D4.1 - Metrics and plan for V&V of the TeamMate HMI software in the 1st cycle

Project Number:	690705
Classification	Public
Deliverable No.:	D4.1
Work Package(s):	WP4
Document Version:	Vs. 1.0
Issue Date:	31.01.2017
Document Timescale:	Project Start Date: September 1, 2016
Start of the Document:	Month 3
Final version due:	Month 5
Deliverable Overview:	**Main document:** D4.1 **Annex I:** Excel table "AutoMate_WP4_Metrics_HMI"
Compiled by:	Stefan Suck, HMT
Authors:	Stefan Suck, HMT Maria Giulia Losi, REL Ebru Burcu Dogan, VED Jürgen Pichen, ULM Zoltan Kovacs, BIT
Technical Approval:	Fabio Tango, CRF
Issue Authorisation:	Andreas Lüdtke, OFF

DISTRIBUTION LIST		
Copy type[1]	Company and Location	Recipient
T	AutoMate Consortium	all AutoMate Partners

RECORD OF REVISION		
Date	Status Description	Author
19.12.2016	Deliverable structure	Stefan Suck, HMT
31.01.2017	Final version	all authors

[1] Copy types: E=Email, C=Controlled copy (paper), D=electronic copy on Disk or other medium, T=Team site (Sharepoint)

Table of Contents

1 Introduction

For the TeamMate car the human driver and the automation are seen as partners who collaborate to achieve a common goal, e.g., safe and comfortable driving. In order to implement this concept an HMI is required which supports the teamwork and cooperation between the human driver and the machine. The HMI should enable the driver to perceive the automation as a transparent and comprehensible teammate.

The goal of WP4 is the development of such an HMI. Core aspects of the HMI are a shared situation representation between driver and automation, and a flexible task distribution between the human driver and the automation system. This includes the design and implementation of an *information structure* and a *multi-modal interaction concept*, a *safe and robust strategy for the hand-over of vehicle control*, as well as the visualizations of data provided by other AutoMate modules, software etc.

This deliverable describes the initial plan on the verification and validation of the HMI concepts and software developed within WP4. The goal is to show the basic approach to ensure the quality of the developments of this work package. The verification and validation approach will be refined and adapted after cycle 1 and cycle 2.

2 General approach to TeamMate HMI verification and validation, and metrics specification

In WP4 a Human-Machine Interaction concept and a respective Human Machine Interface will be developed and implemented as software. The concepts and their software implementations have to be verified and validated.

The interaction concept and the interface have to be tested with human participants and they should be accepted by them. Additionally, the implementation of the concept and the interface has to work performant and correctly. Thus, verification and validation in WP4 are seen from technical and human factors' perspectives.

Verification should be understood as the evaluation of whether the TeamMate HMI complies with certain requirements. Which includes that it is determined if the HMI has the required functionalities and if these functionalities are working intended, without errors, considering certain constraints.

The validation is the evaluation of how appropriate the HMI is for the intended use. This includes requirements like user acceptance, appropriate workload, increased safety etc.

For each aspect that has to be verified or validated a related metric is necessary to determine to which degree a requirement is fulfilled. The type of the metric depends on the requirement. So, a metric is not necessarily a numerical value. So, it is also possible that qualitative expert judgement sometimes is considered as an appropriate outcome of verification and validation.

The mentioned requirements are based on the use-cases described in D1.1. For the interaction concept these are mainly high-level requirements from which further, also lower-level, requirements for the interface and the software are derived. The initial requirements and possible metrics are described in section 6. The requirements and the related metrics will be revised during further cycles.

3 AutoMate HMI expected results

Considering the AutoMate use-cases, described in D1.1, and with respect to the Society of Automotive Engineers (SAE) scheme of automation levels, the drivers will sometimes find themselves as “passenger” of their own vehicles and sometimes they will have to resume control to manual driving depending on the situation. Hence, there will be a continuous shift between levels of automation. Furthermore, due to automation, drivers will be even more engaged in supervision and intervention rather than manual direct control of the vehicle [Saffarian 2012].

Drivers need to interact and cooperate with the automation system with respect to two main functions: (1) Authority transition, and (2) Human-automation instruction and feedback. Authority transition refers to the timing and procedure of transferring responsibility of the driving task from the human to the automation system, and vice versa. Examples of situations requiring such kind of transitions are: road block, critical whether conditions, unexpected maneuver of other vehicles, as well as the overriding of the driver of automation maneuver.

Human-automation instruction and feedback refers to the timing and procedure of communication between the driver and the automation.

Considering the interaction between the driver and the automation, the Cooperation and Communication within the Team, made of Human and Automation, plays a major role in HMI design, as depicted in the attached image.

