the usefulness of the system, and a scale designating satisfaction.

	Dimension addressed: Attitude	Type of metric: Self-report Subjective	Expertise required: Novice	Time requirements: Medium
	Material necessary: Paper/pencil When: After the task	Commentary/limitations: Items 1, 2, 4, 5, 7 are scored on a scale from [+2] to [-2] whilst items 3, 6, 8 are mirrored and should be scored from [-2] to [+2]. The Usefulness scale is the average of item 1, 3, 5, 7 and 9 (so it has a range from -2 to +2) whilst the Satisfying scale is the average of items 2, 4, 6, and 8. Here, acceptance is measured by direct attitudes towards a system. Practical aspects of the system are reflected in the usefulness score, while the pleasantness is mirrored in the satisfying score.		
	Name: User acceptance of automation scale The user acceptance scale is a 7-item scale to which users provide their feedback on a 7-point Likert scale from "1. Totally disagree" to "7. Totally agree". This scale evaluates the average acceptability of a system or tool by a user. It can be adapted to both systems and artificial agents.			
	Dimension addressed: Attitude	Type of metric: Self-report Subjective	Expertise required: Novice	Time requirements: Medium
	Material necessary: Paper/pencil When: After the task	Commentary/limitations: It is adapted from a longer 26-item scale and selects specifically trust, safety and perceived ease of use related items which are fundamental to acceptance measure. This scale can share redundancies with trust, performance and usability measures.		

5 Conclusions

This deliverable provides guidance for people interested in quantifying and characterising the impact of the introduction of artificial agents on human operators. To guide this choice, this deliverable proposes a general process to select adequate metrics depending on the artificial agent evaluated but also depending on the expertise of the evaluators and the material available. This framework first defines relevant dimension of interest when considering the introduction of an artificial partner. For each dimension of interest identified, the deliverable proposes a set of metrics. The deliverable also specifies, for each metric, the expertise required to administer this metric, the time required to administer and/or compute this metric, the material necessary to administer this metric and the moment of the moment when it should be administered. Commentaries and limitations regarding each metric are added.

Not all measures are intended to be used simultaneously. For each assessment campaign, it will be necessary to identify the most relevant dimensions of interest, particularly with regard to the risks identified beforehand (see D2.1). It will then be a matter of choosing the most relevant metrics for the dimensions of interest according to the measurement resources available, the nature of the task, the nature of the interaction between the crew and the virtual assistants, the time available and the experimenters' own skills. The measures presented here therefore constitute a pool of metrics from which experimenters can pick and choose to best quantify the impact of the tool being evaluated for the use case under study.

Two limitations of the evaluation framework have been identified here. The first one is the fact that many measures (self-report, questionnaires, behavioural or physiological) reported in the literature are tested in lab-based environments where independent variables are more or less easy to identify and to set up. In these cases, biases and confounding variables are also easier to control for. In more ecological or operational contexts, tasks are more complex and individual factors variables or processes taking part into these tasks are more difficult to isolate. Thus, confounding variables can render the interpretations of variations in measures trickier. Similarly, as processes engaged in the realisation of such operational tasks are intricated, some measures or variables tend to interaction with each other. As an example, the XAI trust scale or pupillometry are two very different measures (subjective vs. objective, self-report vs. physiological) but both of them have been shown to be modulated by trust as well as workload. Showing the very high entanglement of these two concepts as variables. We have tried as much as possible to display the known correlations in the tables reporting measures. Unfortunately, unknown entanglements can still appear.

The second limitation relates to the systems considered in the evaluation framework of the SafeTeam project. Several dimensions identified in the metric selection process refer to several processes that may arise during collaboration, cooperation and interaction. These processes are often based on what we know from human-human interaction. Dimensions such as agency, trust, explainability or communication may require the system to have more intelligent or agentic abilities. Yet, the systems evaluated in SafeTeam lack these abilities. They can be considered more as Level 1 AI (as described in the EASA concept paper on AI) and are providing assistance to human operators' decision making, compared to Level 2 AI which refers to cooperative and collaborative AI. Still, several measures are either relevant for both (e.g., workload, mental models, team performance) or can be adapted to decision making tools, as well as artificial agents. This limit was also adressed whenever necessary in Tables providing measures for Human-Autonomy Teaming evaluation. In the SafeTeam evaluations though, we have had to select only those appropriate according to the type of system.

