TECHNISCHE UNIVERSITÄT BERLIN

Institut für Psychologie und Arbeitswissenschaften

Masterarbeit

zur Erlangung des Grades

Master of Science

(M. Sc.)

über das Thema

Remote Operation in Teleoperated Railways: Impact of a Supplementary Task, Attention-Guiding Take-Over Requests and Work Experience on Situational Awareness and Workload

Erstprüferin: Prof. Dr. Linda Onnasch

Technische Universität Berlin

Zweitprüferin: Dr. Annika Dreßler

Deutsches Zentrum für Luft- und Raumfahrt

Eingereicht von David Schackmann

am 20. Juni 2025

Zusammenfassung

Hintergrund Teleoperierter Bahnbetrieb ermöglicht Effizienzsteigerungen durch den Einsatz von Automationssystemen. Da das Eingreifen von Menschen jedoch weiterhin notwendig ist, wenn diese Systeme an ihre Grenzen stoßen, ist ein hohes Maß an Situationsbewusstsein (Situational Awareness, SA) erforderlich. Allerdings kann SA insbesondere durch längere Phasen passiven Monitorings negativ beeinträchtigt werden. Diese Studie untersucht daher, ob eine Nebenaufgabe (Supplementary Task, ST) dazu beitragen kann, die Aufmerksamkeit aufrechtzuerhalten, gegebenenfalls unter der Abnahme von SA. Desweiteren wird überprüft, ob aufmerksamkeitslenkende Übernahmeaufforderungen (AGTORs) und berufliche Vorerfahrung als Triebfahrzeugführer:innen (Tf) diese Effekte kompensieren können. **Methoden** Es wurde eine Online-Studie im Mixed-Design durchgeführt, an der 160 Personen teilnahmen, davon 28 aktive Tfs. Alle Teilnehmenden bearbeiteten zwei Blöcke mit je vier kurzen Übernahmeszenarien, einmal mit ST, einmal ohne. Zusätzlich erhielt eine Gruppe AGTORs und die andere einfache Warntöne als Übernahmeaufforderungen. Die Aufgabe der Teilnehmenden war es, Übernahmegründe zu identifizieren und relevante Informationen aus der Umgebung zu erfassen. SA wurde über die erkannten Übernahmegründe und SAGAT-Fragebögen erfasst. Zudem wurden weitere Variablen erhoben, wie die mentale Arbeitsbelastung. Die Daten wurden mittels gemischter linearer und logistischer Regressionsmodelle ausgewertet. **Ergebnisse** Der ST verringerte die Anzahl erkannter Übernahmegründe, führte jedoch zu höheren SAGAT-Werten und war zudem mit einer höheren Arbeitsbelastung verbunden. Die Arbeitsbelastung lag jedoch in beiden Bedingungen über dem Optimalwert. AGTORs verbesserten die Erkennungsleistung, unabhängig von der ST-Bedingung. Zwischen Tfs und Nicht-Tfs zeigten sich kaum Unterschiede und mit zunehmendem Alter sank das SA. Die Umgebung wurde von den Teilnehmenden insgesamt positiv bewertet. **Diskussion** Die vorliegende Studie verdeutlicht sowohl das Potenzial als auch die Herausforderungen von STs im teleoperierten Bahnbetrieb. Der ST resultierte in einer verbesserten Situationswahrnehmung. Gleichzeitig wurden in der ST-Bedingung jedoch weniger Übernahmegründe identifiziert, möglicherweise aufgrund von Aufgabenwechselkosten. AGTORs erwiesen sich als wirksame Maßnahme zur Verbesserung der SA, insbesondere bei der Erkennung der Übernahmegründe. Frühere Erfahrung zeigte keinen Effekt, könnte aber in zukünftigen Studien mit größeren Stichproben und realitätsgetreueren Szenarien verstärkt untersucht werden.

Abstract

Background Remote railway operations can improve efficiency and reduce costs, but operators must still intervene when onboard systems reach their limits, requiring high levels of situational awareness (SA). Passive monitoring can affect SA, especially under low workload conditions. This study on SA and workload in remote train operation explores the effects of a supplementary task (ST) intended to maintain engagement, potentially at the cost of SA. It also examines whether attention-guiding takeover requests (AGTORs) and train-driving experience can mitigate these effects. **Methods** A mixed design online study was conducted with 160 participants, including 28 active train drivers. Participants completed two blocks, one with an ST and one without, and were randomly assigned to receive either AGTORs or simple warning tone takeover requests. Each block featured four eight-second video scenarios in which participants had to identify the reaons for the takeover and assess additional information from a dynamic remote environment. SA was measured using takeover reason recognition and SAGAT questionnaires. Additionally, workload after each block and a number of additional variables were assessed. Data were analyzed using mixed linear and logistic regression models. **Results** Engagement in the ST reduced recognition of takeover reasons but increased SAGAT scores. The ST condition also led to higher workload scores, though levels were above optimal across both conditions. AGTORs significantly improved recognition performance regardless of the ST condition. No significant performance differences were found between train drivers and non-train drivers. Additionally, age was negatively associated with SA, and participants evaluated the remote environment favourably. **Discussion** The study highlights the benefits and challenges of STs in remote train operations. Working on an ST led to higher SAGAT scores, suggesting a benefit for maintaining perception-level SA. However, it also resulted in fewer recognized takeover reasons, potentially reflecting task-switching costs and reduced comprehension-level SA. A promising design direction to tackle this might be AGTORs, that, in this study, improved takeover reason detection without affecting overall SA. Previous experience did not show beneficial effects, but this might be due to the small train driver sample size. Future research might investigate the effects of experience, using a larger sample and more complex scenarios. Overall, this study provides a valuable insight into the challenges associated with remote train operation and offers solutions in the form of maintaining engagement through STs and compensating switching costs with AGTORs to support operators in takeover situations.

Contents

1 Introduction **1**

2 Background **2**

2.1 Status quo and future directions of railway systems . . . 2

2.2 Human factors aspects in automation . . . 5

2.2.1 Situational awareness . . . 6

2.2.2 Workload and task switching . . . 9

2.3 Challenges of increasing automation . . . 10

2.4 Remote operation . . . 12

2.4.1 Supplementary tasks . . . 14

2.4.2 TOR design . . . 16

2.5 Rationale for the study . . . 19

2.6 Hypotheses . . . 20

3 Methods **21**

3.1 Study design and procedure . . . 21

3.2 Measures . . . 23

3.3 Online-study setup . . . 27

3.3.1 The remote environment . . . 27

3.3.2 Supplementary task . . . 30

3.3.3 Scenarios . . . 31

3.4 Participants . . . 33

3.5 Data analysis . . . 35

3.6 Preregistration . . . 37

4 Results **37**

4.1 Control variables . . . 38

4.2 Situational awareness . . . 40

4.3 Workload . . . 45

4.4 Evaluation of the remote environment . . . 47

5 Discussion **48**

5.1 Limitations . . . 55

6 Conclusion **57**

References **58**

A Appendix **70**

A.1 Karolinska Sleepiness Scale (Åkerstedt & Gillberg, 1990) 70

A.2 DLR-WAT (Grippenkoven et al., 2018) 70

A.3 Telepresence questionnaire (Kim & Biocca, 1997) 71

A.4 System Usability Scale (Brooke, 1996) 72

A.5 Overview of all scenarios with corresponding reasons for the TOR. 73

A.6 Form to fill in the reason for the takeover 77

A.7 Participants' *n* for key variables in the study. 78

A.8 Correlation table of all variables used in the study. 79

A.9 Single SAGAT items' recognition rates and importance ratings 80

A.10 Results of the analyses of all statistical models 80

List of Figures

1	Driver's cabs of two generations of German high speed trains.	4
2	Model of situational awareness.	7
3	Theoretical impact of degrees of automation on several variables.	11
4	Exemplary setup of an RCC.	13
5	Study design and structure.	22
6	SAGAT form with exemplary answers selected.	24
7	Both versions of the monitoring state of the remote environment.	28
8	The operating state of the remote environment.	29
9	Both conditions of the VST.	31
10	Z-standardized values of four performance measures by age.	38
11	SAGAT distributions for the ST and no-task condition.	40
12	Main effects on SAGAT scores and distributions.	41
13	Interaction of ST condition and experience on SAGAT scores.	42
14	Percentages of correct detection for the eight takeover reasons.	43
15	Odds ratios regarding takeover reason recognition.	44
16	Main effects on workload scores and distributions.	46
17	Usability ratings for single items between both TOR groups.	47

List of Tables

1	Overview of the task distribution at different GoA	5
2	Possible STs for a TO.	15
3	Description of the sample.	34

List of Abbreviations

AGTOR	Attention-guiding takeover request
ATO	Automatic train operation
DB	Deutsche Bahn (German state railway operator)
ERTMS	European Rail Traffic Management System
ETCS	European Train Control System
OSF	Open Science Framework
RCC	Remote Control Center
RSI	Relative completion speed indicator
SA	Situational awareness
SAGAT	Situation Awareness Global Assessment Technique
ST	Supplementary task
SWT	Simple warning tone
TO	(Remote) Train operator
TOR	Takeover request

1 Introduction

National European rail systems are increasingly facing considerable structural and operational challenges (Broughton et al., 2024). Staff shortages in particular have become a major concern in the rail industry. In Germany, for instance, recent reports from Deutsche Bahn (DB) show that critical roles such as train drivers and signallers remain difficult to fill (Deutsche Bahn AG, 2024). Other challenges of the DB include ageing infrastructure, like interlocking systems from the 1960s and a rail network that often lacks the redundancy and flexibility necessary for resilient operations (Deutsche Bahn AG, 2024). This is also reflected in declining punctuality and passenger satisfaction, (Deutsche Bahn AG, 2024).

A growing emphasis on automation, such as fully automated train operations, reflects one strategy to address these issues (Digitale Schiene Deutschland, 2024). Highly automated trains, alongside advanced digital infrastructure, will reduce the number of required personnel by removing the train driver from the cab and enable more efficient use of the limited rail network, for example by allowing shorter headway distances (Grippenkoven et al., 2020).

While high levels of rail automation offer considerable benefits, they also present a number of challenges, particularly with regard to the transformation of operational roles (Broughton et al., 2024). In *automatic train operation* (ATO), the tasks and responsibilities of most railway personnel will shift, but this will be especially fundamental for train drivers, whose core duties and job characteristics will be redefined (Naumann & Thomas-Friedrich, 2024). This shift is caused by a transition from continuous manual train operation to situational, *remote train operation* (RTO), that entails a redistribution of responsibilities between the automated systems and human operators (Naumann & Thomas-Friedrich, 2024). While the automated train control system will manage routine operations, train drivers will act as *remote train operators* (TOs) from a control centre, instead of being on board the train. This means that they will serve as a fallback layer, resolving unforeseen or safety-critical situations that the systems cannot handle (Grippenkoven et al., 2020), such as unexpected obstacles on the track, medical emergencies among passengers or technical issues, such as sensor failures.

In such events, control of the train is transferred to the TO, who must quickly assess the nature of the issue, understand its implications and initiate appropriate countermeasures. However, resuming manual control can be particularly challenging if the TO is “out of the loop”, a condition marked by reduced *situational awareness* (SA) resulting from prolonged periods of passive monitoring and disengagement (Endsley, 2021). Consequently, a key challenge in

RTO is ensuring that the TO can quickly regain SA when manual control becomes necessary (Thomas-Friedrich et al., 2024a).

As part of the German Aerospace Center's *Teleoperation ATO* project[1], this thesis explores how job and system design can support the TO during takeover situations. Specifically, it examines how engagement in *supplementary tasks* (STs), that are intended to enhance job attractiveness and maintain engagement during periods of low demand (Thomas-Friedrich et al., 2024a), affects SA in takeover situations and overall workload. Furthermore, it investigates whether semantic and directional takeover-requests can help direct attention to the most relevant aspects of the remote environment first and thereby facilitate faster SA recovery. To investigate these questions, an experimental online study was conducted with 160 participants, 28 of whom were train drivers.

2 Background

To establish an understanding of the expected automation-induced changes to railway systems, the following section will provide an overview over the status quo of the railway system in Germany and an outlook on anticipated technological developments and their implications for the role and tasks of the train driver. Since railway systems in different countries vary in their configurations and degrees of complexity, the following section will only consider the German system. Given the limited research in the railway domain, relevant findings from other transport domains, especially automotive and aviation will be incorporated to provide a more comprehensive overview.

2.1 Status quo and future directions of railway systems

The current railway system in Germany comprises several occupational roles, each with distinct duties and responsibilities. At the administrative level, personnel such as CEOs, civil servants, and HR staff are responsible for strategic planning and the overall coordination of business operations under the oversight of the German government (Janicki, 2018).

On the operational side, employees can broadly be divided into two categories: infrastructure or onboard staff. The former includes traffic controllers, dispatchers, and construction

[1]For more information on the project, see https://www.dlr.de/de/ts/forschung-und-transfer/projekte/teleoperation-ato

workers, while the latter consists of train drivers, shunters, and onboard service personnel (Janicki, 2018).

Currently, the functioning of the railway system relies on the close collaboration between those managing train movements and those operating the trains (Janicki, 2018). Traffic controllers, often located in centralized signalling centres, are responsible for setting routes, managing switches, and signalling in accordance with schedules. Their role involves frequent communication with both train drivers and other controllers to ensure safe and efficient operations.

Train drivers on the other hand, need to make sure they drive their trains in a safe, punctual and energy saving manner according to their schedule (Janicki, 2018). Depending on the train protection system in use, they are required to actively confirm and follow signal aspects or deal with warnings. Furthermore, they continuously monitor locomotive parameters such as brake force, traction and overall vehicle condition, diagnosing and resolving technical or operational malfunctions if necessary (Janicki, 2018). Prior to departure, train drivers additionally perform train preparation procedures, such as brake tests and, in emergencies, they are responsible for coordinating passenger care and ensuring that safety procedures are followed.

Overall, train drivers bear significant responsibility for the safety risks that come with railway operations, transporting high amounts of people or cargo, while simultaneously being constrained to long braking distances. As in other transport domains such as aviation and the automotive industry (Salmon et al., 2005), human error contributes substantially to incidents in the railway sector (Evans, 2011). Most railway accidents can be traced back to either driver inattentiveness, such as passing signals at danger, or mistakes made by traffic controllers (Stene, 2018).