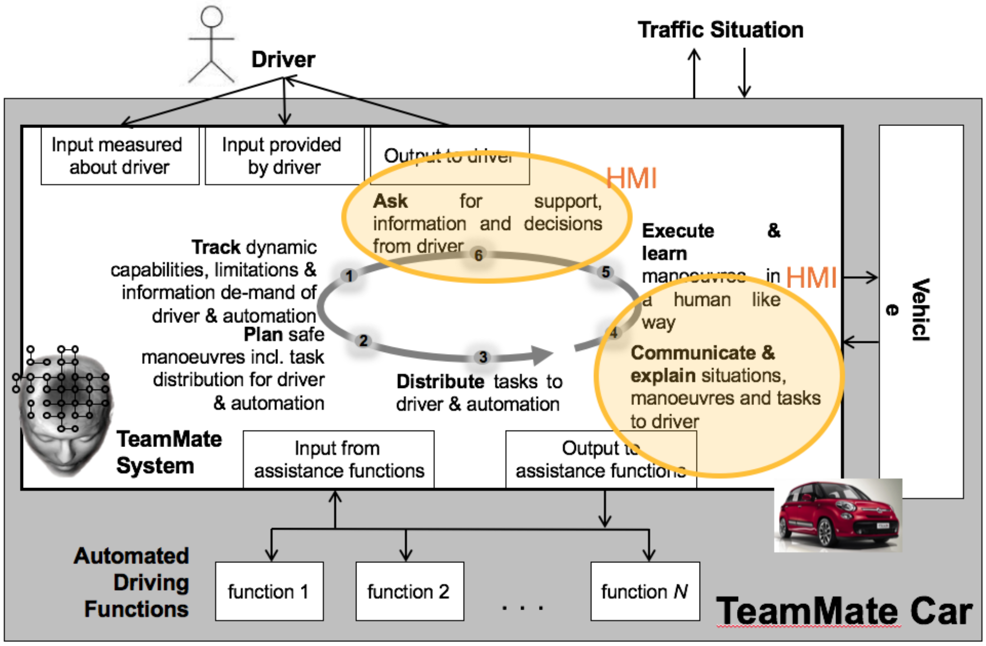

Figure 1: AutoMate Concept

How can a highly reliable automated driving system that users can understand, accept, trust, and eventually regularly use, be developed within the AutoMate project?

The HMI is expected to "communicate and explain situations, maneuvers, and tasks to the driver", as well as to "ask for support, information and decisions from the driver".

More precisely, as depicted in the scheme below, the HMI is expected to:

- Communicate situation:
  - Explaining traffic status (current and foreseen)
- Communicate capabilities:
  - Explaining automation status (current and foreseen)
  - Explaining driver status (current and foreseen)
- Communicate maneuver

- Communicating suggested behaviour (i.e. maneuver) in a way that it is personalized, adaptive to the driver status, and multimodal (visual, acoustic, haptic).
- Communicate and assign tasks:
  - Enabling bidirectional communication between driver and automation
  - Suggesting and requesting support or decision from the driver
  - Enabling multimodal user inputs

HMI is also expected to:

- Improve trust
- Improve acceptance
- Improve safety

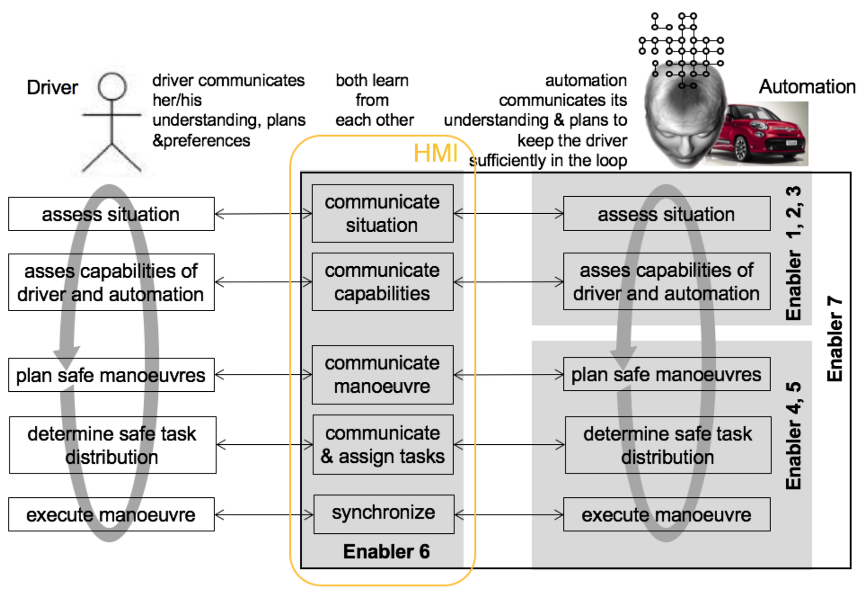

Figure 2: AutoMate HMI expected functionalities

The key question for successful automation is not "who has control over what or how much". It is "how do we get along together”. Indeed, a shared situation representation has to be maintained by means of bi-directional communication.

On the one side, the TeamMate Car needs to communicate to the human driver its situation representation, goals, and plans in a way that does not overload the driver. All communication takes driver's situational awareness into account to prevent annoying the driver, consequently it provides only information that is not yet known.

On the other side, there has to be mechanisms that allow the driver to communicate goals and plans and relevant aspects of the situation to the TeamMate Car.

4 Human-Automation Cooperation

To ensure a successful cooperation four basic requirements must be continuously fulfilled. These four requirements are the establishment and maintenance of a “Basic Compact”, a shared situation representation, mutual predictability and directability (e.g., Christoffersen & Woods, 2002; Klein, Woods, Bradshaw, Hoffman, & Feltovich, 2004; Walch, Lange, Baumann, & Weber, 2015). During the cooperation both partners agree to hold up to these agreements. In case a partner wants to drop out of the cooperation the other partner must be informed intelligible and in time.

The basic compact represents a principal agreement between agents “to facilitate coordination, work toward shared goals, and prevent breakdowns in team coordination” (Klein et al., 2004, p. 91). It has to be continually renewed and maintained. The basic compact includes the expectation that partners invest effort in activities to enhance the integrity of the compact and to repair faulty mutual knowledge, beliefs or assumption. Both partners must be able to inform the other partner if oneself is not capable of fulfilling the assigned and agreed role in the current activity. For safety reasons it is good if each partner has the ability to finish said activity in case the partner fails. In case he is not able to finish or fulfill the role of the failing partner, an alternative way of finishing the activity must be available – even if it means that the underlying goal will not be satisfied – to get to a safe state.

Another important requirement is the shared situation representation. Both partners should have a compatible view of the current situation and should possess knowledge about the other partner’s plans, activities and status. This shared situation representation together with general knowledge about partner’s general strategies and behavior across contexts allows an efficient and effective coordination of activities distributed over the partners (Christoffersen & Woods, 2002). It needs to be continually updated. For this it is necessary that the machine partner provides information to the driver that is related to the current task goals, and easy to process. The information or feedback has to support the establishment and maintenance of an integrated understanding of the situation, the partner's activities and the future evolvement of them (Christoffersen & Woods, 2002).