6 References

- Akash, K., McMahon, G., Reid, T., and Jain, N. (2020). Human trust-based feedback control: dynamically varying automation transparency to optimize human-machine interactions. IEEE Control Syst. Magazine 40, 98–116. doi: 10.1109/MCS.2020.3019151
- Alhadreti, O., & Mayhew, P. J. (2018). Rethinking thinking aloud: a comparison of three think-aloud protocols. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Paper no. 44. DOI: 10.1145/3173574.3173618
- Anderson, A., Dodge, J., Sadarangani, A., Juozapaitis, Z., Newman, E., Irvine, J., et al. (2020). Mental models of mere mortals with explanations of reinforcement learning. ACM Trans. Inter. Intell. Syst. 10, 1–37. doi: 10.1145/336648
- Baron, S. (1988). Pilot control. In: Wiener EL, Nagel DC, eds. Human factors in aviation. San Diego, CA: Academic Press. pp 347–386.
- Berberian, B., Sarrazin, J. C., Le Blaye, P., & Haggard, P. (2012). Automation technology and sense of control: a window on human agency. PLoS One, 7(3), e34075.
- Berberian, B., Le Blaye, P., Schulte, C., Kinani, N., & Sim, P. R. (2013). Data transmission latency and ense of control. In International conference on engineering psychology and cognitive ergonomics (pp. 3–12).
- Berberian, B. (2019). Man-Machine teaming: a problem of Agency. IFAC-PapersOnLine, 51(34), 118-123.
- Berka, C., Levendowski, D. J., Lumicao, M. N., Yau, A., Davis, G., Zivkovic, V. T., Olmstead, R. E., Tremoulet, P. D., & Craven, P. L. (2007). EEG Correlates of Task Engagement and Mental Workload in Vigilence, Learning, and Memory Tasks. Aviation, Space, and Environmental Medicine, 78(5), section II, pp. B231-B244.
- Billings, C. E. (1991). Human-Centered Aircraft Automation: A Concept and Guidelines (NASA Tech. Memo. 103885). Moffet Field, CA: NASA-Ames Research Center
- Boucsein, W. (2012). Electrodermal Activity. 2nd ed. Springer: New York, NY.
- Brooke, J. (1995). SUS: A quick and dirty usability scale. Usability Eval. Ind.. 189.
- Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), Handbook of research on teaching (pp. 171–246). Chicago: Rand McNally.
- Carmody, M. A., & Gluckman, J. P. (1993). Task specific effects of automation and automation failure on performance, workload and situational awareness. In Proceedings of the Seventh International Symposium on Aviation Psychology (Vol. 1, pp. 167-171).
- Casner, S. M., & Schooler, J. W. (2013). Thoughts in Flight Automation Use and Pilots' Task-Related and Task-Unrelated Thought. Human Factors: The Journal of the Human Factors and Ergonomics Society, 0018720813501550.
- Casner, S. M., & Schooler, J. W. (2015). Vigilance impossible: Diligence, distraction, and daydreaming all lead to failures in a practical monitoring task. Consciousness and Cognition, 35, 33–41.
- Caspar, E.A., Christensen, J.F, Cleeremans, A. and Haggard, P. (2016). Coercion Changes the Sense of Agency in the Human Brain. Current Biology. 26 (5): 585-592.
- Castano, D., & Parasuraman, R. (1999). Manipulation of pilot intent under Free Flight- A prelude to not-so-free flight. In International Symposium on Aviation Psychology, 10 th, Columbus, OH (pp. 170-176).
- Chen and G. Fragomeni (Cham: Springer), 199–213. doi: 10.1007/978-3-030-21565-1_13
- Christoff, K. (2012). Undirected thought: Neural determinants and correlates. Brain Research, 1428, 51–59.
- Christoffersen, K., & Woods, D. D. (2002). How to make automated systems team players. Advances in human performance and cognitive engineering research, 2, 1-12.
- Cooke, N. J. (1994). Varieties of knowledge elicitation techniques. International Journal of Human-Computer Studies, 41(6), 801–849. DOI: 10.1006/ijhc.1994.1083
- Davies, D. R., & Parasuraman, R. (1982). The psychology of vigilance.