Automation efforts are aiming to reduce the number of human errors and increase rail network efficiency. While all railway staff will be affected by automation-related changes, operational roles in particular will experience a drastic change in responsibilities and work environment (Naumann & Thomas-Friedrich, 2024). One of the most current examples of this change is the implementation of advanced train control systems, notably the *European Train Control System* (ETCS), which forms a central component of the broader *European Rail Traffic Management System* (ERTMS) (UNIFE, 2021). ETCS enhances interoperability across national railway systems in Europe and facilitates continuous train monitoring, thereby enabling automated acceleration and braking. Furthermore, the system allows train operations to proceed independently of conventional trackside signals, which allows for shorter headway distances and increases network capacity (UNIFE, 2021).

Through these ongoing automation efforts, the roles and work environments of train drivers have already undergone significant changes compared to traditional, non-automated railway systems (Brandenburger et al., 2018). One indication of this transformation is the evolution of the driver's cab in German trains (see Figure 1), where numerous mechanical levers and analogue gauges have been reduced to few multifunctional displays and simplified interfaces.

(a) BR 103

(b) BR 411 (ICE T)

Figure 1. Driver's cabs of two generations of German high speed trains. Images created in Train Sim World 4 (Dovetail Games, 2023).

The changes brought about by increasing automation have been formalized in the IEC-62267 standard (IEC, 2009). It uses the *Grades of Automation* (GoA) categorisation system to classify four types of railway operation. For an overview over the differences between the GoA levels with regard to five key train driving tasks, see Table 1.

Numerous metro lines are already operating fully automated under GoA4, for example the entire Copenhagen metro network and metro lines in major cities like Honolulu, Rome or Lima (Hitachi Rail, 2025). The difference to mainline railway operations is mainly due to metro lines' inherent controllability advantages, since metro trains often run on secluded railway structures that allow most external influences to be controlled (Grippenkoven et al., 2020). On the other hand, mainline railways often consist of more complex networks, as well as heavier trains with longer braking distances, higher speeds, and higher passenger numbers per train (Stene, 2018). This significantly increases the safety risks and, consequently, safety requirements for ATO systems on mainline railways.

A notable exception is the Rio Tinto network in Australia, where a coal train has been upgraded to GoA4 operation (Wardrop, 2019). But this change to automated operation could only be implemented because of the area's vast, unpopulated landscapes and because the

Table 1. Overview of the task distribution at different GoA for five main train driving tasks. Adapted from IEC (2009).

Task	GoA 1	GoA 2	GoA 3	GoA 4
Accelerating and braking	Train driver	System	System	System
Monitoring guideway	Train driver	Train driver	System	System
Supervise train status	Train driver	Train driver	Train attendant or TO	System
Supervising passenger transfer	Train driver	Train driver	Train attendant or system	System
Management of emergency situations	Train driver	Train driver	Train attendant or TO	System

train does not carry passengers, thereby reducing safety requirements (Wardrop, 2019).
For most mainline railway operations, transitioning to GoA3 and GoA4 presents considerable challenges, both from a technical and a human factors perspective. However, solutions to human factors challenges can build upon a substantial body of research that has examined the impact of varying degrees of automation on human performance.

2.2 Human factors aspects in automation

One of the main aims of automation is to support a human in performing routine tasks, thereby reducing human workload and increasing overall system performance (Onnasch et al., 2014). To achieve this, automation can be applied at all stages of information processing, i.e. sampling information from the environment, integrating this information, making a decision and finally performing an action based on the decision (Parasuraman et al., 2000).
By reallocating responsibilities across these stages, automation offers several benefits. Supporting the human adequately may help maintain better performance by reducing task and subsequent cognitive workload, allowing the human to allocate more cognitive resources to other tasks or advanced information processing steps (Parasuraman et al., 2000).
These benefits depend on the reliable performance of the automation (Parasuraman et al., 2000), but real-world systems, particularly in traffic situations on roads or railways, will presumably never be completely reliable. Thus, when a system experiences technical problems or unforeseen incidents, human action will still be required to resolve the situation. This dynamic is associated with a number of issues, that have been described as *ironies of automation* (Bainbridge, 1983).

At the heart of these ironies lies the idea that, although automation is designed to reduce or eliminate human involvement, the human operator's performance becomes increasingly critical as automation advances (Bainbridge, 1983). This is because the most complex tasks are often the most difficult to automate. Therefore, in cases of imperfect automation, they remain the responsibility of the human operator. However, if these tasks occur very infrequently, the operator's manual and cognitive skills required for dealing with them can deteriorate due to a lack of practice (Bainbridge, 1983). Moreover, long periods of low demand may result in decreased vigilance and a loss of SA (Endsley, 2019).

2.2.1 Situational awareness

SA can be understood as an individual's ability to understand their surroundings and anticipate potential changes (Endsley, 2021). In a more formal definition by Endsley (1995), SA involves three levels, namely the perception of elements in the environment (level one), the comprehension of their meaning (level two), and the projection of their status into the near future (level three).

These three levels form the core of Endsley's (1995) SA model (Figure 2), and serve as a fundamental requirement for effective decision-making and action-taking. Performing actions, in turn, will alter the state of the environment, resulting in a continuous feedback loop (Endsley, 1995).

In addition to this core structure of the model, various external factors influence the development and maintenance of SA (Endsley, 1995). These can be broadly categorized into individual and system- or task-related factors. Individual factors are shaped by both innate abilities and prior experiences and include a person's motivations, expectations, cognitive processing abilities, and working and long-term memory capacity. They primarily influence SA through the allocation of attentional resources and working memory capacity, which determine what elements of the environment are perceived and how this information is integrated. Furthermore, according to the model, prior knowledge, expectations, and motivational goals play an important role in distinguishing relevant from irrelevant information. They facilitate the development of SA by enabling the formation and application of mental models and schemata derived from previously encountered or similar situations (Endsley, 2006).

System- or task-related factors include the level of cognitive demand imposed by the task, the design of the human-machine interface, and the overall complexity and level of automation within the operational environment (Endsley, 1995). These factors influence both the quan-

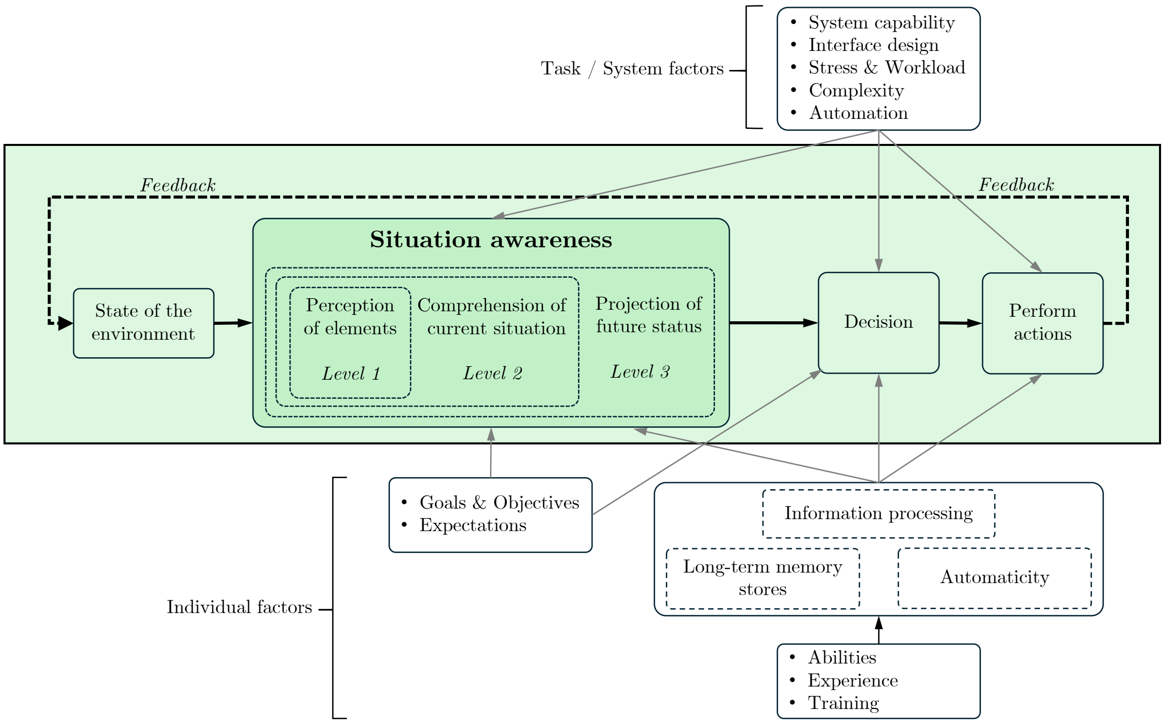


Figure 2. Model of situational awareness. Adapted from Endsley (1995).

tity and salience of information available to the operator, which shape attentional demands and workload distribution. Furthermore, automation in particular may influence all levels of SA by aiding in the perception of elements, facilitating their integration into a coherent situation assessment, and assisting in anticipating future states of the system. Furthermore, beyond SA, automation may assist in decision-making processes or even support or take over the execution of actions (Endsley, 1995).

In the context of train driving, the application of this model requires that drivers first perceive all relevant cues from within the cab and the external environment. Inside the cab this involves monitoring elements such as sensors, gauges and digital displays as well as any non-visual cues, such as sounds, vibrations or smells. Outside the cab, drivers must observe the track and the surrounding landscape. All of the perceived individual elements must then be integrated to form an accurate mental model of the train's current operational state and how this aligns with the expected or desired condition. In cases of deviation, for example, a reduction in braking power, the projection component of SA becomes critical, as the driver must anticipate potential implications like increased stopping distances or delays to the schedule.

The influence of additional individual and system-level factors becomes evident in such situations. Experts can draw on prior encounters with similar failures (Endsley, 2006), enabling them to recognize the issue more quickly than novices and implement effective countermeasures. Furthermore, if an onboard system detects the braking issue and issues a clear, timely alert, perceptual and integration workload may be reduced. This support then frees up cognitive resources compared to manual operation where failure detection relies entirely on the human operator (Endsley, 2006).

By nature, acquiring and maintaining SA is a dynamic process, because environments change constantly and so do the systems' and the humans' states (Endsley, 2015). Thus, SA is consistently updated by new and changing information. Furthermore, although the three levels of SA can be understood as forming an ascending order, they do not necessarily develop linearly (Endsley, 2015). For instance, experience may support understanding of the current situation and subsequently guide the perception of relevant elements in a top-down fashion. Many findings across various domains have indeed linked experience to enhanced SA (Endsley, 2006). This is largely attributed to a reduced working memory load in more experienced individuals, as advantages in long-term memory guide information search. Additionally, mental models and schemata help match the current situation to prior experiences and allow experts to perform actions automatically (Endsley, 2021).

In practice, a variety of challenges have been identified to affect the development and maintenance of SA (Endsley, 2021). One such challenge is attention tunnelling, which occurs when an individual focuses too narrowly on a single element of their environment, reducing their awareness of other potentially important aspects (Endsley, 2021). This is particularly likely when a particular stimulus is highly salient and captures attention in a bottom-up manner. Additionally, if a person must retain too many elements in their working memory simultaneously, their ability to integrate new information may be hindered due to limitations in cognitive capacity. A number of additional factors are known to interfere with effective SA development but are not directly relevant to this thesis and will thus not be discussed here (for a more comprehensive overview, see Endsley, 2021).

In the context of RTO, one of the most relevant challenges is the *out-of-the-loop effect*, whereby reduced engagement leads to a decline in SA (Endsley, 2021; Parasuraman et al., 2000). The implications of this effect will be discussed in greater detail in the following sections.

2.2.2 Workload and task switching

Another important human factors concept in the context of automation is *cognitive workload* (herein after referred to as workload). It refers to "the operator's allocation of limited processing capacity or resources to meet task demands" (Matthews & Reinerman-Jones, 2017, as cited in Hancock et al., 2021, p. 205). However, instead of workload drawing from a single undifferentiated resource as proposed in early models, newer theories, such as the *Multiple Resource Theory* (MRT; Wickens, 1980), conceptualize workload as being sustained by multiple distinct and individually limited resource pools. According to MRT, these resource pools are separated along four (partly nested) dimensions: stages of processing (perception/ cognition vs. response execution), modalities (visual vs. auditory), types of code (spatial vs. verbal), and visual channels (focal vs. ambient) (Wickens, 2008).

While MRT has been used to explain multi-tasking effects of working on simultaneous tasks, that share the same modalities, in the RTO context, additional tasks are more likely to be performed sequentially than simultaneously with monitoring. Switching from an ST to monitoring will, however, incur task-switching costs, arising from the need to reconfigure the cognitive system when switching between tasks (Shi & Bengler, 2022). This reconfiguration includes the task goal, the type of stimuli involved in the task, and the required responses (Vandierendonck et al., 2010). Unlike MRT, where interference is primarily determined by concurrent tasks drawing on the same sensory or cognitive resources, task switching theories emphasize that interference is driven by the degree of cognitive reconfiguration required when transitioning between tasks (Vandierendonck et al., 2010).

For instance, in an RTO setting, a visually demanding ST involving a structured search task may necessitate a goal-oriented, focused scanning strategy, which contrasts with the broad and continuous vigilance required for monitoring of a remote environment. This mismatch in attentional focus settings may hinder regaining SA quickly. As a result, operators may be slower to detect critical anomalies or may struggle to update their mental model of a train's current status.

Reducing the difference in cognitive demands between STs and monitoring tasks, or incorporating interface cues that facilitate attentional re-engagement, may help to mitigate switching costs (Vandierendonck et al., 2010) and preserve SA following a *take-over request* (TOR). Therefore, when designing future work environments such as RTO settings, tasks that are performed in close succession should ideally be aligned in terms of their cognitive demands to help minimize task-switching costs, support the rapid restoration of SA, and reduce overall operator workload (Shi & Bengler, 2022).