To efficiently and safely plan the future actions that must be taken, the mutual predictability must be assured. The TeamMate car and the driver should openly communicate their current state so the opposite partner can predict which future actions will be taken. This shared knowledge serves both partners as a base to understand and predict future actions of each other. It is also crucial to know about the other party's future goals to plan one's own goals (e.g., Klein et al., 2004).

The mutual directability in the car is a central point of the TeamMate car concept. It means that in every situation there is the possibility to assess and modify the actions taken by a partner. Consequently, a smooth transition of task control is one of the results of an interactive directability.

5 AutoMate Use-cases

HMI concept and the Validation of the Concept itself will be also based on process steps of the use cases. In this phase of the project PETER's use cases were developed. From every PETER use case an UML process flow diagram was created (UML – Unified Modelling Language).

Peter's use cases:

The driver indicates that he wants to overtake, system overtakes

Use-Case 1.pdf

The indication from the driver to overtake is possibl but the system needs driver's full attention because of bad road/wheather conditions

Use-Case 2.pdf

The indication from the driver to overtake is in contradiction with traffic laws

Use-Case 3.pdf

The indication from the driver to overtake is in contradiction with the road/weather condition, system refuses

Use-Case 4.pdf

The driver does not know if he can overtake the leading vehicle, the system offers him to overtake

Use-Case 5.pdf

The indication from the driver to overtake is in contradiction with the traffic situation

Use-Case 6.pdf

6 Validation of TeamMate HMI: HMI Hypothesis, Performance Indicators (PI), and related Metrics

As already mentioned the validation is about the appropriateness of the TeamMate HMI for the intended use and to which extend the users accept the HMI. So the validation TeamMate HMI is mainly focused on human factor aspects. Several general requirements have been defined which should be addressed by the HMI to support Human-Automation Cooperation. Multiple functional requirements (HMI functionalities) are derived from the general requirements. For each functionality some hypotheses were made, which have to be tested during the validation experiments with human drivers. The requirements, hypotheses, and measures to determine if the hypotheses are accepted or rejected are described in this section.

The approach to the definition of the AutoMate metric framework in D4.1 has been focused on the definition of HMI hypothesis, performance indicators (PI) and Metrics that will guide the HMI design phase (in WP4) as well as the identification of the most suitable metrics to be applied in the phase of HMI validation (in WP6) as well as the tools and methods needed to collect those metrics (in WP6).

6.1 HMI hypothesis

HMI hypothesis have been identified for each HMI expected functionality. The hypothesis concerns how the HMI is supposed to be working in the context of human-automation cooperation.

HMI hypothesis have been derived from the objectives of the TeamMate car approach. For each hypothesis, a performance indicator (PI) has been identified.

The PIs have associated metrics and measures, derived from Human Factors' literature, which can be applied for validating the hypothesis.

In order to validate the hypothesis, specific validation tests with end users, either in simulated driving or in real vehicles, will be setup and identified for each demonstrator.

The selection of the use-cases applied in each of the three AutoMate demonstrators as well as the hypothesis tested, will be performed during cycle 2 and 3 of the project.

6.1.1 Hypothesis for Shared situation representation

Concerning shared situation representation, both partners should have a compatible view of the current situation and should possess knowledge about the other partner's plans, activities and status.

HMI expected functionality: As far as Shared situation representation is concerned, the HMI is expected to *communicate to the driver the driving situation.*

Hypothesis for validating Shared Situation representation are:

1. The driver understands TeamMate status correctly and at the right time
2. The driver understands if TeamMate is active or inactive
3. The driver understands what TeamMate is doing (current behaviour)

4. The driver understands why TeamMate is doing what is doing (current behaviour)
5. The driver understands environment/traffic situation correctly and at the right time
6. The HMI communicates to the driver mission status correctly and at the right time

The above mentioned hypothesis can be validated by measuring the driver **Situation Awareness** (see paragraph 6.1.2.1).

6.1.2 Hypothesis for Mutual predictability

Mutual predictability implies that Actions (either Human or Automation) need to be mutually predictable (usually based on previous experience of cooperation).

HMI expected functionality: As far as Mutual predictability is concerned, the HMI is expected to communicate (1) capabilities (either of the driver and the automation) as well as (2) maneuvers (including current maneuver and next maneuver).

Hypothesis identified for validating mutual predictability are**:**

1. The driver understands automation's capabilities
2. The driver understands TeamMate status (i.e. it’s working/not working)
3. The driver understands TeamMate suggested maneuver in an efficient and effective way
4. The driver understands what TeamMate is going to do next (Prevision)
5. The driver understands why TeamMate is going to do it (Prevision)
6. The driver predicts correctly how TeamMate will behave in different driving situations

The above mentioned hypothesis can be validated by measuring the driver **Situation Awareness** (see paragraph 6.1.2.1).

6.1.3 Hypothesis for Directability

Directability implies that Actions (either Human or Automation) need to be mutually directable.

HMI expected functionality: As far as Directability is concerned, the HMI is expected to communicate and assign tasks.

Hypothesis identified for validating directability are**:**

1. The driver understands task distribution between human driver and automation. This hypothesis can be validated by measuring the driver Situation Awareness (see paragraph 6.1.2.1).
2. The HMI communicates task assignment to the driver in an efficient way taking into consideration the driver status,
3. The driver understands TOR (take-over request) coming from TeamMate system (i.e.The driver interprets and executes the intervention requested by TeamMate and it executes it correctly),
4. The driver communicates TOR (take-over request) to the TeamMate system (i.e. override) in an efficient and effective way.

Hypothesis (2), (3), and (4) can be validated by measuring the **Driver Workload** (see paragraph 6.1.2.2).

6.1.4 Hypothesis for Trust, Acceptance, and perceived Safety

HMI expected functionality:

The HMI is expected to increase driver trust in the system, in terms of willingness to rely on an automation system, as well driver acceptance of the system.

The **hypothesis** are that following:

(1) the driver trusts the TeamMate system correctly,

(2) the driver accepts the TeamMate system.