- Degani, A., & Heymann, M. (2000). Some formal aspects of human automation interaction. NASA Technical Memorandum number 209600. Moffett Field, CA: NASA Ames Research Center. v
- Denzin, N. K. (2017). The Research Act: A Theoretical Introduction to Sociological Methods. Routledge. New York, NY. DOI: 10.4324/9781315134543
- Dekker, S. W., & Woods, D. D. (2002). MABA-MABA or abracadabra? Progress on human–automation co-ordination. Cognition, Technology & Work, 4(4), 240-244. DOI: 10.1007/s101110200022

- Dewey, J. A., & Carr, T. H. (2013). When dyads act in parallel, a sense of agency for the auditory consequences depends on the order of the actions. Consciousness and cognition, 22(1), 155-166.
- Diaz-Piedra, Carolina, Emilo Gomez-Milan, and Leandro L. Di Stasi. 2019. 'Nasal Skin Temperature Reveals Changes in Arousal Levels Due to Time on Task: An Experimental Thermal Infrared Imaging Study'. Applied Ergonomics 81 (November): 102870. https://doi.org/10.1016/j.apergo.2019.06.001.
- Dodge, J., Khanna, R., Irvine, J., Lam, K. H., Mai, T., Lin, Z., ... & Fern, A. (2021). After-action review for AI (AAR/AI). ACM Transactions on Interactive Intelligent Systems (TiiS), 11(3-4), 1-35.
- Dzedzickis, Andrius, Artūras Kaklauskas, and Vytautas Bucinskas. 2020. 'Human Emotion Recognition: Review of Sensors and Methods'. Sensors 20 (3): 592. https://doi.org/10.3390/s20030592.
- Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. International Journal of Human-Computer Studies, 58(6), 697-718.
- Eloy, L., Doherty, E. J., Spencer, C. A., Bobko, P., & Hirshfield, L. (2022). Using fNIRS to Identify Transparency- and Reliability-Sensitive Markers of Trust Across Multiple Timescales in Collaborative Human-Human-Agent Triads. Frontiers in Neuroergonomics, 3, 838625.
- Endsley, M. R. (1996). Automation and situation awareness. Automation and human performance: Theory and applications, 163-181.
- Endsley, M. R. (1988, May). Situation awareness global assessment technique (SAGAT). In Proceedings of the IEEE 1988 national aerospace and electronics conference (pp. 789-795). IEEE.
- Endsley, M. R. (2000). Direct measurement of situation awareness: Validity and use of SAGAT. In M. R. Endsley & D. J. Garland (Eds.), Situation awareness analysis and measurement (pp. 147–174). Mahwah, NJ: Lawrence Erlbaum
- Endsley, M. R. (2021). A systematic review and meta-analysis of direct objective measures of situation awareness: a comparison of SAGAT and SPAM. Human factors, 63(1), 124-150.
- Endsley, M.R., & Kiris, E. O. (1995). The out-of-the-loop performance problem and level of control in automation. Human Factors: The Journal of the Human Factors and Ergonomics Society 37(2), 381-394.
- FAA (1997). ATS Concept of Operations for the National Airspace System in 2005. Department of Transportation, Federal Aviation Administration. Washington, D.C.: Author.
- Flemisch, Frank, Matthias Heesen, Tobias Hesse, Johann Kelsch, Anna Schieben, and Johannes Beller. 2011. "Towards a Dynamic Balance between Humans and Automation: Authority, Ability, Responsibility and Control in Shared and Cooperative Control Situations." Cognition, Technology & Work 14 (1): 3–18.
- Fryer, D. (1939). Post quantification of introspective data. Am. J. Psychol. 52, 367–371. doi: 10.2307/141674
- Galera, C., Orriols, L., M'Bailara, K., Laborey, M., Contrand, B., Ribereau-Gayon, R., ... Lagarde, E. (2012). Mind wandering and driving: responsibility case-control study. BMJ, 345(dec13 8), e8105–e8105.
- Galster, S. M., Duley, J. A., Masalonis, A. J., & Parasuraman, R. (2001). Air traffic controller performance and workload under mature free flight: Conflict detection and resolution of aircraft self-separation. The International Journal of Aviation Psychology, 11(1), 71-93.