Both SA and workload are closely interconnected, as workload influences the attentional resources available to develop SA (Endsley, 2021). While cognitive overload can especially hinder the ability to effectively perceive important elements in the environment and make decisions, cognitive underload, caused by a low frequency of necessary engagement during routine performance, may result in disengagement, boredom, and fatigue, eventually leading to being out-of-the-loop (Endsley, 2021). Remaining out-of-the-loop for extended periods can deteriorate the operator's mental model of the system state. As a result, it may take longer for the operator to reorient and "get back into the loop" when re-engagement becomes necessary. This delay in regaining SA can be critical, particularly in time-sensitive or high-risk situations. Additionally, low engagement frequency at higher levels of automation may contribute to an overreliance on the automated system. (Endsley, 2021). This overreliance may lead to complacency, meaning operators overestimate the system's capabilities, which makes them insensitive to anomalies or system failures.

Taken together, the effects of low engagement, suboptimal workload levels and inadequate SA development suggest that, while increasing automation offers benefits, it also presents challenges.

2.3 Challenges of increasing automation

The effects of increasing degrees of automation are summarized in Figure 3, depicting the trade-off between the advantages, like improving routine performance and reducing workload and the disadvantages, namely losing SA and deteriorating performance in the event of a failure (Wickens et al., 2010). Each system will have an optimal point concerning the degree of automation (point *a* in Figure 3), where the advantages are maximized while the detrimental consequences remain within an acceptable range. However, there is no universal solution as to where this point will lie within a system (Onnasch et al., 2014).

For example, in a plane, the highest degree of automation might lead to pilots completely losing SA and understanding of the systems. This has indeed been the case in a number of severe aviation accidents in recent years, such as Air France Flight 447, in which the pilots were unable to regain control of the aircraft after an unforeseen outage of the systems, resulting in the loss of 228 lives (Müller & Drax, 2014). Conversely, in manufacturing, for example, automation might improve efficiency without the risk of catastrophic safety risks. Robotic arms on assembly lines, for instance, can operate with minimal oversight as errors

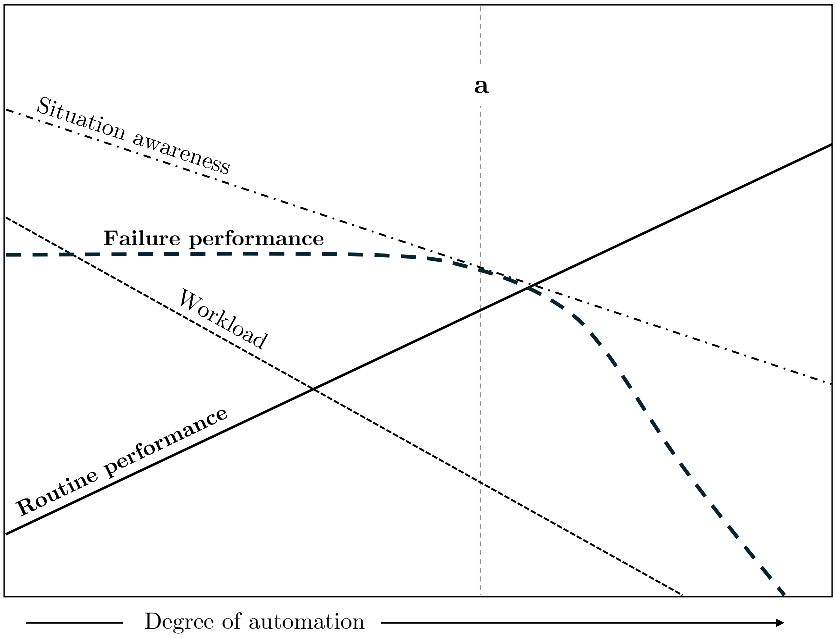


Figure 3. Theoretical impact of degrees of automation on several variables. Adapted from Wickens et al., 2010 and Onnasch et al., 2014.

typically result in downtime rather than loss of life.

In practice, with regard to automated driving, studies suggest that conditional automation, intended to enhance safety and comfort by allowing the system to handle routine driving tasks, can result in poorer handling of critical situations when human intervention becomes necessary (Weaver & DeLucia, 2022). Switching back from automated to manual driving has been found to lead to increased reaction times and greater lane deviation (Lu et al., 2021), reduced accuracy in judging the criticality of situations (Greenlee et al., 2022), and longer durations required to regain SA comparable to that during manual control (Vogelpohl et al., 2018). These findings might be explained by evidence showing that semi-autonomous driving was associated with reduced vigilance (Biondi et al., 2023; Greenlee et al., 2022), in some cases even after short periods of 15 to 45 minutes (Feldhütter et al., 2019).

In the railway domain, driving at GoA2 compared to GoA1 (the key difference being that, under GoA2, acceleration and braking is taken over by the train) resulted in deteriorating SA, fatigue and suboptimal workload (Brandenburger & Naumann, 2019b). Additionally, driving under GoA2 led to decreases in attention and vigilance, as well as worse takeover performance, indicated by inadequate handling of the brake after the train switched back to GoA1 (Du & Zhi, 2022; Spring et al., 2009). Therefore GoA2, while offering some operational benefits, entails many disadvantages from a human factors perspective (Brandenburger et al., 2021). These findings are especially relevant, since fatigue and decreasing vigilance have long been known to be key issues in all forms of train driving (Brandenburger et al., 2021; Dorrian et al., 2007; Filtness & Naweed, 2017; Sharp Grant, 1971). They can be attributed primarily to the often monotonous nature of the train driving task.

In the automotive sector, various countermeasures have been explored against these adverse effects of mid-level automation. These include time-based break reminders, human-centred interface design (Cabrall et al., 2019), the development of innovative HMI concepts for driver engagement (Wang et al., 2023) and driver state monitoring systems that detect signs of reduced attention and fatigue to trigger real-time interventions (Bosch et al., 2025; X. Zhang et al., 2017).

Another promising approach might be skipping continuous supervision-level automation operations altogether (Cabrall et al., 2019). For railway operations this would be achieved by switching to GoA3 and GoA4 operations.

2.4 Remote operation

In GoA3 rail operations, the primary driving tasks are handled by onboard ATO systems, eliminating the need for a train driver in the cab (Naumann & Thomas-Friedrich, 2024). However, since human intervention remains necessary when the ATO system encounters unexpected challenges, teleoperation has been proposed as a viable solution to support remote handling of such situations (Grippenkoven et al., 2020). In this setup, train drivers no longer actively control individual trains but instead passively supervise one or more trains in a designated area of the rail network remotely as TOs (Naumann & Thomas-Friedrich, 2024). They take over only when a TOR is issued by a train due to an unforeseen event that the onboard systems cannot resolve (Naumann & Thomas-Friedrich, 2024). In such cases, the TO must quickly assess the situation that triggered the TOR and enact countermeasures, for example by driving the train manually via a remote connection. Once the situation has been resolved, the TO can hand over control to the onboard system again (Naumann & Thomas-Friedrich, 2024).

In initial concepts for RTO, the TOs are stationed in a *remote control centre* (RCC), where they work in teams alongside dispatchers, signallers, and other operational staff (Brandenburger & Naumann, 2019a). An example RCC configuration, initially developed by Brandenburger and Naumann (2019a) and adapted to the current study, is shown in Figure 4.

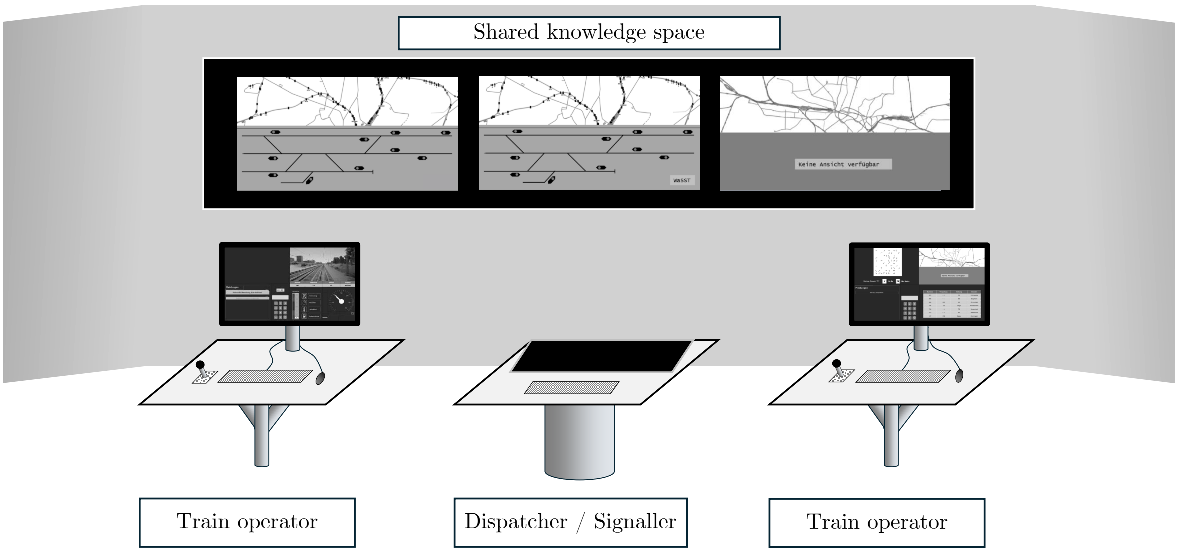


Figure 4. Exemplary setup of an RCC. Adapted from Brandenburger and Naumann (2019b).

Advances toward GoA3 operation on mainline railways have been made in recent years. In Germany, for example, a real-life trial is underway on the S-Bahn Hamburg network (Deutsche Bahn AG, 2025), where one line is operating at GoA3. However, due to the novelty of the operation and legal requirements, a safety driver is still present in the train cab. Similar projects have been implemented in other regions (see, for example, Masson et al., 2019; Schöne et al., 2023), reflecting a growing interest among railway operators in ATO. However, while research into technical challenges has already provided a number of findings, research into the human factors challenges imposed by GoA3 and GoA4 operations remains limited. Initial findings from a series of studies suggest that working at GoA3 level leads to reduced fatigue and workload closer to optimal levels than at GoA2 (Brandenburger et al., 2021). Even though these findings are encouraging, they may be partly due to the novelty of the task and remote environment, which could temporarily increase engagement, and the relatively short duration of the experimental trials. Further research is needed to better understand the long-term effects of this type of operation on workload, vigilance and SA.

From a technical perspective, to enable effective remote intervention, several prerequisites must be fulfilled. First, onboard automation systems must provide a high degree of reliability in handling routine tasks and operational scenarios autonomously (Schöne et al., 2023). This is essential to prevent frequent TORs that overwhelm the TO and disrupt overall network performance. Additionally, reliable sensors and AI-assisted analysis of camera data are needed to monitor the train and its surroundings for obstacles (Grippenkoven et al., 2020). This data and the results must be provided via a continuous, stable and low-latency communication link between the train and the RCC at all times (Masson et al., 2022).

From a human factors perspective, several additional requirements must be addressed to ensure safe and effective RTO. The type, amount, and timing of information presented to the TO during both routine supervision and takeover events must be carefully designed to avoid mental overload (Hely et al., 2015). This can be supported through an HMI design that structures the necessary data in ways that facilitate rapid SA development (Naumann & Thomas-Friedrich, 2024).

Moreover, the supervisory nature of GoA3 operations can lead to task monotony (Brandenburger & Naumann, 2018), increasing the risk of reduced vigilance, and deteriorated SA due to the TO being out-of-the-loop. To mitigate these risks, it is not sufficient to only optimize the design of the remote control workstation. Instead, the overall job design of the TO must also be taken into account. This includes ensuring cognitive engagement and supporting the development and maintenance of SA even during extended periods of passive monitoring (Naumann & Thomas-Friedrich, 2024). One way to achieve this, could be the introduction of STs for the TO.

2.4.1 Supplementary tasks

STs refers to any activity that a TO performs during routine operations when no train has issued a TOR (Thomas-Friedrich et al., 2024a). They may be considered measures of job enlargement and enrichment, which make the work more varied and increase its attractiveness by offering greater responsibility, challenges and meaning (Naumann & Thomas-Friedrich, 2024). Furthermore, if a TO is able to fulfil several roles simultaneously, fewer personnel are required, thereby enhancing the RTO's overall efficiency (Grippenkoven et al., 2020). Possible STs for a TO have been elaborated as part of the *Teleoperation ATO* project and include other train driving tasks, training and education or operational tasks (Thomas-Friedrich et al., 2024a). For a full list of the proposed STs, see Table 2.

Besides these operational benefits of introducing an ST to the TO's work, there is a number of associated risks. As cognitive processing abilities are limited (see Section 2.2.2), introducing an additional task will increase mental workload (Hancock et al., 2021). This is particularly relevant when an ST is performed at the same time as the main task and draws from

Table 2. Possible STs for a TO. Adapted from Thomas-Friedrich et al. (2024).

Type of Activity	Possible STs
Train driving tasks	• Remote-controlled shunting and delivery runs • Routine track observation • Mutual support in high workload situations
Training and education	• Professional development through simulation tasks, driving dynamics training, or technical courses
Operational tasks	• Taking over dispatching tasks • Personnel planning • Other administrative tasks • Supporting fault management or train maintenance • Training other employees • Assignment of delay reasons and creation of incident reports • Preparation of operational and construction instructions • Support in timetable and capacity management

the same cognitive resources, bearing the risk of exceeding processing capacity limits. On the other hand, if STs are performed sequentially, this introduces the risk of task-switching costs, which may impair the development of SA and subsequently degrade return-to-manual performance (Shi & Bengler, 2022).

A multitude of studies have observed the detrimental effects of STs on manual car driving (Ferdinand & Menachemi, 2014). More recently, STs have also attracted significant attention in the context of conditionally automated car driving (B. Zhang et al., 2019). For instance, working on a cognitively demanding ST led to an increased initial reaction time to a TOR (Amre & Sun, 2021; Choi et al., 2020; Weaver & DeLucia, 2022), later fixation on the road (Weaver & DeLucia, 2022) and a higher steering angle variance (Choi et al., 2020).

Interestingly, while many studies suggest that STs have a detrimental effect on driving performance, some report the opposite. For instance, working on a cognitively demanding n-back task led to improved longitudinal vehicle control after a TOR (Lu et al., 2021). One possible explanation is that engaging in an ST may help counteract the monotony and reduced vigilance associated with mid-level automation. This interpretation is supported by findings that participants who performed various tasks during conditionally automated driving reported lower fatigue levels and experienced fewer vigilance decrements than during manual driving without an ST (Miller et al., 2015). In the rail domain, a high-demand ST implemented as a mathematical speed restriction task even resulted in reduced driving errors (Dunn & Williamson, 2012).