6.2 Performance indicators and metrics

6.2.1 Situation awareness

Situation awareness (SA) is the “perception of the elements in the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in the near future” (Endsley, 1995).

Traditionally SA consist of three levels and every one encompasses the previous one (Endsley, 1996).

SA has three levels (Endsley, 1991): level 1, perception of the elements in the environment; level 2, comprehension of the current situation; and level 3, projection of future status.

In other words SA can be defined as an own independent interpretation of the world.

For its own definition and properties, it is a crucial factor in the decision making process and, as a consequence, in the control of a dynamic complex system.

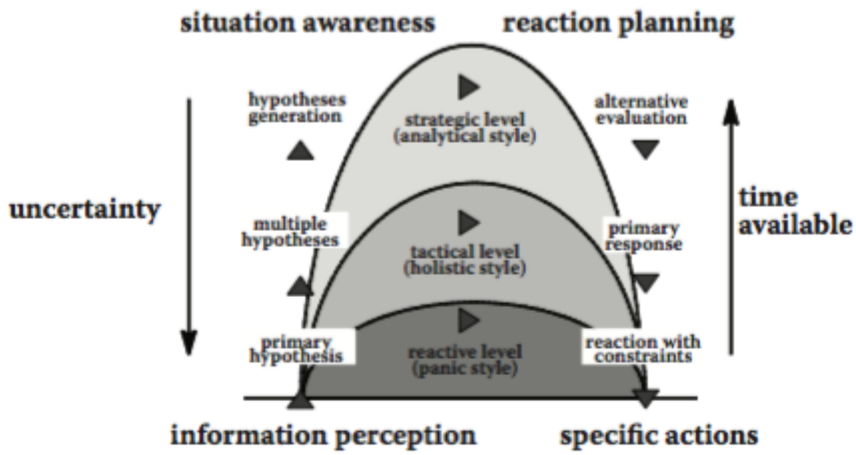

Figure 3: Decision making under uncertainty and time pressure

A decrease of SA level may be determined by several factors: i.e. distraction, interaction with automatic system, sleepiness, complacency.

During driving, usually people are placed in SA Level 1 when they are detecting events occurring around their vehicle. When a decrease of situation awareness happens, it is usually a loss of SA Level 3, that origins an incorrect expectation on traffic events.

The effects of these cognitive shifts of attention, arousal and situation awareness are reflected by driving behaviour, expressed in terms of performance quality.

A guide to select the appropriate SA measure is depicted in the figure below [Gawron 2008]:

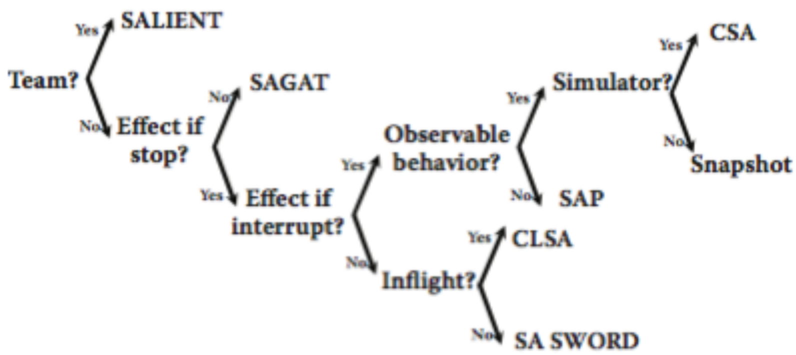

Figure 4: SA measures

There are four types of SA measures: performance, subjective ratings, simulation, and physiological measures.

A selection for each type of measures is the following:

SA performance measure:

6.2.1.1 SAGAT (Situation Awareness Global Assessment Technique)

"Using SAGAT, the simulation is stopped at random times, and the operators are asked questions to determine their SA at that particular point in time. Subjects' answers are compared with the correct answers" (Endsley, 1988a).

It is an online measure.

In AutoMate the driver can be asked if he/she is aware in which "mode" the system currently is and what it can do in this mode and what not.

Possible results are:

a) driver thought that he understood the situation, and processing not too demanding

b) check if s/he really got the needed information and made the right interpretation and measure effort

c) mixture

Measure limitations:

- The freeze method of SAGAT may compromise driver behavior,
- It might not precisely measure SA but what drivers can recall.

SA subjective measure (self-ratings):

6.2.1.2 SART (Situational Awareness Rating Technique)

It is an offline measure.

According to (Taylor, 1990) the 10 scales of the SART is a questionnaire method that focus on three areas:

- Demands on attentional resources (D): a combination of complexity, variability and instability of the situation,
- Supply of attentional resources (S): a combination of arousal, focusing of attention, spare mental capacity and concentration of attention.
- Understanding of the situation (U): a combination of information quantity, information quality and familiarity of the situation.

SART scale is depicted in the following table:

		LOW						HIGH
		1	2	3	4	5	6	7
Demand	Instability of Situation							
	Variability of Situation							
	Complexity of Situation							
Supply	Arousal							
	Spare Mental Capacity							
	Concentration							
	Division of Attention							
Under	Information Quantity							
	Information Quality							
	Familiarity							

Figure 5: SART scale

SART indicator is computed as: U-(D-S), where U, D, and S are the sum of the answers to the SART scales belonging to these aggregations.

Measure limitations:

- It is a subjective measure.

6.2.2 Driver workload

Driver workload refers to the relationship between the amount of effort required by the driving task and driver's capacity. More precisely, it is the cost and difficulty experienced by the driver to meet the driving task demands.

Introduction of automation changes driver workload. Indeed, automation is intended to reduce the demand on the driver as the driver would be relieved from the cognitive (i.e. information processing) and physical (i.e. vehicle control) demands of driving. Nonetheless, automation seems to induce mental underload, which is associated with reduced attentional capacity and impoverished performance. This is of particular importance when there is a sudden increase in situational demands, such as at the moment of a takeover request.