- Gawron, V. (2019). Nothing can go wrong: A review of automation-induced complacency research. MITRE Technical Report, no. MTR190017. MITRE Center for Advanced Aviation Systems Development. McLean, VA, USA. Available at https://www.mitre.org/sites/default/files/2021-11/pr-16-3426-lessons-lost-nothing-can-go-wrong-automation-induced-complacency.pdf
- Gouraud, J., Delorme, A., & Berberian, B. (2018). Influence of automation on mind wandering frequency in sustained attention. Consciousness and cognition, 66, 54-64.
- Grosz, B. J. (1981). Focusing and description in natural language dialogues. In: A. K. Joshi, B. L. Webber & I. A. Sag (Eds), Elements of Discourse Understanding. Cambridge, MA: Cambridge University Press
- Halawi, L., Miller, M., and Holley, S. (2024). Fostering trust in artificial intelligence in commercial aviation: an exploratory study. Issues in Information Systems, 25(2), 397-407. DOI: 10.48009/2_iis_2024_131
- Harding, Richard. 1987. Human Respiratory Responses during High Performance Flight. Edited by NATO. AGARDograph 312. Neuilly-sur-Seine: AGARD.
- Hart, S. G. & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock and N. Meshkati (Eds.) Human Mental Workload. Amsterdam: North Holland Press.
- Hart, S. G. (2006). NASA-task load index (NASA-TLX); 20 years later. In Proceedings of the human factors and ergonomics society annual meeting (Vol. 50, No. 9, pp. 904-908). Sage Publications: Los Angeles, CA.
- Heard, Jamison, Caroline E. Harriott, and Julie A. Adams. 2018. 'A Survey of Workload Assessment Algorithms'. IEEE Transactions on Human-Machine Systems 48 (5): 434–51. https://doi.org/10.1109/THMS.2017.2782483.

- Hergeth, S., Lorenz, L., Vilimek, R., Krems, J.F., 2016. Keep your scanners peeled: gaze behavior as a measure of automation trust during highly automated driving. Hum. Factors 58 (3), 509–519. DOI: 10.1177/0018720815625744
- Hilbert, D. M., & Redmiles, D. (2000). Extracting usability information from user interface events. ACM Computing Surveys (CSUR), 32(4), 384-421. DOI: 10.1145/371578.371593
- Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2023). Measures for explainable AI : Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. Frontiers in Computer Science, 5. DOI: 10.3389/fcomp.2023.1096257
- Hollingsed, T., & Novick, D. G. (2007). Usability inspection methods after 15 years of research and practice. Proceedings of the 25th annual ACM international conference on Design of communication (SIGDOC'07), 249-255. DOI: 10.1145/1297144.1297200
- Holzinger, A., Carrington, A.M., & Müller, H. (2019). Measuring the Quality of Explanations: The System Causability Scale (SCS). Kunstliche Intelligenz, 34, 193 - 198.
- Howitt, D. (2013). Introduction to Qualitative Methods in Psychology (2nd ed.). Pearson.
- Inagaki, T. (2007). Smart collaborations between humans and machines based on mutual understanding. IFAC Proceedings Volumes, 40(16), 12-22.
- ISO. (2019). ISO 9241-210:2019 Ergonomics of human-system interaction — Human-centred design for interactive systems. International Organization for Standardization
- Izzetoglu, M., K. Izzetoglu, S. Bunce, H. Ayaz, A. Devaraj, B. Onaral, and K. Pourrezaei. 2005. 'Functional Near-Infrared Neuroimaging'. IEEE Transactions on Neural Systems and Rehabilitation Engineering 13 (2): 153–59. https://doi.org/10.1109/TNSRE.2005.847377.
- Jahedi, S. & Méndez, F. (2014). On the advantages and disadvantages of subjective measures. Journal of Economic Behavior & Organization, 98, pp. 97-114. DOI: 10.1016/j.jebo.2013.12.016
- Jeannerod, M. (2003). The mechanism of self-recognition in humans. Behavioural brain research, 142(1-2), 1-15.
- John, B. E., & Kieras, D. E. (1996). Using GOMS for user interface design and evaluation: which technique?. ACM Transactions on Computer-Human Interaction, 3(4), 287-319. DOI: 10.1145/235833.236050
- Jonker, C. M., Van Riemsdijk, M. B., & Vermeulen, B. (2010, August). Shared mental models: A conceptual analysis. In International Workshop on Coordination, Organizations, Institutions, and Norms in Agent Systems (pp. 132-151). Berlin, Heidelberg: Springer Berlin Heidelberg.