Thus, if implemented appropriately, STs may serve as a strategy to maintain TO engagement, stabilize workload closer to optimal levels, and help mitigate fatigue (Thomas-Friedrich et al., 2024a). For an ST to be effective, however, it must strike a careful balance: it should be sufficiently engaging to sustain attention, yet not overly demanding, causing cognitive overload or impairing SA.

In line with task-switching theories, the potential impact of STs should not be evaluated based on a single defining feature such as sensory modality (Shi & Bengler, 2022). For instance, both passively monitoring the remote environment and performing a visual search task engage the visual channel, yet they may differ substantially in their effects on SA. While the former may support SA by aiding in the continuous maintenance of a mental model of the environment's state, the latter may impair SA by diverting attention away from task-relevant cues (Shi & Bengler, 2022). This indicates that the cognitive demands and attentional requirements of a task are likely more meaningful predictors of potential interference than modality alone, particularly in settings where tasks are performed sequentially rather than simultaneously (Shi & Bengler, 2022).

In summary, while STs offer benefits, such as reducing monotony and fatigue, they also introduce challenges, including task-switching costs and out-of-the-loop effects that can deteriorate SA (Xu et al., 2024). Thus, since STs are likely to be implemented due to their operational advantages (Thomas-Friedrich et al., 2024a), strategies that facilitate the rapid re-establishment of SA following a TOR are needed. One promising approach is to design the TOR itself to actively support efficient SA recovery.

2.4.2 TOR design

In semi-automated driving, a TOR is a means of indicating a transition of control between human and system (Colley et al., 2021). It may be initiated by the system or by the human, and either be scheduled or unscheduled. Additionally, the TOR may either mark the immediate transfer of control, or allow for a gradual transition (Walch et al., 2015).

TORs are already used in trains. For example, if a train in Germany is operating under GoA2 rules using the automated train protection system *LZB* and a switch back to the conventional GoA1 system *PZB* is announced, this will be signalled by an acoustic and visual alert and the driver will be required to acknowledge it by pushing a toggle switch (DB InfraGO AG, 2009).

Recent advances in automated car driving have led to extensive research on two key aspects of TORs: the time needed to complete the takeover and designing TORs to support the human driver in becoming alert and regaining SA as quickly as possible. Regarding the necessary time to takeover completion, studies on conditionally automated car driving have proposed takeover timeframes ranging from three to ten seconds, depending on factors such as the driver's current state and the surrounding traffic conditions (Deng et al., 2024).

However, these timeframes may not be directly applicable to the railway domain, because in contrast to cars, in critical situations trains will automatically transition into a safe state by initiating an emergency brake (Naumann & Thomas-Friedrich, 2024). As a result, if a TO exceeds the takeover timeframe, there is no immediate risk of collision or loss of control, since no active longitudinal or lateral driving input is required. Nevertheless, it is important for a TO to establish SA effectively and quickly in the event of a safety-critical situation to minimize operational disruptions and ensure timely intervention (Thomas-Friedrich et al., 2024a). For instance, in cases such as passenger medical emergencies, a swift response may be life-saving and prompt action in response to obstacles on the tracks could prevent derailments of oncoming trains. Thus, the level of support that the design of a TOR provides to the TO in establishing SA is also a concern in RTO (Thomas-Friedrich et al., 2024b).

The design of TORs can be oriented along two main criteria. The first is the modality in which the TOR is presented, and the second is the level of abstraction of the TOR. With regards to modality, TORs might be presented as auditory, visual, tactile, or multimodal stimuli (B. Zhang et al., 2019). Research from TORs in conditionally automated car driving shows that using multimodal stimuli compared to unimodal ones can reduce takeover time (Huang & Pitts, 2022). This is in accordance with the MRT (Wickens, 1980), which predicts that presenting information through different modalities than the ones required by the primary task is expected to reduce interference and support performance. For example, if a driver needs to detect an obstacle on the road, a visual TOR may interfere with the execution of the task, but an auditory alert may not. Building on this idea, the use of multimodal TORs may be useful, because they introduce informational redundancy (Colley et al., 2021), enabling the system to better accommodate varying driver states and concurrent activities during the drive (Yoon et al., 2019).

If a TOR is understood as a way of transmitting information to the driver, enhancing the amount and level of detail of that information could prove useful in supporting SA development (Liu et al., 2023). In car driving and RTO, a promising approach would be to use informationally enhanced auditory TORs to convey system states or takeover situation details, as they do not rely on the visual channel and are therefore less likely to interfere with the visual monitoring tasks.

One method of enriching auditory TORs is the use of semantic information, for example using speech, instead of *simple warning tones* (SWTs) (Liu et al., 2023). In two studies, investigating the acceptance of different TOR formats in conditionally automated car driving, speech-based alerts were rated more favorably than generic tones and were perceived as more intuitive and useful (Bazilinskyy et al., 2018; Brandenburg & Epple, 2019). Another study found similar results, but reported that semantic TORs led to slower reaction times (Chai et al., 2024).

A potential limitation of these implementations of semantic TORs lies in how speech was used. In the cited studies, verbal alerts were limited to basic commands such as "Danger ahead; please take over" (Chai et al., 2024) or "Please take over!" (Bazilinskyy et al., 2018). Such phrasing may have imposed cognitive demands on participants, who had to listen to, comprehend and interpret the message, but did not contain information supporting situation understanding.

Another approach of enhancing TORs consists of using directional, unimodal stimuli (Liu et al., 2023). These are designed to either convey information about the state of the system or environment, direct the driver's attention towards the most critical elements of the visual scene, or suggest a driving action. The primary aim of such TORs is to support the development of level two and three SA, rather than merely prompting the driver to perceive environmental elements (Liu et al., 2023). For example, when the position of an obstacle on the road that needed to be detected following a TOR was indicated using LED strips and vibrations as directional cues, participants showed fewer lateral lane deviations and fewer collisions compared to non-directional cues, although their reaction times increased (Gruden et al., 2022). In other studies, an auditory directional TOR led to less accidents and similar (Chen et al., 2023) or better reaction times (Huang & Pitts, 2022).

Only few studies have examined the combined use of directional and semantic information in TORs (Liu et al., 2023). These *attention-guiding TORs* (AGTORs; Chen et al., 2023) may be particularly effective in supporting the development of level two SA by guiding the operator's attention toward the most critical aspects of the scene first, thereby facilitating efficient attention allocation (Liu et al., 2023). This approach appears especially promising in the RTO context, where targeted guidance could help the TO prioritize essential information, especially after long periods of passive supervision or involvement in STs. Moreover, while the integration of semantic content may slightly increase reaction times due to additional cognitive processing demands, this trade-off is less critical in RTO than in car driving, as such delays are unlikely to result in immediate safety-critical consequences.

2.5 Rationale for the study

This study extends existing research on STs and TOR-design by exploring their relevance in the context of RTO, an area that has received little attention in human factors research. It addresses the differences between conditionally automated car driving and RTO, where TOs are physically separated from the train, have limited control authority, and receive less immediate and less multimodal system feedback. In addition to the inherent differences between trains and cars, such as the continuous, track-bound movement of trains and higher safety requirements, situational remote control in RTO fundamentally alters the dynamics of monitoring the environment and maintaining SA. This is further highlighted by the fact that STs are typically performed sequentially rather than simultaneously, unlike in conditionally automated car driving. Together, these differences limit the transferability of findings from car-based automation to RTO and underscore the need for domain-specific research, instead of relying on the direct application of car-related findings to rail. Furthermore, research on SA in rail automation is limited, especially at GoA3 and GoA4 and often includes small sample sizes. In contrast, this study uses a comparatively large and diverse sample.

Lastly, the study examines the novel use of AGTORs to support rapid SA buildup after attention has been diverted from monitoring trains and the environment. Given the risk of TOs becoming out-of-the-loop, AGTORs offer a promising approach to mitigate the adverse effects associated with GoA3 and GoA4 operations.

2.6 Hypotheses

Based on this rationale, the following hypotheses will be tested in this study:

Supplementary task

Performing a demanding ST is hypothesized to deteriorate SA in subsequent takeover situations due to the cognitive costs associated with task switching. This is indicated by differences in *Situation Awareness Global Assessment Technique* (SAGAT; Endsley, 1988) scores (for further explanation of the SAGAT technique, see Section 3.2). At the same time, due to greater cognitive demands, it is also expected to increase overall workload. Therefore:

- **H1a**: Participants achieve lower SAGAT scores after performing an ST compared to when no ST is performed. *(Main effect of ST on SAGAT scores)*
- **H1b**: Participants less frequently recognize takeover reasons after performing an ST compared to when no task is performed. *(Main effect of ST on takeover reason recognition)*
- **H2**: Participants report higher workload values after performing an ST compared to when no ST is performed. *(Main effect of ST on workload)*

TOR condition

AGTORs help to focus attention on the most critical element of the scene, i.e. the takeover reasons, thus they are expected to increase the rates at which these are detected. They are expected to be particularly beneficial when an ST has been performed before, because they help to mitigate task-switching costs. Therefore:

- **H3**: Participants more frequently recognize takeover reasons in the AGTOR compared to the SWT condition. *(Main effect of TOR condition on takeover reason recognition)*
- **H4**: The improvement in takeover reasons recognized due to the AGTORs is more pronounced when an ST was previously performed. *(Interaction of TOR condition and ST on takeover reason recognition)*

Driving experience

Train driving experience is hypothesized to facilitate restoration of SA and reduce workload due to transfer effects from previous experiences and domain-specific knowledge. AGTORs may help to offset this advantage of train drivers by providing greater support for non-train drivers. Therefore:

- **H5a**: Train drivers obtain higher SAGAT scores than non-train drivers. *(Main effect of driving experience on SAGAT scores)*
- **H5b**: Train drivers more frequently recognize takeover reasons than non-train drivers. *(Main effect of driving experience on takeover reason recognition)*
- **H6**: The improvement in takeover reasons recognized facilitated by the AGTORs is greater for non-train drivers compared to train drivers. *(Interaction effect of driving experience and TOR condition on takeover reason recognition)*

3 Methods

To test these hypothesized effects, a novel remote environment was developed and implemented in an experimental online study. The following section describes the study design, procedure, sample and the online remote environment and its development.

3.1 Study design and procedure

Three independent variables were used in the study, of which the TOR condition (SWT vs. AGTOR) and experience (non-train drivers vs. train drivers) were between variables and the ST condition (no-task vs. ST) was a within variable. This resulted in a $2 \times 2 \times 2$ mixed design. The design and structure of the study can be found in Figure 5.

At the beginning of the online experiment, participants were provided with essential information regarding the study and data protection. They could only proceed if they confirmed that they had read and understood the information, were over 18 years of age, had a stable internet connection, and were able to listen to audio. Participants then completed demographic questions, including age, gender, and railway-related experience (see Section 3.2). If they indicated that they worked for a railway company or were active train drivers, they were asked additional questions to describe their professional role in more detail.

Following this, participants were guided through a comprehensive introduction to the experimental environment and tasks. This explanatory section, presented via text and a narrated voice-over, covered the structure and function of the remote environment, the ST, handling of TORs, and the subsequent questionnaires. It also included one test trial for the ST and a

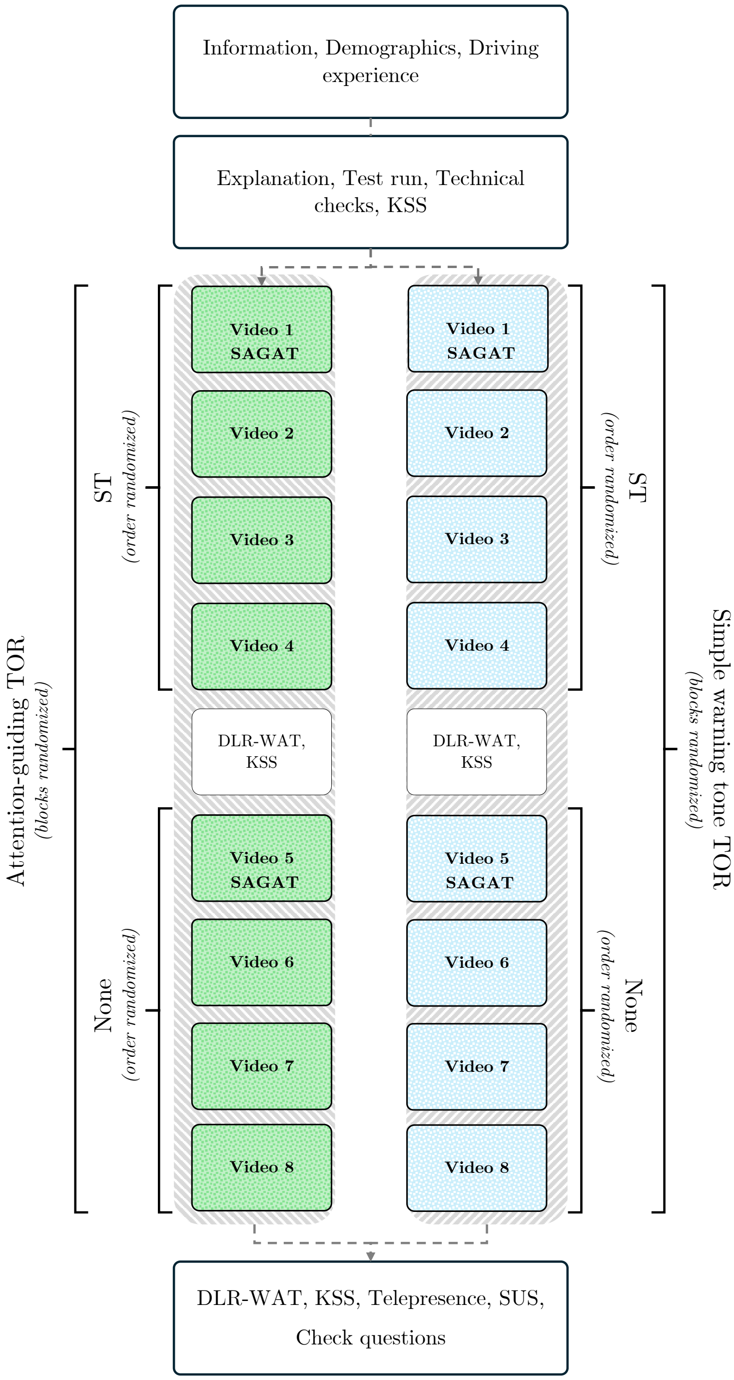


Figure 5. Study design and structure.

sample TOR scenario with a SAGAT questionnaire to familiarize participants with the procedure. Finally, participants were required to complete an auditory check (listening to words and identifying which one was presented) and a screen resolution check (indicating whether a small cross appeared on the left or right half of the screen).