De Waard (1996) proposes six regions to explain the inverted-U relationship between the workload and performance (Figure X). The region D in the figure refers to the tasks that require very low workload. Driving task during the automated driving period, for instance, may fall into this region. The regions, namely, A1, A2, and A3 indicate a stable performance within the limited capacity of the driver. Note that, however, in region A3, human driver can compensate for increased workload thanks to human flexibility. If the required effort persists, the region-B is entered, moving to performance deterioration in region C.

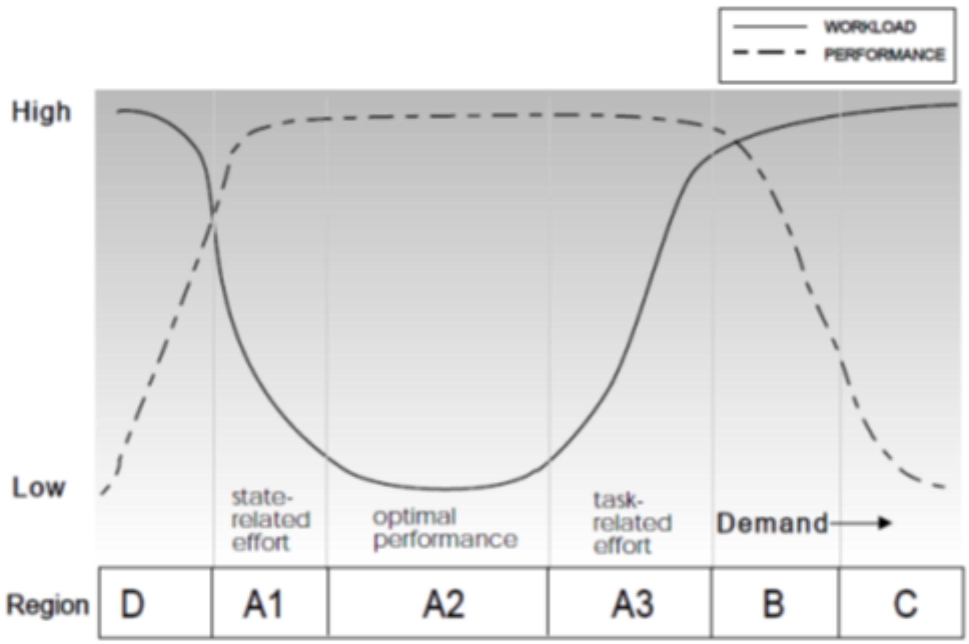

Figure 6: Workload and performance in 6 regions (from de Waard, 1996)

In the case of automated driving, drivers suddenly move from a very low level of workload during automated driving to a relatively high level of workload when the vehicle issues a takeover request. The aim of the AutoMate concept is to prepare the driver to this transition.

The common measures of driver workload are:

Performance measures

6.2.3 Driving task performance

These measures are explained in detail in Section 6.1.2.5 Intervention and 6.1.2.7 Driving maneuver quality.

Non-driving related task performance

Drivers' performance on the non-driving related tasks (NDRT) could be used as an indicator of workload. Beyond level 3 automation, drivers will have the chance to engage in the NDRT presented via the vehicle infotainment system. The performance criteria will depend on the NDRT used. For instance, for drivers who read news on the infotainment system, total number of questions answered correctly about the news could be an indicator. This can as well be a disadvantage of the current measure.

6.2.3.1 Detection Response Task (DRT)

DRT is a standardized indicator to assess the effects of cognitive load on driver attention while performing a secondary tasks that are not related to the driving task (ISO 17488: 2016). DRT can be in visual-manual, voice-based or haptic modes.

In AutoMate, DRT can be used in combination with the NDRT in order to study the attentional effects of cognitive load linked to the NDRT.

Measure limitations

- The equipment is obtrusive.

Subjective measures

6.2.3.2 NASA Task Load Index

NASA TLX (Hart & Staveland, 1988) is a subjective, multidimensional scale consisting of six subscales on mental demand, physical demand, temporal demand, performance, effort, and frustration. Participants are provided with the definition of each dimension to make sure that they can provide accurate ratings.

It is an offline measure.

In AutoMate, drivers can be asked the demand they experienced on the NASA TLX dimensions during the takeover request after they finished an experimental lap.

Measure limitations

- Subjective ratings

6.2.3.3 Rating Scale Mental Effort

RSME (Zijlstra & Van Doorn, 1985) is a unidimensional scale. Participants rate the amount of invested effort into the driving task on a 150-milimetre scale, each 10 mm corresponding to a different anchor.

In AutoMate, RSME can be administered right after the takeover request, decreasing the temporal proximity between the event and the self-report, as it is an easy, unidimensional measure.

Measure limitations:

- Subjective ratings

Physiological measures

6.2.3.4 Eye movements

Visual search is a strategy to collect information about the task and task environment. Gaze behavior is used as an indicator of driver workload. Three measures that will be adopted in the AutoMate are:

- **Blink frequency:** Some studies show that workload increases how frequently drivers blink their eyes, whereas some other studies the opposite. A hypothesis to account for these opposite findings distinguish visual and cognitive workload (Recarte et al., 2008). Accordingly, search for visual information is proposed to inhibit blinks, enabling a maximum information intake. Cognitive workload, on the other hand, is proposed to increase the frequency of the blinks.
- **Horizontal gaze dispersion:** It is another indicator used to quantify visual search. Supposedly, cognitive workload decreases horizontal gaze dispersion, that is, drivers scan a smaller area for information.

In AutoMate, we can compare gaze behavior among baseline and TeamMate groups, as well as during different phases of automated driving in order to see the effect of workload.

Measure limitations:

- Gaze behavior may be sensitive to the light conditions
- High rate of data loss is pertinent problem for gaze behavior

6.2.3.5 Heart rate variability (HRV)

The contraction of heart can be measured in the form of an ElectroCardioGram (ECG) signal. The measures pertinent for time domain, frequency, and amplitude can be inferred from ECG.