- Jordan, B. & Henderson, A. (1995). Interaction Analysis: Foundations and Practice. Journal of the Learning Sciences, 4(1), 39-103. DOI: 10.1207/s15327809jls0401_2
- Kaber, D. B., & Endsley, M. R. (1997). Out-of-the-loop performance problems and the use of intermediate levels of automation for improved control system functioning and safety. Process Safety Progress, 16(3), 126-131.
- Kirwan, B. (2023). The Future Impact of Digital Assistants on Aviation Safety Culture. Human Interaction and Emerging Technologies (IHIET-AI 2023): Artificial Intelligence and Future Applications, 70, 77-87. DOI: 10.54941/ahfe1002932
- Kirwan, B., and Ainsworth, L. K. (Eds.) (1992). A Guide to Task Analysis. Taylor & Francis.
- Kirwan, B., A. Evans, L. Donohoe, A. Kilner, T. Lamoureux, T. Atkinson, and H. MacKendrick, 1997: Human factors in the ATM system design life cycle. FAA/Eurocontrol ATM R&D Seminar, Paris, France, Federal Aviation Administration and EUROCONTROL, 21 pp.
- Klein, G., & Borders, J. (2016). The ShadowBox approach to cognitive skills training: An empirical evaluation. Journal of Cognitive Engineering and Decision Making, 10(3), 268-280. Doi: 10.1177/1555343416636515
- Klein, G., Woods, D. D., Bradshaw, J. M., Hoffman, R. R., & Feltovich, P. J. (2004). Ten challenges for making automation a" team player" in joint human-agent activity. IEEE Intelligent Systems, 19(6), 91-95. DOI: 10.1109/MIS.2004.74
- Kohn, S. C., Kluck, M., and Shaw, T. H. (2020). A brief review of frequently used self-report measures of trust in automation. in Proceedings of the Human Factors and Ergonomics Society Annual Meeting. (Los Angeles, CA: SAGE Publications) 64, 1436–1440.
- Krueger, R. A. & Casey, M. A. (2015). Focus Groups: A Practical Guide for Applied Research (5th ed.). SAGE: Los Angeles, CA.
- Krumpal, I. (2013). "Determinants of social desirability bias in sensitive surveys: a literature review". Quality & Quantity. 47 (4): 2025–2047.

- Kuniavsky, M. (2003). Observing the User Experience: A Practitioner's Guide to User Research. Morgain Kaufmann: San Francisco, CA.
- Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human Factors: The Journal of the Human Factors and Ergonomics Society, 46(1), 50-80.
- Le Goff K, Rey A, Haggard P, Oullier O, Berberian B. Agency modulates interactions with automation technologies. Ergonomics. 2018 Sep;61(9):1282-1297. DOI: 10.1080/00140139.2018.1468493.
- Lewis, J. R. (2018). The System Usability Scale: Past, Present, and Future. International Journal of Human-Computer Interaction, 34(7), 577-590. DOI: 10.1080/10447318.2018.1455307
- Lietz, P. (2010). Research into questionnaire design: a summary of the literature. International Journal of Market Research, 52(2), 249-272. DOI: 10.2501/S147078530920120X
- Limerick, H., Coyle, D., and Moore, J.W. (2014). The Experience of Agency in Human-Computer Interactions: A Review. Frontiers in Human Neuroscience 8.
- Lohani, M., Payne, B. R. & Strayer, D. L. (2019). A Review of Psychophysiological Measures to Assess Cognitive States in Real-World Driving. Frontiers in Human Neuroscience, 13. DOI: 10.3389/fnhum.2019.00057
- Longo, Luca. 2015. 'A Defeasible Reasoning Framework for Human Mental Workload Representation and Assessment'. Behaviour & Information Technology 34 (8): 758–86. https://doi.org/10.1080/0144929X.2015.1015166.
- Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human–human and human–automation trust: an integrative review. Theoretical Issues in Ergonomics Science, 8(4), 277-301.
- Madsen, M., & Gregor, S. (2000, December). Measuring human-computer trust. In 11th australasian conference on information systems (Vol. 53, pp. 6-8).