After completing the instruction phase, participants were required to answer seven comprehension questions about the environment and correct procedures following a TOR. Feedback was provided for each response, and participants could choose to repeat the instructional section. Upon continuing, participants were randomly and secretly assigned to one of the two TOR conditions and began the first experimental block. The order of the blocks was randomized across participants. Each block included a brief explanation and introduction, that was designed to encourage participants to adopt the role of a TO.

An experimental block comprised four trials. A trial consisted of either the ST or no task, both lasting approximately one minute. In the no-task condition, an attention check appeared around 30 seconds into the trial, that required participants to press either the Y-key or N-key to proceed. After approximately one minute, a TOR was issued, followed by a short video sequence of the remote environment including the takeover scenario. This was then followed by either SAGAT questions (once per block) or simple questions about the reason for the TOR (three times per block).

Upon completing the first block, participants filled out workload and fatigue questionnaires. They then proceeded to the second block, which included the condition not presented in the first block (either ST or no-task). After the second block, participants completed workload, fatigue, telepresence, and usability questionnaires, and answered control questions. Finally, participants could enter a lottery to win one of five €20 gift cards. The study concluded with a thank-you message for their participation.

3.2 Measures

Situation awareness

SA was assessed threefold in this study. First, in three of the four takeover scenarios per block, participants were asked, after watching the video sequence, to identify what they believed had triggered the TOR. This approach was chosen because correctly identifying the cause of the TOR was considered essential for implementing appropriate countermeasures in a real world RTO setting and indicative of level two SA (comprehension of the situation). It involved a two-step response: first, selecting the area in which they believed the reason was located (camera view, parameters, or notification area) and second, choosing the specific reason from a predefined list. The form to choose the reason, can be found in Appendix A.6. To reduce the likelihood of correct guesses, the list included distractors that were never correct. Responses were only coded as correct if the specific reason was accurately identified. A total of eight takeover scenarios were presented, allowing for a maximum of eight correct identifications.

Secondly, instead of the simple takeover reason form, participants were presented with a SAGAT (Endsley, 1988) form once per block. It was shown as a blanked-out image of the remote environment with a total of 16 input forms at locations where information was previously present (see Figure 6). The input forms primarily used drop-down menus, each containing one correct and several incorrect response options. An additional "I don't know" option was provided to discourage guessing. For the speedometer and time items, participants had to enter numerical values. Each of the 16 SAGAT items was coded as either correct or incorrect, and the proportion (in %) of correct responses was calculated to yield an overall SAGAT score. Furthermore, in the SAGAT forms, participants could select the assumed reason for the takeover from a drop-down menu, which was recorded as part of the takeover reason recognition variable.

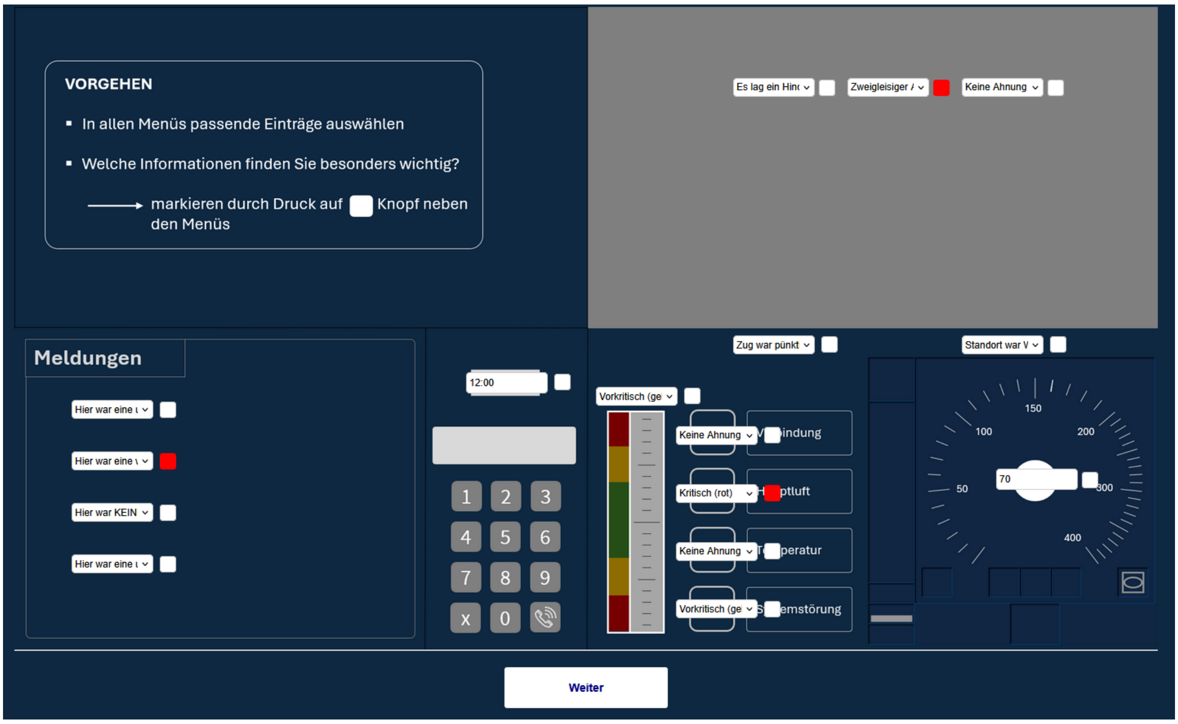


Figure 6. SAGAT form with exemplary answers selected.

The SAGAT method was assumed to reflect level one SA by indicating what participants had perceived from the available information in each scenario.

Participants were also asked to rate the importance of each of the 16 information elements in the SAGAT forms by checking a box next to each input field (see Figure 6). This measure was included to account for the varying relevance of the information presented. A binary variable was obtained for each element, with zero indicating that the information was not considered important and one indicating that it was.

Finally, subjective SA was assessed by asking participants to rate their confidence in having understood the scene on a scale from 0 to 10, with higher values indicating greater confidence in their SA. Out of these ratings a mean subjective SA score was calculated for each block. However, subjective SA was not assessed after the SAGAT trials for consistency reasons and due to technical implementation challenges.

Workload

Following each block, workload was measured using the *DLR-WAT* questionnaire (Grippenkoven et al., 2018). In contrast to widely used workload questionnaires, such as the *NASA-TLX* (Hart & Staveland, 1988), that merely distinguish between varying degrees of workload presence, the DLR-WAT allows for a finer assessment of workload across the range from sub-optimal to optimal to above-optimal levels. It comprises a total of eight scales. Six of these scales (*workload due to information intake, workload due to knowledge retrieval, workload due to decision-making, motor and physical workload, workload due to time pressure*) range from "1 - very heavily underloaded" to "200 - very heavily overloaded) with 100 being the optimum workload level. Additionally, *frustration with the task* is measured on a scale ranging from 100 to 200, with the absence of frustration (100) being regarded as the optimum. In a similar way, *task accomplishment* is measured on a scale from 1 to 100 (optimal), because there is no over-accomplishment. The complete scale (in German language) can be found in Appendix A.2.

For the confirmatory analyses, a mean score was formed using the scales *workload due to information intake, workload due to knowledge retrieval, workload due to decision-making* and *workload due to time pressure and effort*. The remaining scales were not included in the score due to their irrelevance to the situation (*motor and physical workload*) or because of differing scales that did not allow both over- and underload to be measured (*frustration* and *task accomplishment*). However, these scales are evaluated as part of the exploratory analyses.

Additional measures

Several additional measures were collected to gain a more comprehensive understanding of participants' characteristics and to evaluate the novel remote environment. Socio-demographic variables included age, gender, and country of residence. Participants were further asked to indicate their railway-related experience using the following self-constructed scale:

1 - No previous knowledge

2 - Exclusively knowledge from a passenger's perspective

3 - Advanced knowledge (e.g. acquired through simulation games)

4 - Professional knowledge: I work for a rail transport company, but not in transport operations

5 - Professional knowledge: I work for a railway company in driving operations

Participants who selected four or five on the scale were asked to provide additional details about their professional role. Individuals were classified as train drivers for all subsequent analyses only if they indicated that their primary responsibility involved operating locomotives or that they held a valid train driving license and accumulated at least 100 hours of train driving per year.

Since fatigue has been shown to affect SA (Endsley, 2021), participants were asked to state how awake they felt before the experimental part started, after the first block and after the second block using the *Karolinska Sleepiness Scale* (KSS; Åkerstedt and Gillberg, 1990). The KSS is a one item questionnaire with a ten-point verbally anchored scale ranging from "10 - extremely sleepy, falls asleep all the time" to "1 - extremely alert", with each step being verbally anchored. The complete scale (in German language) can be found in Appendix A.1.

In the last block of the experiment, the usability of the teleoperation environment was measured using the *System Usability Scale* (SUS; Brooke, 1996). Usability was assessed, because of the novelty of the environment and the limited user feedback available during its development. The SUS consists of ten items rated on a scale from 1 to 5, used to calculate a global evaluation score of the environment, following the formula provided by Brooke (1996):

$$SUS_{score} = 2.5 \cdot \Big(\sum_{i \in \{1,3,5,7,9\}} (Item_i - 1) + \sum_{j \in \{2,4,6,8,10\}} (5 - Item_j) \Big)$$

This resulted in a range from 0 to 100 with 100 being the best possible usability score. The complete scale (in German language) can be found in Appendix A.4.

To assess the extent to which participants felt part of the environment and immersed in the simulation, a short questionnaire on subjective telepresence (Schloerb, 1995) was implemented at the end of the study. The scale was taken from Kim and Biocca (1997) and adapted to fit the context of the study. All of the eight items could be answered on a five point Likert scale ranging from "0 - not at all applicable" to "5 - completely applicable". Because of the scarcity of research on telepresence in remote operation contexts, this scale was used to measure immersion exploratory. A mean score across all items was calculated, with higher values indicating higher levels of immersion. The full telepresence scale (in German language) can be found in Appendix A.3.

Finally, to assess the data quality, control questions were implemented at the end of the study. Participants were asked to rate the diligence and fun with which they worked on the experiment on a scale of "0 - very little" to "1000 - very much". They were informed that their responses would not affect their chances of winning the lottery. Additionally, participants were asked to rate the difficulty of the ST on a scale from 1 to 10, with higher scores indicating higher subjective difficulty. ST performance was measured as the percentage of correct responses across all ST trials. Participants could also state whether they were engaged in any other activity during the no-task condition in an open text field at the end of the study.

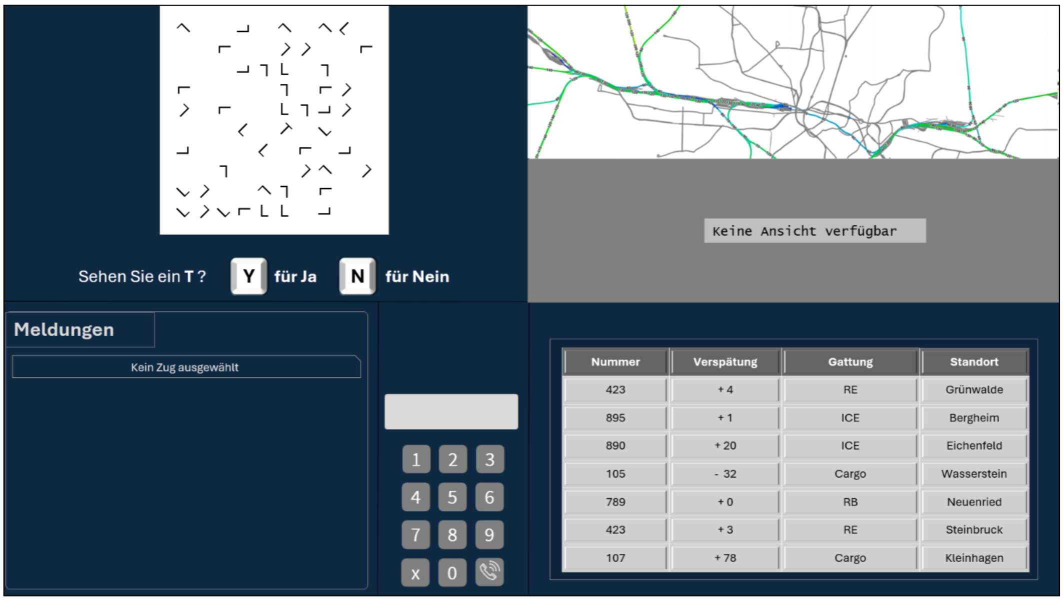
The total time that was needed to finish the experiment as well as the time needed to acknowledge the attention checks in the no-task condition were logged to be used as an exclusion criterion.