A frequently used time domain measure is heart rate variability. HRV coefficient or index is obtained by dividing the standard deviation of inter-beat-interval (IBI) by the average IBI.

As a measure of driver workload, HRV decrease is sensitive to increases in mental workload.

In AutoMate, HRV can be used through the automated driving as well as during takeover periods.

Measure limitations:

- Sometimes we fail to link HRV and mental workload, mostly due to the physical demands of an activity.
- It is obtrusive, although more and more unobtrusive HRV measurement tools are becoming available in the market.

6.2.3.6 Electrodermal activity (EDA)

Electrodermal activity refers to the electrical changes in the skin and expressed in terms of skin conductance or skin resistance. EDA is an indicator of arousal and emotion in general.

The most common measure of skin response is Galvanic Skin Resistance (GSR). In the driving context, it is measured on the palm of the driver as the glands here are sensitive. GSR is considered as an indicator of information processing rather than mental workload.

In AutoMate, similar to HRV, GSR can be measured especially during driver response to a takeover request.

Measure limitations:

- It is a measure of global sensitivity, that is to say, emotional and physical.

6.2.3.7 Attention Allocation Efficiency

In order to evaluate if the driver understands TOR (take-over request) coming from TeamMate system (i.e. The driver interprets and executes the intervention requested by TeamMate and it executes it correctly), attention allocation efficiency can be used "for assessing if operators know where to find the information or the functionality needed, and also when to look for a given piece of information and when to execute a given function" (Madhavan et al. 2009, Chapter 2).

In supervisory control, users divide their attentional resources into a series of dynamic processes (i.e. driving task and not-driving related tasks), thus receiving information coming from multiple channels and looking for critical events.

To evaluate Attention Allocation Efficiency there are different types of measures:

Performance indicators	P.I. definition	Type of metrics	Metric	Online/ Offline metric	Tools/ sensors/ techniques
Attention Allocation Efficiency	It is used for assessing if operators know where to find the information or the functionality needed, and also when to look for a given piece of information and when to execute a given function.	Performance measures	Proportion of time that the visual gaze is within each "area of interest" of an interface	Offline	Eye tracking
		Performance measures	Average number of visits to each "area of interest" of an interface	Offline	Eye tracking
		Performance measures	Switching time for multiple tasks	Offline	Human interface-inputs
		Performance measures	Information used	Offline	Human interface-inputs

Figure 7: Measures for Attention Allocation efficiency

The length and frequency of eye fixations on a specific area is an indicator of the level of attention on that area/element.

Concerning the following performance indicators (i.e. Reaction Time, Intervention, Remaining time, and Reaction Quality) a useful table on measures has been derived from Gold et al. 2016.

Dependent Variable	Unit	Category	Explanation
Hands-on time ($t_{hands-on}$)	Seconds	Timing aspects	Time between TOR and hands on steering wheel
Takeover time ($t_{takeover}$)	Seconds	Timing aspects	Time between TOR and start of maneuver
Maximum longitudinal accelerations (a_{long})	m/s²	Quality aspects	Maximum braking acceleration during situation
Maximum lateral accelerations (a_{lat})	m/s²	Quality aspects	Maximum lateral acceleration during situation
Minimal time to collision (TTC)	Seconds	Quality aspects	Minimal TTC during situation
Horizontal gaze dispersion (HGD)	Pixels	Workload (due to task)	Deviation of horizontal gaze position

Note. TOR = takeover request.

Figure 8: Measures for Intervention, Reaction Time, Remaining Time, Reaction Quality

6.2.4 Reaction time

Concerning take-over request (TOR), the common situation is the the automation has taken over longitudinal and lateral control of the vehicle and the driver is supervising, usually not continuously monitoring the system. When the system requests a take-over, the driver needs to be given sufficient time for take-over and for re-gaining the correct driving position.

The driver reaction time is considered the point in time when the subjects start to brake or steer consciously.

It has been verified that braking generally occurs 1 second before the steering input [Gold et al. 2013].

With longer TOR times usually reactions (either on the steering wheel or the brake pedal) occur later in time.

The metrics that can be considered for the reaction time are:

Performance measures

- **Gaze reaction,** i.e. the time the first saccade stirs from the not-driving related task,
- **Road fixation,** i.e. the time the first gaze look at the driving scenario,
- **Hands on,** i.e. the time of the TOR before the driver puts the hands on the steering wheel.

Gaze reaction and road fixation require the installation of an eye tracking system on-board.

6.2.5 Intervention

For intervention is intended when the subject arrives at a decision and starts the maneuver, either initially braking or steering.

There are limits defined from which an input is assumed to be a conscious maneuver.

Useful measures for intervention (described in the attached table) are:

- **Braking pedal position**
- **Force applied to the brake pedal**
- **Steering wheel angle**

Performance indicators	P.I. definition	Type of metrics	Metric	Online/ Offline metric	Tools/ sensors/ techniques	Success criteria
Intervention	The subject arrives at a decision and starts the maneuver. There are limits defined from which an input is assumed to be a conscious maneuver.	Performance measures	Braking pedal position	Online	CAN BUS data	At least 10 percent braking pedal position
		Performance measures	Force applied to the brake pedal	Online	CAN BUS data	
		Performance measures	Steering wheel angle	Online	CAN BUS data	At least 2 degrees steering wheel angle

Figure 9: Intervention measures

In order to distinguish between intentional and unintentional intervention, success criteria/thresholds have been defined for:

- Braking pedal position → 10% braking pedal position
- Steering wheel angle → 2 degrees

Below these thresholds, the the breaking and the steering inputs do not have the purpose of stabilizing the vehicle [Gold et al. 2013].

6.2.6 Remaining action time

After the driver intervention, the remaining action time, defined as the **Time-to-collision (TTC)**, is the time the driver has to perform a maneuver (either braking until full stop or taking over).

The drivers' inputs after the intervention can be considered an intentional maneuver.

This action implies that the driver has re-gained situation awareness of the driving scenario and takes a decision.