- Mahatody, T., Sagar, M., & Kolski, C. (2010). State of the Art on the Cognitive Walkthrough Method, Its Variants and Evolutions. International Journal of Human-Computer Interaction, 26(8), 741-785. DOI: 10.1080/10447311003781409
- Mandrick, Kevin, Vsevolod Peysakhovich, Florence Rémy, Evelyne Lepron, and Mickaël Causse. 2016. 'Neural and Psychophysiological Correlates of Human Performance under Stress and High Mental Workload'. Biological Psychology 121 (December): 62–73. https://doi.org/10.1016/j.biopsycho.2016.10.002.
- Mathavara, K., and Ramachandran, G. (2022). Role of Human Factors in Preventing Aviation Accidents: An Insight. In Ali, Z. A. and Cvetković, D. (eds.), Aeronautics - New Advances. IntechOpen. DOI: 10.5772/intechopen.106899
- Matthews, Gerald, Lauren E. Reinerman-Jones, Daniel J. Barber, and Julian Abich. 2015. 'The Psychometrics of Mental Workload: Multiple Measures Are Sensitive but Divergent'. Human Factors: The Journal of the Human Factors and Ergonomics Society 57 (1): 125–43. https://doi.org/10.1177/0018720814539505.
- McCarthy, J. C., Miles, V. C., & Monk, A. F. (1991, March). An experimental study of common ground in text-based communication. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 209-215).
- McEneaney, J.E. (2013). Agency Effects in Human–Computer Interaction. International Journal of HumanComputer Interaction 29 (12): 798–813.
- Merritt, S. M. (2011). Affective processes in human–automation interactions. Human Factors 53, 356–370. doi: 10.1177/0018720811411912
- Metzger, U., & Parasuraman, R. (2001). The role of the air traffic controller in future air traffic management: An empirical study of active control versus passive monitoring. Human Factors: The Journal of the Human Factors and Ergonomics Society, 43(4), 519-528.
- Millsap, R. E. & Meydeu-Olivares, A. (Eds.) (2009). The SAGE Handbook of Quantitative Methods in Psychology. London: SAGE Publications
- Mizuno, T., Mito, K., & Itakura, N. (2020). Investigation of Psychological Evaluation and Estimation Method Using Skin Temperature of Lower Half of Face. In HCI International 2020-Posters: 22nd International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part I 22 (pp. 372-377). Springer International Publishing.
- Morrison, J. G., Cohen, D., & Gluckman, J. P. (1993). Prospective Principles and Guidelines for the Design of Adaptively Automated Crewstations. In R. S. Jensen & D. Neumeister (Eds.), Proceedings of the Seventh International Symposium on Aviation Psychology (pp. 172–177). Columbus, Ohio: Department of Aviation, The Ohio State University.
- Muckler, F. A. & Seven, S. A. (1992). Selecting Performance Measures: "Objective" versus "Subjective" Measurement. Human Factors, 34(4), pp. 441-455. DOI: 10.1177/001872089203400406

- NASA. 2014. 'HUMAN INTEGRATION DESIGN HANDBOOK (HIDH) BASELINE'. NASA/SP-2010-3407/REV1.
- NATO. 2004. 'Operator Functional State Assessment'. Technical Report TR-HFM-104. RTO. Neuilly-sur-Seine, France: NATO.
- Nielsen, J., & Levy, J. (1994). Measuring usability: preference vs. performance. Communications of the ACM, 37(4), 66-75.
- Nielsen, J. & Molich, R. (1990). Heuristic evaluation of user interfaces. Proceedings of the 1990 SIGCHI Conference on Human Factors in Computing Systems, 249-256. DOI: 10.1145/97243.97281
- Norman, D. A. (1990). Cognitive artifacts. Department of Cognitive Science, University of California, San Diego.
- Obhi, S.S., & Hall, P. (2011). Sense of agency and intentional binding in joint action. Experimental Brain Research, 211 (3-4), 655–662.
- O'Connor, P., Campbell, J., Newon, J., Melton, J., Salas, E., & Wilson, K. A. (2008). Crew Resource Management Training Effectiveness: A Meta-Analysis and Some Critical Needs. International Journal of Aviation Psychology, 18(4), 353-368. DOI: 10.1080/10508410802347044
- O'Hanlon, J. F. (1981). Boredom: Practical consequences and a theory. Acta psychologica, 49(1), 53-82.
- Oppenheimer, A. N. (2000). Questionnaire Design, Interviewing and Attitude Measurement (2nd ed.). Bloomsbury Academic.