3.3 Online-study setup

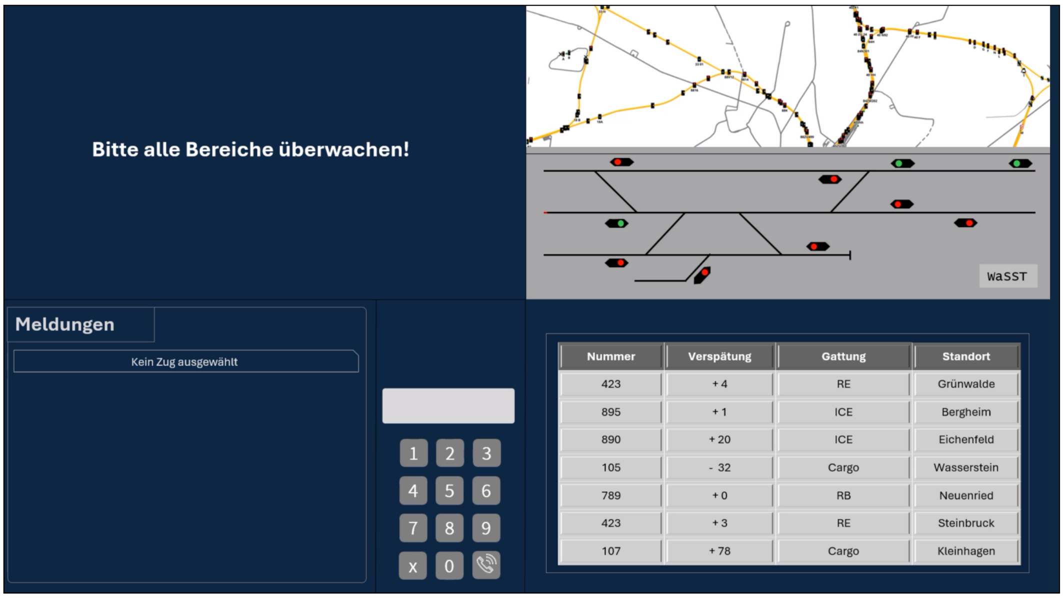
3.3.1 The remote environment

The remote environment was adapted from Brandenburger and Naumann (2019a) and modified to fit the present study. It was presented to participants either as static images or as animated video sequences. However, the difference was not clearly visible to participants, because apart, from the ST, no element of the environment was interactive. This was done intentionally, to increase immersion and create the illusion of a somewhat dynamic and unpredictable environment.

The environment featured two distinct states. The first, herein after referred to as the monitoring state, was visible whenever no TOR was issued. Both versions of this state (in either the ST and no-task conditions) are shown in Figure 7. The display was divided into four quadrants: the upper left quadrant presented either the ST or, in the no-task condition, a static instruction to monitor the environment. The upper right quadrant showed a static picture during the ST trials to minimize distraction. In the no-task condition, it displayed animated, simplified signal box activities designed to enhance immersion and focus on the environment without being overly engaging. The lower right quadrant showed a list of trains currently operating in the monitored sector, which was periodically updated in the no-task condition. In contrast, no animations were present during the ST, as participants were expected to focus on the ST. The lower left quadrant contained a notification panel, which remained empty in both conditions.



(a) ST condition



(b) No-task condition

Figure 7. Both versions of the monitoring state of the remote environment.

When a TOR was issued, the environment transitioned into another state, herein after referred to as operational state. This state always had the same layout, regardless of the ST condition. Now, the upper left quadrant was blanked out. The upper right quadrant displayed a front-facing camera view from the train, showing an eight second long video clip of a railway environment (see Section 3.3.3 for the rationale behind the video duration and the content of the videos). Directly below the camera view, a table provided information about the observed train, including its number, delay, type, and current location, that was structured similarly to the train list in the monitoring state. The bottom right area presented several parameters, namely a bar graph for friction status, icons for data link status, wheel temperature, air brake pressure, and general system faults, as well as a speedometer and a digital clock. These parameters were chosen to represent important real-life indicators of trains, while remaining comprehensible to non-train drivers. The lower left area continued to display the notification panel, but now included up to five message entries. The TOR alert was always present as a notification, accompanied by either relevant messages (e.g., a passenger triggering the emergency brake) or irrelevant messages (for instance a message displaying that the train was started up properly) shown below it. An example of the operational state is shown in Figure 8.

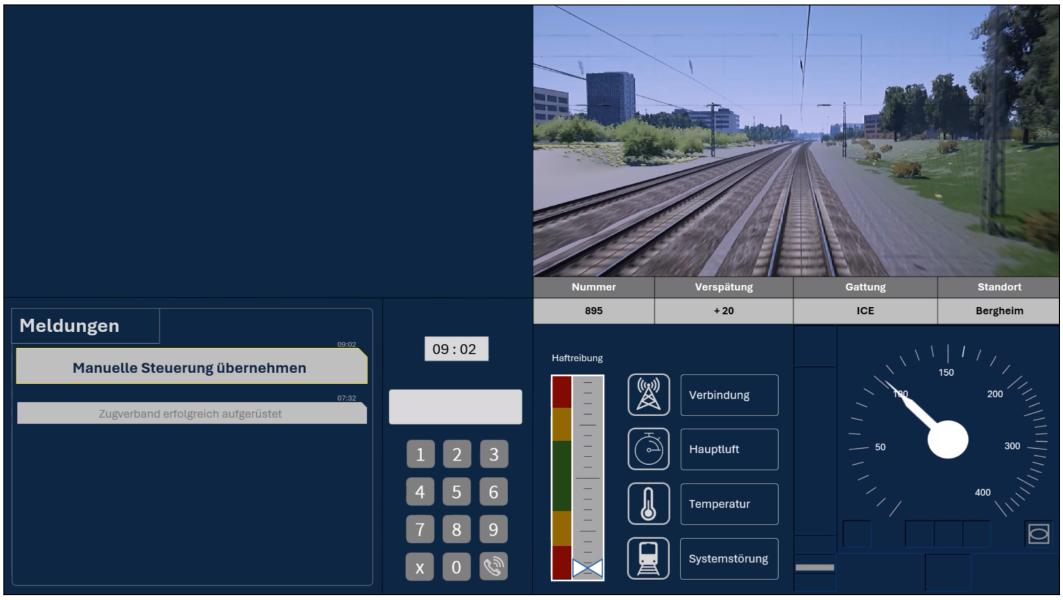


Figure 8. The operating state of the remote environment.

The online study was configured to ensure that the environment was consistently displayed at a resolution of exactly 1400 × 788 pixels, independent of the participant's actual screen size. This standardization aimed to prevent visibility advantages that could arise from higher screen resolutions, thereby maintaining experimental control across devices. Participants' screen size was logged, and those who did not meet the required resolution were excluded from the dataset afterwards.

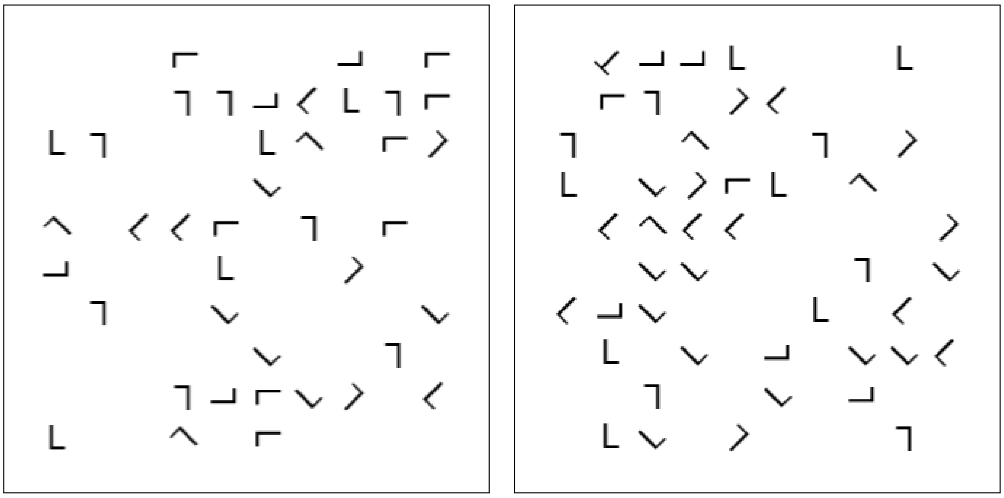
TORs were issued either as a SWT or a verbal AGTOR. The SWT consisted of alternating simple triangle shaped tones with 890 and 1760 Hz frequency and 1.5 seconds length. This alarm tone design was modelled after Yun and Yang (2020) and modified to fit the length of the AGTOR alarm. The AGTOR alarm was created using synthetic voices from the Elevenlabs Web application (ElevenLabs, 2025). The voice was imitating a male voice to resemble alarms currently used in trains to notify the train driver of irregularities, for instance missed acknowledging of train protection. The following AGTORs were created in this manner: "Takeover. Focus on camera", "Takeover. Focus on parameters" or "Takeover. Focus on messages", depending on where the takeover reason could be found. To enhance immersion, additional short audio clips simulating brief conversations with a colleague were included at the end of the takeover scenarios. All alarm and conversation sound files used in the study are available in the *Open Science Framework* (OSF) repository of the study (Schackmann, 2025a).

The experimental part of the online study was designed using lab.js (Henninger et al., 2021). This enabled the creation of a dynamic and interactive environment through custom HTML and JavaScript logic, as well as the measurement of reaction times and the logging of all relevant variables. The lab.js file used in this study is also available in the OSF repository (Schackmann, 2025a).

3.3.2 Supplementary task

In one of the two blocks, participants were asked to perform a *visual search task* (VST) as an ST. The task design and structure were taken from the standard task repository for the lab.js experiment builder (Henninger et al., 2021). The JavaScript source code was modified to fit the online study setup. In the VST, participants had to search a square box in the top left area of the environment for the presence of a divergent stimulus. Specifically, the box contained the letter L either 40 or 45 times in different orientations and positions, and in half of the cases an additional T, which acted as the divergent stimulus that had to be detected. The participants were instructed to press the N-key if there was no T or the Y-key if there was. There were a total of 15 boxes to scan in each trial and a total of 60 boxes for the whole ST block. To increase engagement in the VST, participants received feedback on whether their response was correct or not. In addition, there was a time limit of five seconds after which, if no response was recorded, a text was displayed asking participants to respond more quickly. An example square box for both VST conditions can be found in Figure 9.

The VST was selected to represent an engaging ST, similar to those TOs might engage in to counteract monotony and fatigue during low-demand phases (see Section 2.4). Based on



(a) Stimulus not present **(b)** Stimulus present

Figure 9. Both conditions of the VST.

task-switching theories, switching between task sets can lead to detriments in SA and performance (Shi & Bengler, 2022; Vandierendonck et al., 2010). The VST was therefore expected to interfere with the primary task, which was monitoring the remote environment following a TOR. While it also involves visual search, it differs from the remote monitoring task in that it requires focused attention on clearly defined and simple targets. In contrast, the monitoring task is more open-ended, requiring participants to shift from a known search (letters) to unknown, unpredictable targets. This transition in visual and attentional demands was expected to induce task-switching costs. Additionally, as the VST places minimal demands on working memory, it was not anticipated to interfere with the performance on the SAGAT or the takeover reason questionnaires and reduce their validity through carry-over effects.

3.3.3 Scenarios

A total of eight scenarios were presented to the participants. They were derived from previous work in the German Aerospace Centre's *Teleoperation ATO* project. However, relevance to actual train operations was only one of several criteria used to select the scenarios to be included in the study. First, the feasibility of presenting the scenarios in the online study was considered. For example, a broken overhead power line might be a reason for the onboard system to issue a TOR, but it would not have been feasible to depict this in a way and at a size that participants could see on the screen. In addition, scenarios that relied on overly complex concepts or required extensive prior knowledge were excluded to minimize bias against participants without train-driving experience. Ethical considerations were also taken into account during scenario selection. For example, a person standing too close to the tracks would be a realistic scenario, but was not implemented to avoid potential re-traumatisation of the train drivers within the sample.

Three categories of events could trigger a TOR. First, an object could appear in the camera view. In the study this was implemented as a line-side fire, palettes on the opposite track, a container on the opposite track and a cow standing close to the track. Second, a train parameter could reach a critical state, which was indicated either by a red icon or, in the case of the friction graph, by the needle entering the red zone. Two scenarios displayed such issues in the form of the friction and temperature parameter being in the critical range. Third, a safety-related or emergency message could be displayed in the notification area. This was implemented as a passenger initiated emergency brake and a notification for a necessary balise maintenance. In each ST condition block, two of the four reasons for the TOR could be found in the camera view, one in the parameter area, and one in the notification area. All of these scenarios can be found in Appendix A.5.

In each scenario, the critical reason remained visible for five seconds following the TOR. This duration was chosen to align with existing estimates of headway distances in train obstacle detection systems. Due to limited publicly available data, the value was approximated using specifications from the automated Hong Kong metro system (Zhangyu et al., 2021). With a maximum obstacle detection range of 120 meters and an estimated train speed of 100 km/h, which represents an average speed between typical German commuter and cargo trains, the system would require approximately 4.32 seconds to detect a pedestrian-sized object. This value was rounded up to five seconds. An additional 1.5 seconds were added to allow for the full playback of the TOR audio, as comprehension of the AGTOR was only possible if the complete message was heard. A further 1.5 second buffer was included to accommodate potential latency due to varying internet connection quality.

In total, participants thus had eight seconds to scan the environment after the TOR was initiated. This duration aligns with estimates from the conditional automated car driving domain regarding the time needed to regain SA comparable to manual driving (Deng et al., 2024). After these eight seconds, the environment was blanked out and participants were presented with the questionnaires about the reason for the takeover or with the SAGAT forms.

Each scenario featured a video sequence in the camera view area. These sequences were created using the *Unreal Engine* (Epic Games, 2023) based editor of the train simulation game *Train Sim World 4* (Dovetail Games, 2023). The visual environment shown in the videos was custom-built to depict varied railway settings. Key variables included the type of environment (urban vs. natural, flat terrain vs. valley, presence vs. absence of a river), the number of tracks, and whether the tracks were electrified. To maintain consistency and avoid confounding factors, no humans, level crossings, or related elements were included. Additionally, due to licensing constraints, no other trains were depicted. Still images of all scenarios with the reason for the takeover visible can be found in Appendix A.5.

The animations for the speedometer, train parameters and notifications were created using animations in *PowerPoint* (Microsoft Corporation, 2024).

3.4 Participants

The required sample size for the study was calculated using *G*Power* (Faul et al., 2007) through an a priori power analysis. The analysis aimed to achieve a statistical power of .80 to detect a medium effect size (f = .3) at the conventional alpha level of .05. This effect size represents a best estimate based on findings from the automated car driving literature and was considered appropriate for testing the main effect hypotheses of the study. It resulted in a target sample size of 128 participants, of which 64 should be active train drivers. To account for potential dropouts and unusable data, the recruitment goal was increased to 140 participants, with 70 of them being train drivers.

Participants were primarily recruited from the participant pools of the German Aerospace Centre and the Human Factors department at the Technical University of Berlin. Other recruitment methods included a newsletter, rail-related Reddit pages, and rail forums.