6.2.7 Driving maneuver quality

The quality of driver performance upon taking over control could be evaluated by several indicators. Although there is no established criteria or standardization for such measures, frequently used measures are:

- **Maximum longitudinal accelerations:** The maximum value of longitudinal acceleration during a period of certain distance or time after the takeover. It can be positive (speed up) as well as negative (brake) depending on the driver performance and maneuver.
- **Maximum lateral accelerations:** The maximum value of lateral acceleration during a period of certain distance or time after the takeover. It can provide information about the crash potential with the vehicles on adjacent lanes.
- **Acceleration potential:** It is the combined vector of longitudinal and lateral accelerations and an indicator of maneuver quality especially during a lane change maneuver.
- **Steering wheel angle:** Two steering wheel angle measures of interest are micro-corrections and frequency content of the steering wheel angle (AIDE, 2005). Micro-correction is the steering wheel reversals smaller than 1°. It is an indicator of driver's effort to lane keeping. Frequency component of the steering wheel signals are obtained by using a Fast Fourier Transform (FFT) in order to create frequency spectra for the steering wheel angle. The amplitude of the spectra would reveal driver's lateral vehicle control precision.
- **Minimum time-to-collision:** It is the minimum time left before the vehicle would collide with a lead vehicle or an obstacle if the velocities and accelerations of both vehicles were frozen in time. It is a useful indicator to determine the safety of a given traffic situation.

By recording and analyzing these measures, the quality of the driving maneuver can be assessed. Note that there are no widely accepted values of these measures to quantify takeover quality. However, if the performance on these measures reaches the physical limits, the driving can become unstable [Gold et al. 2013].

Performance indicators	P.I. definition	Type of metrics	Metric	Online/ Offline metric	Tools/ sensors/ techniques	Success criteria
Driving manouver quality	It is the evaluation of the quality of the manouver taken by the driver after TOR.	Performance measures	Maximum lateral acceleration of the vehicle	Online	CAN BUS data	
		Performance measures	Maximum longitudinal acceleration of the vehicle	Online	CAN BUS data	
		Performance measures	Acceleration potential	Online	CAN BUS data	
		Performance measures	Steering wheel angle	Online	CAN BUS data	

Figure 10: Driving manouver quality measures

6.2.8 Trust

Lee and See (2004) define trust as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" (p.51). Increased trust is not per se a goal of HMI design for autonomous driving as this could result in misuse of automation. Rather, a calibrated level of trust meeting the situational capabilities of the driving assistance fosters a safe and efficient driver vehicle interaction. Thus, a HMI concept should be designed in a way to not unconditionally increase trust, but to foster a calibrated level of trust. Thus, it should provide information on system reliability and provide feedback on current system states and planned maneuvers.

Trust in automation is a multifaceted construct that can be operationalized on different psychological levels (see Hoff & Bashir, 2015). Dispositional trust describes an enduring and general attitude towards automation while situational trust is directed at a certain automated system in a certain situation. In order to reflect the full picture in Automate, we will measure trust at different points in the interaction process with the highly automated driving assistance. This endeavor should provide answers to the question how cooperative assistance influences the process of trust development.

Thus, trust will be measured using a single item percentage slider bar indicator, psychometric scales (adapted from Jian, Bisantz, Drury & Llinas, 1998) and also eye tracking correlates indicating percentage of monitoring behavior in dual task conditions.

6.2.9 Acceptance

Driver's acceptance of the TeamMate system would influence their adoption of this new technology. Existing measures on acceptance focus on intention to use a new technology and/or its usage.

One of the commonly applied models to study acceptance is Technology Acceptance Model (TAM, Davis, 1989), which is tailored to the context of information systems. TAM is a subjective measure based on a questionnaire approach. The questionnaire aims to disentangle two dimensions:

- **Perceived usefulness** is the degree to which a person believes that using a certain technology would improve performance
- **Perceived ease of use** is the degree to which a person believes that using a certain technology would be easy and free of effort

TAM has been developed further by combining other socio-cognitive constructs focusing on:

- **Perceived behavioral control**, that is, perceived difficulty of using a new technology
- **Attitudes**, that is, individual's evaluations about a new technology
- **Subjective norms**, that is, individual's perception of how others would expect him or her to behave

In AutoMate, we will develop a questionnaire complying with the TAM and its extended version.

Another scale that could be used to measure drivers' acceptance of the TeamMate concept is the acceptance scale developed by Van der Laan and colleagues (1997). This is a system acceptance scale on two dimensions: usefulness (performance dimension) and satisfaction (affective dimension). The scale consists of nine items in dichotomy and asks drivers' evaluations of a new technology on each dichotomy.

An objective measure that will be used to study TeamMate acceptance is the percentage of the time that drivers spend on a non-driving-related task while TeamMate is actively driving.

6.3 Measure selection: costs and benefits

In the process of selecting the appropriate measure for a metric, costs and benefits need to be taken into consideration [Madhavan et al. 2009].

Some metric evaluation criteria are included in the attached table:

Evaluation criteria	Example
Experimental constraints	Time required to analyze a metric
Comprehensive understanding	Causal relations with other metrics
Construct validity	Power to discriminate between similar constructs
Statistical efficiency	Effect size
Measurement technique efficiency	Intrusiveness to subjects

Figure 11: Metric selection criteria

In the second cycles, measures will be selected in order to validate AutoMate HMI.

In this process of identification, time as well as budget constraints will be considered, as well as the stage of HMI development in which the system is in order to be evaluated and the testing environment (e.g. simulator, real vehicle).

Each metric will be evaluated based on how much this will give insights with respect to the research question that will be tested (e.g. workload can be measured by monitoring pupil dilation over time or by collecting subjective responses at the end of the experiment but the latter does not provide the same comprehensive understanding).