- Paliga, M. (2022). Human–cobot interaction fluency and cobot operators' job performance. The mediating role of work engagement: A survey. Robotics and Autonomous Systems, 155, 104191. Doi: 10.1016/j.robot.2022.104191
- Palmer, E. (1995). Oops, it didn't arm-a case study of two automation surprises. In Proceedings of the Eighth International Symposium on Aviation Psychology (pp. 227-232). Columbus, Ohio: Ohio State University.
- Palmer, E. (1999). Murphi busts an altitude: A murphi analysis of an automation surprise. In Digital Avionics Systems Conference, 1999. Proceedings. 18th (Vol. 1, pp. 4-B). IEEE.
- Parasuraman, R. (1987). Human-computer monitoring. Human Factors: The Journal of the Human Factors and Ergonomics Society, 29(6), 695-706.
- Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance consequences of automation induced 'complacency'. The International Journal of Aviation Psychology, 3(1), 1-23.
- Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. Human Factors: The Journal of the Human Factors and Ergonomics Society, 39(2), 230-253.
- Parasuraman, R., & Wickens, C. D. (2008). Humans: Still vital after all these years of automation. Human Factors: The Journal of the Human Factors and Ergonomics Society, 50(3), 511-520.
- Patterson, E. S., Watts-Perotti, J. C., & Woods, D. D. (1999). Voice Loops as Coordination Aids in Space Shuttle Mission Control. Computer Supported Cooperative Work, 8, 353–371.
- Regan, M. A., Horberry, T., & Stevens, A. (2014). Driver Acceptance of New Technology: Theory. Measurement and Optimisation.
- Rovira, E., McGarry, K., & Parasuraman, R. (2007). Effects of imperfect automation on decision making in a simulated command and control task. Human factors, 49(1), 76-87.
- Roy, Q., Zhang, F., & Vogel, D. (2019). Automation Accuracy Is Good, but High Controllability May Be Better. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 1-8. Doi: 10.1145/3290605.3300750
- Rubin, J., Chisnell, D., & Spool, J. (2008). Handbook of Usability Testing : How to Plan, Design, and Conduct Effective Tests (2nd ed.). John Wiley & Sons, Incorporated. ISBN: 978-0-470-18548-3
- Salas, E., Fowlkes, J. E., Stout, R. J., Milanovich, D. M., & Prince, C. (1999). Does CRM Training Improve Teamwork Skills in the Cockpit?: Two Evaluation Studies. Human Factors, 41(2), 326-343. DOI: 10.1518/001872099779591169
- Salvendy, G. (Ed.) (2012). Handbook of Human Factors and Ergonomics, 4th ed. Wiley.
- Sakamoto, Ryo, Akio Nozawa, Hisaya Tanaka, Tota Mizuno, and Hideto Ide. 2006. 'Evaluation of the Driver's Temporary Arousal Level by Facial Skin Thermogram-Effect of Surrounding Temperature and Wind on the Thermogram-'. IEEJ Transactions on Electronics, Information and Systems 126 (7): 804–9. https://doi.org/10.1541/ieejeiss.126.804.
- Sarter, N.B., Mumaw, R.J., Wickens, C.D., (2007). Pilots' monitoring strategies and performance on automated flight decks: an empirical study combining behavioral and eye-tracking data. Hum. Factors 49 (3), 347–357. Doi: 10.1518/001872007X196685

- Sarter, N. B., & Woods, D. D. (1994). Pilot interaction with cockpit automation II: An experimental study of pilots' model and awareness of the flight management system. The International Journal of Aviation Psychology, 4(1), 1-28.
- Sarter, N. B., & Woods, D. D. (1995). How in the world did we ever get into that mode? Mode error and awareness in supervisory control. Human Factors: The Journal of the Human Factors and Ergonomics Society, 37(1), 5-19.
- Sarter, N. B., Woods, D. D., & Billings, C. E. (1997). Automation surprises. Handbook of human factors and ergonomics, 2, 1926-1943.
- Sauro, J. & Lewis, J. R. (2011). When designing usability questionnaires, does it hurt to be positive?. Proceedings of the 2011 SIGCHI Conference on Human Factors in Computing Systems, 2215-2224. DOI: 10.1145/1978942.1979266
- Sheridan, T. B., and Verplank, W. L. (1978). Human and Computer Control of Undersea Teleoperators. MIT Man-Machine Systems Laboratory
- Sinha, R., Curran, P., Merritt, S., & Ilgen, D. (2008, August). Role of trust in decision making: trusting humans versus trust-ing machines. Paper session at the Academy of Management annual meeting, Anaheim, CA.