A total of 192 participants completed the study, which exceeded the target sample size. All met the minimum screen resolution requirement of 800 × 450 pixels, ensuring correct display of the environment. In addition, all participants successfully passed the audio and visual system checks, as well as the comprehension checks for understanding the environment and tasks (requiring at least 50% of test questions to be answered correctly). Furthermore, no participants were excluded based on their relative completion speed indicator (RSI; Leiner, 2019), as none exceeded the threshold value of two, indicating that across all experiment pages, no participant responded at a speed twice as fast as the median time. Eight participants were excluded due to self-reported diligence of working on the study below 50%. A further 16 participants were excluded, because they had a median of more than 5 seconds for all attention checks combined, or did not have a single trial where the attention check time was less than 5 seconds, meaning that they could potentially have missed any takeover reason due to inattention. Finally, eight participants were removed due to missing data caused by logging problems, resulting in a final sample of n = 160 participants.

This final sample was predominantly male (65.0%) and based in Germany (98.6%). Participants had a mean age of 33.1 years (SD = 12.7), with ages ranging from 18 to 75. 48.8% of the participants reported having only passenger-level railway knowledge. 26 participants (16.3%) were active train drivers, and 16 (10.0%) were employed in other railway related jobs. Two participants in the latter group held a valid train driving license and reported driving trains at least 100 hours per year. Therefore, they were also included in the train driver group, resulting in a total of 28 train drivers (17.5%). This number was lower than the desired 64 train drivers. A detailed overview of the sample is provided in Table 3.

Table 3. Description of the sample.

Demographics	**Attributes**	***M***	***SD***
Age		33.1	12.7
		n	**%**
Gender	Female	54	33.8
	Male	104	65.0
	Other	2	1.3
Country	Germany	158	98.8
	Switzerland	1	0.6
	Denmark	1	0.6
Rail experience	None	10	6.3
	Only as passenger	78	48.8
	Advanced, but not professional	30	18.8
	Professional, not (mainly) train driving	16	10.0
	Professional, train driver	26	16.3
Railway Staff	Train Driver	28	17.5
	Traffic Controller	7	4.4
	Operational Staff	6	3.8

3.5 Data analysis

Data was analysed using R (v4.2.2; R Core Team, 2022) and Python (v3.12). Preprocessing included calculating SAGAT, workload, usability and telepresence scores as well as coding recognition performance and the importance ratings of the single SAGAT elements. The experimental and questionnaire data were merged, and participants were excluded based on the exclusion criteria stated in the previous section. Due to the randomized assignment, slightly more participants were allocated to the AGTOR group (n = 81) than to the SWT group (n = 79). Additionally, data from three participants in the ST block was not recorded, resulting in a final sample of n = 157 for all analyses in the ST condition, compared to n = 160, in the no-task condition. A detailed overview over participants' n for key variables in the study can be found in Appendix A.7.

Subsequently, descriptive and then confirmatory analyses were conducted. Additionally, a number of exploratory analyses were conducted. These included the variables age, subjective SA, experience, ST performance, telepresence, usability, subjective difficulty of the ST, as well as fun and diligence working on the study. Group differences were analyzed using independent and paired samples t-tests. When the assumption of normality was not met, non-parametric alternatives were applied: Wilcoxon rank-sum tests for comparisons between the TOR and experience conditions, and Wilcoxon signed-rank tests for comparisons between ST conditions (Wilcoxon, 1945). However, as these analyses were exploratory in nature, no adjustment for multiple comparisons was performed. Consequently, the results should be interpreted with caution and regarded as preliminary until they can be confirmed in future studies.

For all inferential tests, the standard α error level of .05 was used. To conduct the confirmatory analyses, three models were specified. For models 1 and 3, mixed linear effects models with random intercepts for each participant were fitted. For model 2, which included eight binary outcome variables (whether the takeover reasons were correctly identified or not), a generalized linear mixed-effects model was used. The models were specified as follows for each observation i of participant j:

Model 1: SAGAT Scores (H1a, H5a, H6)

This model uses SAGAT scores for the ST and no-task condition as dependent variables.

$$SAGAT_{ij} = \alpha + \beta_1 \cdot \mathrm{ST}_{ij} + \beta_2 \cdot \mathrm{TOR}_{ij} + \beta_3 \cdot Experience_{ij} + u_j + \varepsilon_{ij}$$

Statistical implications:

- **H1a**: If $\beta_1 < 0$, performing an ST will reduce SAGAT scores compared to the no-task condition.
- **H5a**: If $\beta_3 > 0$, train drivers will show higher SAGAT scores compared to non-train drivers.

Model 2: Recognition performance (H1b, H3, H4, H5b, H6)

This model uses takeover reason recognition for the eight scenarios as dependent variables.

$$\ln\left(\frac{P(Recognition_{ij} = 1)}{1 - P(Recognition_{ij} = 1)}\right) = \alpha + \beta_1 \cdot \mathrm{ST}_{ij} + \beta_2 \cdot \mathrm{TOR}_{ij} + \beta_3 \cdot Experience_{ij} + \beta_4 \cdot (\mathrm{ST}_{ij} \times \mathrm{TOR}_{ij}) + \beta_5 \cdot (\mathrm{TOR}_{ij} \times Experience_{ij}) + u_j$$

Statistical implications:

- **H1b**: If $\beta_1 < 0$, performing an ST will reduce recognition compared to the no-task condition.
- **H3**: If $\beta_2 > 0$, participants in the AGTOR condition will show better recognition compared to those in the SWT condition.
- **H4**: If $\beta_4 > 0$, the improvement in recognition due to AGTORs is more pronounced when an ST was previously performed.
- **H5b**: If $\beta_3 > 0$, train drivers will have better critical recognition than non-train drivers.
- **H6**: If $\beta_5 < 0$, the improvement in performance facilitated by the AGTOR condition is stronger for non-train drivers compared to train drivers.

Model 3: Workload (H2)

This model uses the mean workload scores for the ST and no-task condition as dependent variables.

$$Workload_{ij} = \alpha + \beta_1 \cdot ST_{ij} + \beta_2 \cdot Experience_{ij} + \beta_3 \cdot (ST_{ij} \times Experience_{ij}) + u_j + \varepsilon_{ij}$$

Statistical implications:

- **H2**: If $\beta_1 > 0$, performing an ST will lead to an increase in reported workload.

3.6 Preregistration

The study design, including all hypotheses, was pre-registered in the OSF prior to data collection (Schackmann, 2025b). One deviation from the original protocol was made in the data analysis of takeover reason recognition (model 2). The pre-registration stated that *"All of these models are constructed as mixed linear models with a set of fixed factors and a random intercept for each participant."*. However, this followed the idea that for each participant, a sum score would be calculated for both experimental blocks, and thus these two sum scores, which would have ranged from zero to four, would be compared. However, in order to obtain a more adequate and precise analysis, it was considered more appropriate to look at recognition for each one of the reasons, resulting in a total of eight binary measures, grouped by ST condition into groups of four measures. This change required the statistical analysis of model 2 to be performed using a logistic approach rather than a linear mixed effects approach.

The preprocessed data set, containing data of 160 participants has been uploaded to the OSF and is publicly available (Schackmann, 2025a). For data protection reasons, it does not contain the variables age, gender, country of residence, professional railway experience and the free text answers.

4 Results

Three main variables were examined in this study: SAGAT scores (model 1), takeover reason recognition (model 2) and workload (model 3). For all models, including participant-specific slopes in addition to random intercepts did not improve model fit, as indicated by non-significant likelihood ratio tests (Model 1: $\chi^2(1) = 0.09$, p = .762; Model 2: $\chi^2(1) = 0.21$, p = .899; Model 3: $\chi^2(1) = 0.14$, p = .701).

In addition to these three measures, several additional variables were included to assess participants' responses to the novel remote environment and to explore other potential factors influencing SA and workload. A comprehensive overview of the correlations among all measured variables can be found in Appendix A.8.

The results are presented in the following order: first, analyses of the control variables; second, findings related to SA and workload, including the confirmatory analyses; and finally, results concerning the evaluation of the remote environment.

4.1 Control variables

Age, attention checks and ST performance

Age was moderately to strongly associated with performance measures. Increasing age was associated with less takeover reasons recognized (r = -.29, p < .001), worse ST performance (r = -.39, p < .001) and higher median reaction times to the attention checks in the no-task condition (r = .40, p < .001). These associations are depicted in Figure 10 using z-standardized values to facilitate visual comparison. Age was furthermore negatively correlated with all three fatigue scores (r between -.30 and -.25; all $p \leq .001$).

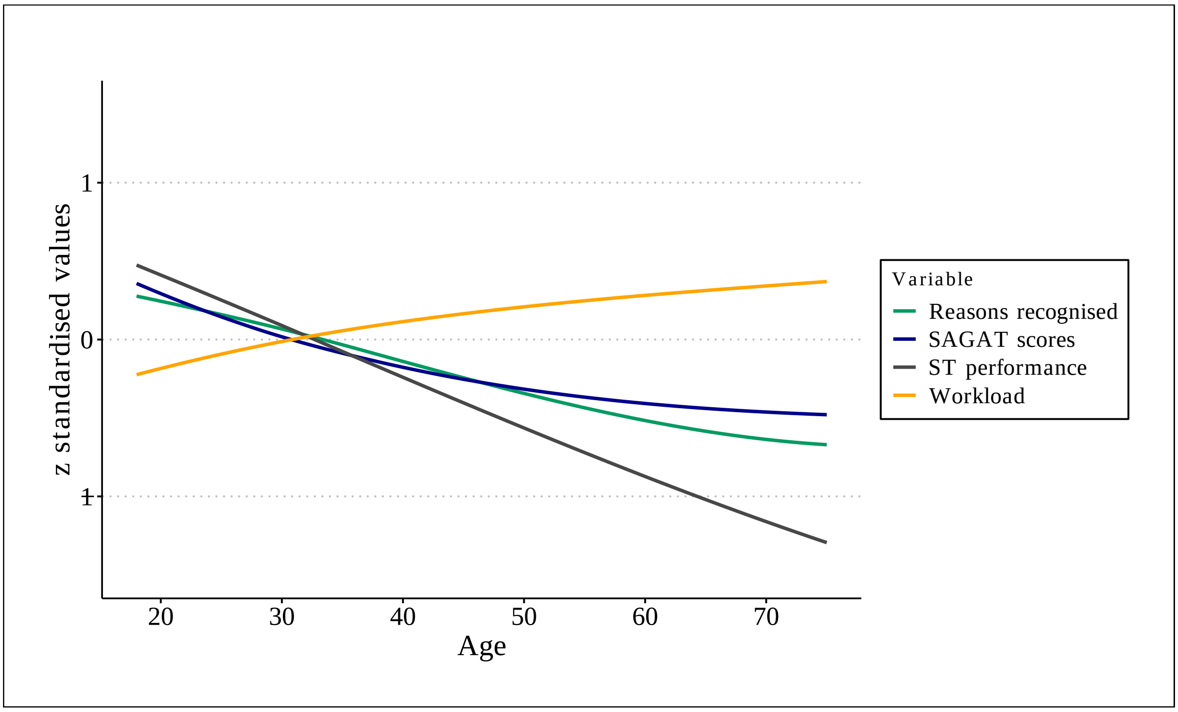


Figure 10. Z-standardized values of four performance measures by age. Lines show smoothed raw values.

To examine the level of inattention during the no-task condition, the median reaction time across four attention checks was recorded. Participants with a median reaction time exceeding five seconds were excluded from the study. The remaining participants needed on average 2.89 s (SD = 1.35 s) to respond to the attention checks. In addition to the positive correlation with age, higher median response times were associated with fewer takeover reasons recognized (r = -.32, p < .001).

Furthermore, the percentage of correct answers in the VST was logged to capture individual differences in participants' performance on a visually demanding task and their level of engagement. Overall, participants rated the difficulty of the VST as 5.3 (SD = 2.3) on a scale from 1 to 10 and achieved a mean accuracy rate of 82.04% (SD = 9.46) across the 60 VST trials. With regards to accuracy, only slight differences were observed between participants in the AGTOR (M = 82.96, SD = 9.33) and SWT group (M = 81.08, SD = 9.57; W = 2735.0, p = .226) , as well as between train drivers (M = 81.98, SD = 9.35) and non-train drivers (M = 82.05, SD = 9.52; W = 1790.0, p = .872).

Fatigue, diligence and fun working on the study

To measure time-on-task effects, participants' self-reported fatigue was measured using the KSS (Åkerstedt and Gillberg, 1990; scale from 1 to 10) before the first experimental block, between both blocks and after the second block. When comparing fatigue scores before and after the ST and no-task condition, median fatigue increased by Mdn = 0.5 in each case. Regardless of the condition, post-experiment fatigue scores (Mdn = 4) were, on a descriptive level, slightly higher than both mid-experiment (Mdn = 3) and pre-experiment scores (Mdn = 3), suggesting a slight increase in fatigue over the course of the study.

Finally, participants were asked how diligently they had worked on the study, and how enjoyable they had found it. Participants reported an average self-rated diligence of 84.21% (SD = 11.95%) while working on the study. However, this value is based on responses after excluding participants who indicated working on the study with less than 50% diligence, as part of the data preprocessing procedure. Those in the AGTOR group reported fractionally higher diligence than those in the SWT group (M_{AGTOR} = 85.30% vs. M_{SWT} = 83.10%; W = 2803.5, p = .177), as did train drivers compared to non-train drivers ($M_{TrainDriver}$ = 85.31% vs. $M_{Non-TrainDriver}$ = 83.98%; W = 1755.0, p = .678).

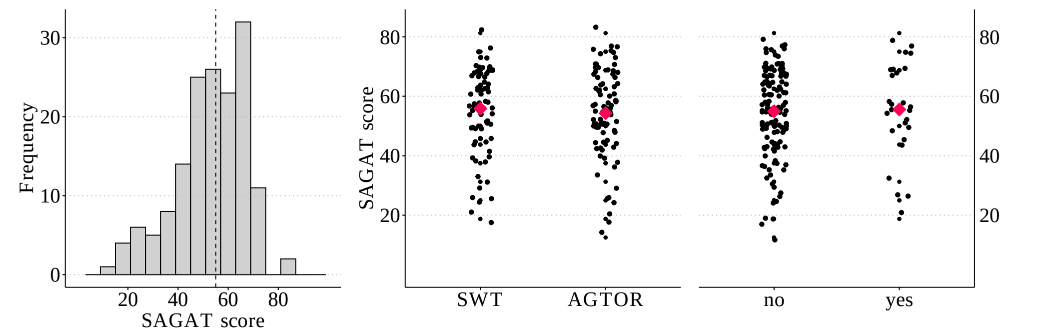
Regarding fun working on the study, participants stated a mean of 59.83% (SD = 26.89%), but ratings differed greatly (Min = 0.5%, Max = 100%). Those in the AGTOR group rated the study fractionally more fun than those in the SWT group (M_{AGTOR} = 60.12% vs. M_{SWT} = 59.53%; W = 3103.5, p = .745), and train drivers more than non-train drivers ($M_{TrainDriver}$ = 61.41% vs. $M_{Non-TrainDriver}$ = 59.49%; W = 1824.5, p = .918).