Cost- benefit parameters for metric selection are described in the attached table:

Costs		
Data gathering	Preparation	Time to setup Expertise required
	Data collection	Equipment Time Measurement error likelihood
	Subject recruitment	Compensation IRB preparation and submission Time spent recruiting subjects
Data analysis	Data storage/transfer	Equipment Time
	Data reduction	Time Expertise required Software
	Statistical analysis	Error proneness given the required expertise Time Software Expertise
Benefits		
Comprehensive understanding		Proximity to primary research question Coverage – Additional understanding given other metrics Causal relations to other metrics
Construct validity		Sensitivity Power to discriminate between similar constructs Inter-subject reliability Intra-subject reliability
Statistical efficiency		Effect size: Difference in means, Error variance Frequency of observations Total number of measures collected
Measurement technique efficiency		Non-intrusiveness to subjects Non-intrusiveness to task nature
Appropriateness for system development phase/testing environment		

Figure 12: Cost-benefit analysis for metric selection

7 Verification of TeamMate HMI software

To ensure quality of TeamMate HMI software two methods should be used where it is possible: unit testing to verify logically coherent parts of the software, and using bug tracking system to receive feedbacks about the whole HMI software.

7.1 Unit testing

In computer programming, unit testing is a software testing method by which individual units of source code, sets of one or more computer program modules together with associated control data, usage procedures, and operating procedures, are tested to determine whether they are fit for use. [Automated Defect Prevention: Best Practices in Software Management, 2007]

Most of the modern computer languages (like *Java* or *C#*) support unit testing. Proper usage of it ensures that different functional parts of the software are working as intended. Unit tests should be based on requirements table.

7.2 Using bug tracking system

Bug tracking systems help to track and manage reported software bugs. Collecting feedbacks from the testers and forwarding them to the software developers create loop which is necessary for any kind of quality assurance.

8 Conclusions

Within this document the process of the definition of the validation metrics for AutoMate Human Machine Interaction has been described.

The validation process has included the identification of HMI hypothesis to be validated in test sessions with final users.

For each hypothesis, Performance Indicators and related Metrics have been identified.

The verification process has been described with the aim of ensuring the quality of TeamMate HMI software.

In cycle 2, according to the Use-cases defined in D1.1, the experimental protocol will be defined for each demonstrator, selecting appropriate metrics to be validated and verified in driving simulation environment.

The description of experiments and the specification of success-criteria for each metric will also be covered in later versions of this deliverable when more information on the upcoming experiments and necessary aspects of HMI validation is available.

Metrics will be further refined and updated during the duration of the project.

9 References

AIDE (2005). Driving performance assessment – methods and metrics. IST-1-507674-IP

Christoffersen, K., & Woods, D. (2002). *How to make automated systems team players. Advances in Human Performance and Cogntive Engineerung Research*, 2, 1–12. http://doi.org/10.1016/S1479-3601(02)02003-9

Davis, F.D., (1986), *A technology acceptance model for empirically testing new end-user information systems: Theory and results*, PhD thesis, MIT

De Waard, D., (1996), *The measurement of drivers' mental workload (Unpublished doctoral thesis).* University of Groningen, Groningen, Netherlands.

Gawron V. J., (2008) *Human performance workload and situational awareness measures handbook,* Second Edition, CRC Press

Gold et al., (2013) *"Take over!" How long does it take to get the driver back into the loop?,* PROCEEDINGS of the HUMAN FACTORS and ERGONOMICS SOCIETY 57th ANNUAL MEETING - 2013

Gold et al., (2016) *Taking Over Control From Highly Automated Vehicles in Complex Traffic Situations: The Role of Traffic Density*, Institute of Ergonomics, Technical University of Munich, Germany, HUMAN FACTORS, Vol. 58, No. 4, June 2016, pp. 642–652

Hart, S. G., & Staveland, L. E. (1988). *Development of a multi-dimensional workload rating scale: Results of empirical and theoretical research*. In P. A. Hancock & N. Meshkati (Eds.), Human mental workload, 139-183. Amsterdam, The Netherlands.

Hoff, K. A., & Bashir, M. (2015). *Trust in automation integrating empirical evidence on factors that influence trust.* Human Factors: The Journal of the Human Factors and Ergonomics Society, 57(3), 407-434.

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. http://doi.org/10.1109/MIS.2004.74

Kolawa A. and Huizinga D., (2007), *Automated Defect Prevention: Best Practices in Software Management*. Wiley-IEEE Computer Society Press. p. 75. ISBN 0-470-04212-5

Jian, J., Bisantz, A. M., & Drury, C. G. (2000). *Foundations for an empirically determined scale of trust in automated systems*. International Journal of Cognitive Ergonomics, 4, 53–71.

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.

Llinas, J., Bisantz, A., Drury, C., Seong, Y., & Jian, J. (1998). *Studies and Analyses of Aided Adversarial Decision-Making.* In Phase 2: Research on Human Trust in Automation, Center for Multisource Information Fusion. State University of New York at Buffalo.

Madhavan R. et al., (2016) *Evaluation Criteria for Human-Automation Performance Metrics*, Chapter 2, in Performance Evaluation and Benchmarking of Intelligent Systems, ISO 17488:2016

Recarte, M. A., Perez, E., Conchillo, A., & Nunes, L. M. (2008). *Mental workload and visual impairment: Differences between pupil, blink and subjective rating.* Spanish Journal of Psychology, 11, 374–385.

Saffarian M. , de Winter J. C. F. , Happee R. (2012), *Automated Driving: Human-factors issues and design solutions,* PROCEEDINGS of the HUMAN FACTORS and ERGONOMICS SOCIETY 56th ANNUAL MEETING - 2012

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.

Walch, M., Lange, K., Baumann, M., & Weber, M. (2015). *Autonomous Driving: Investigating the Feasibility of Car-driver Handover Assistance*. In Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (pp. 11–18). http://doi.org/10.1145/2799250.2799268

Zijlstra, F., & van Doorn, L., (1985) *The construction of a scale to measure subjective effort.* In Technical Report. Delft University of Technology

10 Annex I.

Annex_D4_1_Metrics_HMI_Final.xlsx