- Singh, I. L., Molloy, R., & Parasuraman, R. (1993). Automation-induced" complacency": Development of the complacency-potential rating scale. The International Journal of Aviation Psychology, 3(2), 111-122.
- Stanton, N. A., Salmon, P. M., Walker, G. H., Baber, C., & Jenkins, D. P. (2013). Human Factors Methods: A Practical Guide for Engineering and Design. 2nd ed. CRC Press.
- Stawarczyk, D., Majerus, S., Maquet, P., & D'Argembeau, A. (2011). Neural Correlates of Ongoing Conscious Experience: Both Task-Unrelatedness and Stimulus-Independence Are Related to Default Network Activity. PLoS ONE, 6(2), e16997.
- Strauch, B. (2002). Investigating human error: Incidents, accidents, and complex systems. Burlington, VT:Ashgate.
- Shneiderman, B., & Plaisant, C. (2006). Strategies for evaluating information visualization tools: multi-dimensional in-depth long-term case studies. In Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization (pp. 1-7). ACM.
- Taleb, N. N. (2012). Antifragile: Things that gain from disorder (Vol. 3). Random House Incorporated.
- Tanner Jr, W. P., & Swets, J. A. (1954). A decision-making theory of visual detection. Psychological review, 61(6), 401. doi: 10.1037/h0058700
- Tokadlı, G., & Dorneich, M. C. (2022). Autonomy as a teammate: Evaluation of teammate-likeness. Journal of Cognitive Engineering and Decision Making, 16(4), 282-300. DOI: 10.1177/15553434221108002
- Van Charante, E. M., Cook, R. I., Woods, D. D., Yue, L., & Howie, M. B. (1992). Human-computer interaction in context: Physician interaction with automated intravenous controllers in the heart room. Analysis, design, and evaluation of man-machine systems, 263-274.
- Van der Laan, J.D., Heino, A., & De Waard, D. (1997). A simple procedure for the assessment of acceptance of advanced transport telematics. Transportation Research - Part C: Emerging Technologies, 5, 1-10.
- Vandenbosch, B., & Higgins, C. (1996). Information acquisition and mental models: An investigation into the relationship between behaviour and learning. Information Systems Research, 7(2), 198-214. doi: 10.1287/isre.7.2.198
- Verberne, F., Ham, J. & Midden, C. (2012). "Trust in Smart Systems: Sharing Driving Goals and Giving Information to Increase Trustworthiness and Acceptance of Smart Systems in Cars." Human Factors 54 (5): 799–810. doi: 10.1177/0018720812443825.
- Victor, T.W., Tivesten, E., Gustavsson, P., Johansson, J., Sangberg, F., Ljung Aust, M., (2018). Automation expectation mismatch: incorrect prediction despite eyes on threat.
- Wickens, C. D. (1994). Designing for situation awareness and trust in automation. Proceedings of IFAC integrated systems engineering, 77-82.
- Wiener, E.L. (1988). Cockpit automation. In E.L. Wiener and D.C. Nagel (Eds.), Human factors in aviation (pp.433-461). San Diego: Academic.
- Wierwille, Walter W., and F. Thomas Eggemeier. 1993. 'Recommendations for Mental Workload Measurement in a Test and Evaluation Environment'. Human Factors: The Journal of the Human Factors and Ergonomics Society 35 (2): 263–81. https://doi.org/10.1177/001872089303500205.
- Wixon, D., Holtzblatt, K., & Knox, S. (1990). Contextual design: an emergent view of system design. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 329-336. DOI: 10.1145/97243.97304

- Woods, D. D., & Tinapple, D. (1999). W3: Watching human factors watch people at work. In Presidential address, presented at the 43rd Annual Meeting of the Human Factors and Ergonomics Society, Houston, TX.
- Zenju, Hideyuki, Akio Nozawa, Hisaya Tanaka, and Hideto Ide. 2004. 'Estimation of Unpleasant and Pleasant States by Nasal Thermogram'. IEEJ Transactions on Electronics, Information and Systems 124 (1): 213–14. https://doi.org/10.1541/ieejeiss.124.213.