4.2 Situational awareness

SAGAT scores

SA was measured threefold in this study. First, as a measure of level one SA, participants completed a SAGAT questionnaire once per block, in which they were asked to fill in 16 items of information on a blanked version of the environment (see Section 3.2) and additionally could rate the subjective importance of these items.

Across both conditions, a mean of 52.36% (*SD* = 11.90%) of the 16 SAGAT items were answered correctly by the participants. All SAGAT score distributions are shown in Figure 11, and the percentages of correct responses along with the importance ratings for all 16 SAGAT items are provided in Appendix A.9.



(a) Overall distribution **(b)** TOR **(c)** Train driver

ST condition

(d) Overall distribution **(e)** TOR **(f)** Train driver

No-task condition

Figure 11. SAGAT distributions for the ST and no-task condition. Dashed lines and red diamonds indicate mean values.

Grouped together, camera view items yielded the highest correct response rate (*M* = 73.96%, *SD* = 34.69%), followed by train parameter items (*M* = 65.36%, *SD* = 24.34%), notification items (*M* = 42.19%, *SD* = 34.10%) and train information items (*M* = 18.33%, *SD* = 24.13%).

With regards to subjective importance, participants rated anomalies on the track as the most important information (M = 72.81%, SD = 37.55%) and the fifth message entry as the least important (M = 2.19%, SD = 11.69%). There was only a small correlation (r = .10, p = .18) between recognition rates and importance ratings.

Confirmatory analyses of SAGAT scores were conducted using a mixed-effects linear regression (model 1; detailed results in Appendix A.10), which yielded a marginal R^2 = .28 and a conditional R^2 = .51. This indicates that a substantial portion of the explained variance is attributable to the participant-specific random intercepts. All main effects and distributions for model 1 can be found in Figure 12.

There was a significant main effect of the ST condition on SAGAT scores (b = 5.46, p < .001, r = 0.31), with participants achieving higher SAGAT scores in the ST (M = 55.06, SD = 14.61) compared to the no-task condition (M = 49.65, SD = 14.49). The direction of this effect was opposite to hypothesis 1a.

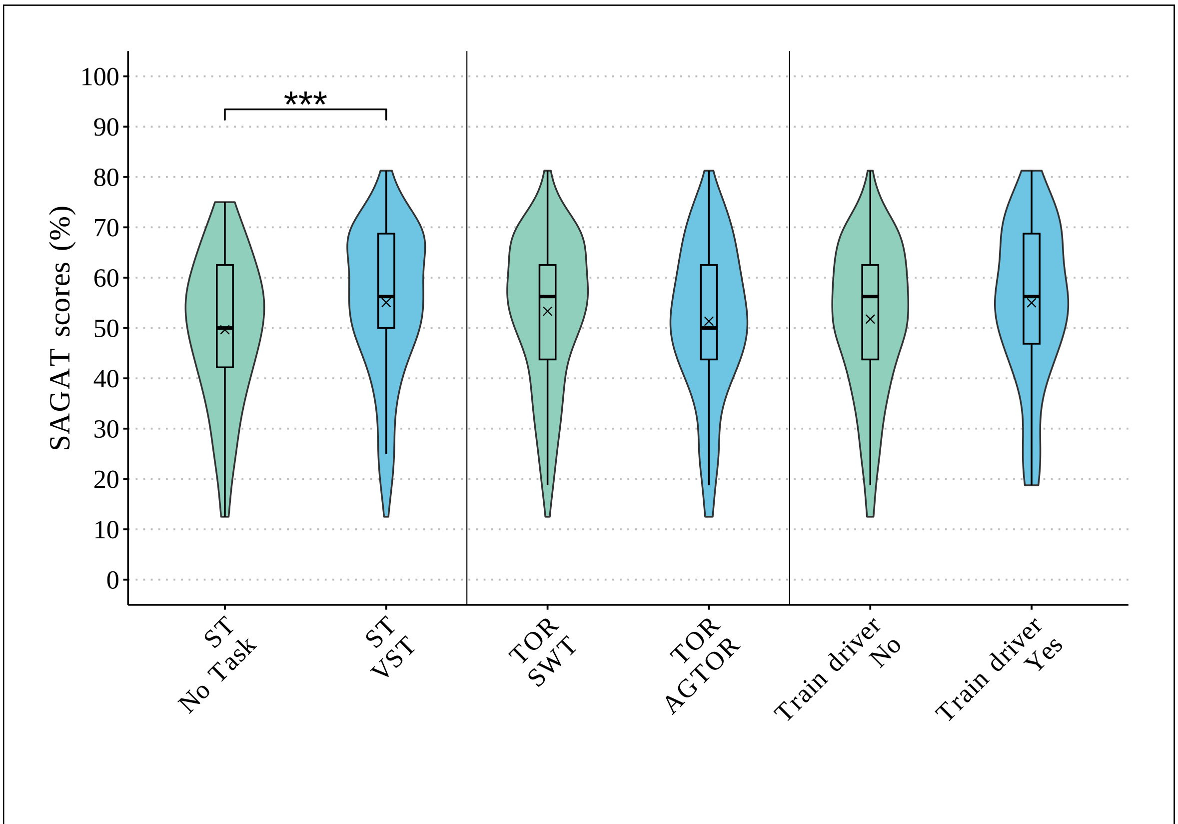


Figure 12. Main effects on SAGAT scores and distributions. *** signifies $p < .001$.

No significant main effect was observed for the TOR condition (b = -2.22, p = .240, r = -.09), with people in the AGTOR condition obtaining slightly lower SAGAT scores (M_{ST} = 54.30, SD_{ST} = 14.84%; $M_{no-task}$ = 48.46%, $SD_{no-task}$ = 14.74%) than those in the SWT condition (M_{ST} = 55.84%, SD_{ST} = 14.42%; $M_{no-task}$ = 50.87%, $SD_{no-task}$ = 14.21%).

Contrary to hypothesis 5a, there was no significant main effect of experience (b = 3.51, p = .160, r = .11), but train drivers achieved higher SAGAT scores than non-train drivers, especially in the no-task ($M_{TrainDriver}$ = 54.46%, $SD_{TrainDriver}$ = 14.42%; $M_{Non-TrainDriver}$ = 48.63%, $SD_{Non-TrainDriver}$ = 14.35%) compared to the ST condition ($M_{TrainDriver}$ = 55.56%, $SD_{TrainDriver}$ = 16.48%; $M_{Non-TrainDriver}$ = 54.95%, $SD_{Non-TrainDriver}$ = 14.26%). However, this interaction (see Figure 13) was not tested in model 1.

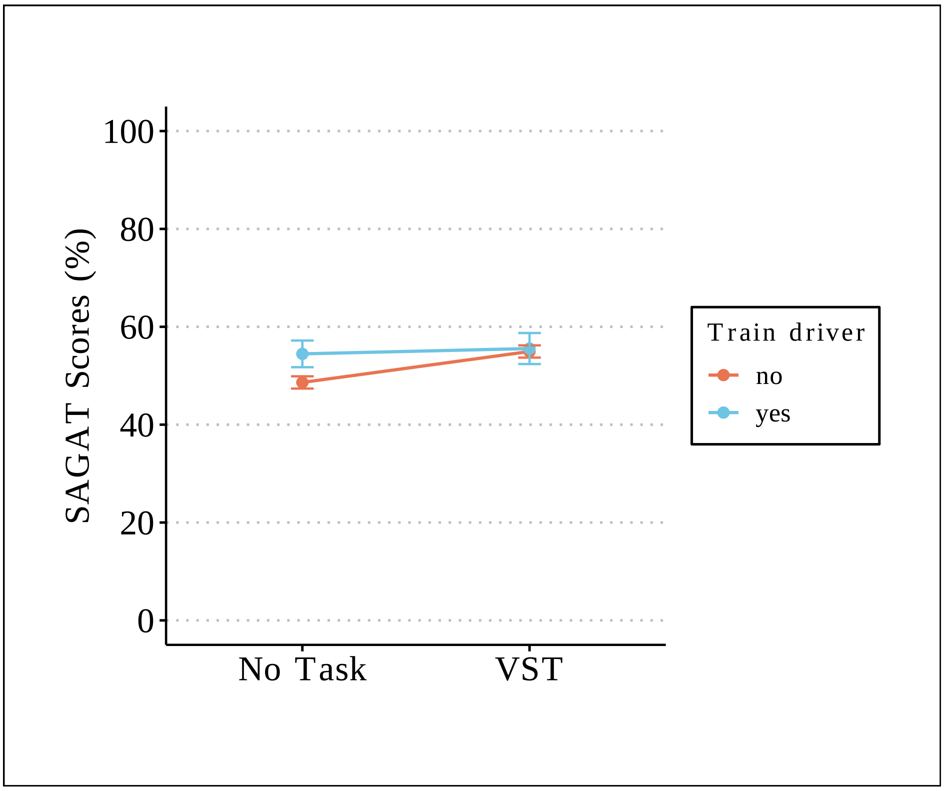


Figure 13. Interaction of ST condition and experience on SAGAT scores. Error bars signify $\pm$ *SE*.

Recognition performance

In addition to the SAGAT forms and to approximate a measure of level two SA, participants were asked to select what they believed was the reason for each takeover from a predefined list following every takeover sequence. This resulted in a total of eight possible reasons identified: four in the ST block and four in the no-task block.

Across all conditions, participants averaged 5.32 (SD = 1.94) of the eight takeover reasons correctly identified. An overview of the percentage of correct identifications for the eight reasons can be found in Figure 14. Participants most frequently recognized the passenger initiated emergency brake notification (M = 81.25%, SD = 39.15%) and least frequently recognized the friction parameter being in the critical range (*M* = 50.32%, *SD* = 50.16%). Overall, recognition was higher for the notification-related takeover reasons (*M* = 75.63%, *SD* = 33.16%), compared to the parameter-related (*M* = 65.00%, *SD* = 33.55%) and camera-related takeover reasons (*M* = 61.88%, *SD* = 34.56%).

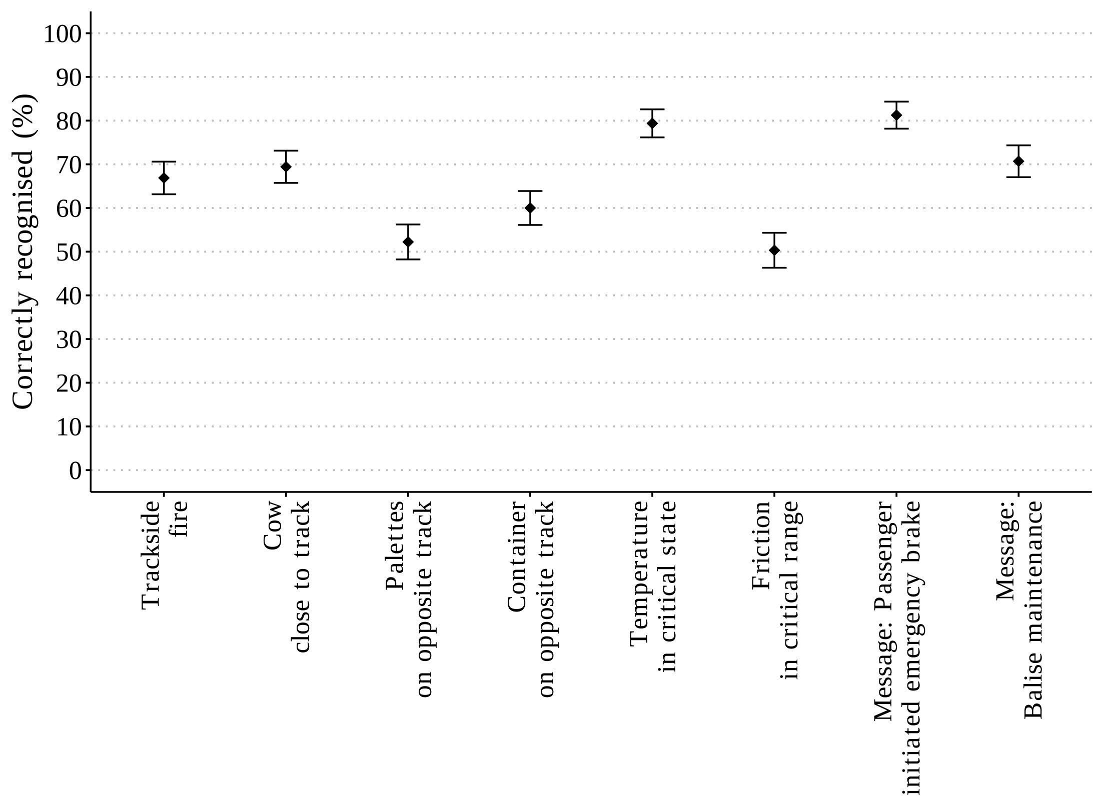


Figure 14. Percentages of correct detection for the eight takeover reasons. Error bars signify ± *SE*.

Confirmatory analyses of takeover reason recognition were conducted using a generalized linear mixed-effects model (model 2; detailed results in Appendix A.10) based on the eight binary reason detection variables. It yielded a marginal R^2 = .05 and a conditional R^2 = .26, suggesting that a large proportion of the explained variance is attributable to the participant-specific random intercepts. Odds ratios (ORs) for all tested variables are presented in Figure 15.

There was a main effect of the ST condition on takeover reason recognition (*OR* = 0.66, *p* = .020, 95% CI [0.46, 0.94]), meaning participants were 34% less likely to recognize the takeover reasons in the ST condition (*M* = 2.43 of the four reasons, *SD* = 1.17), compared to in the no-task condition (*M* = 2.88, *SD* = 1.09). This result supports hypothesis 1b.

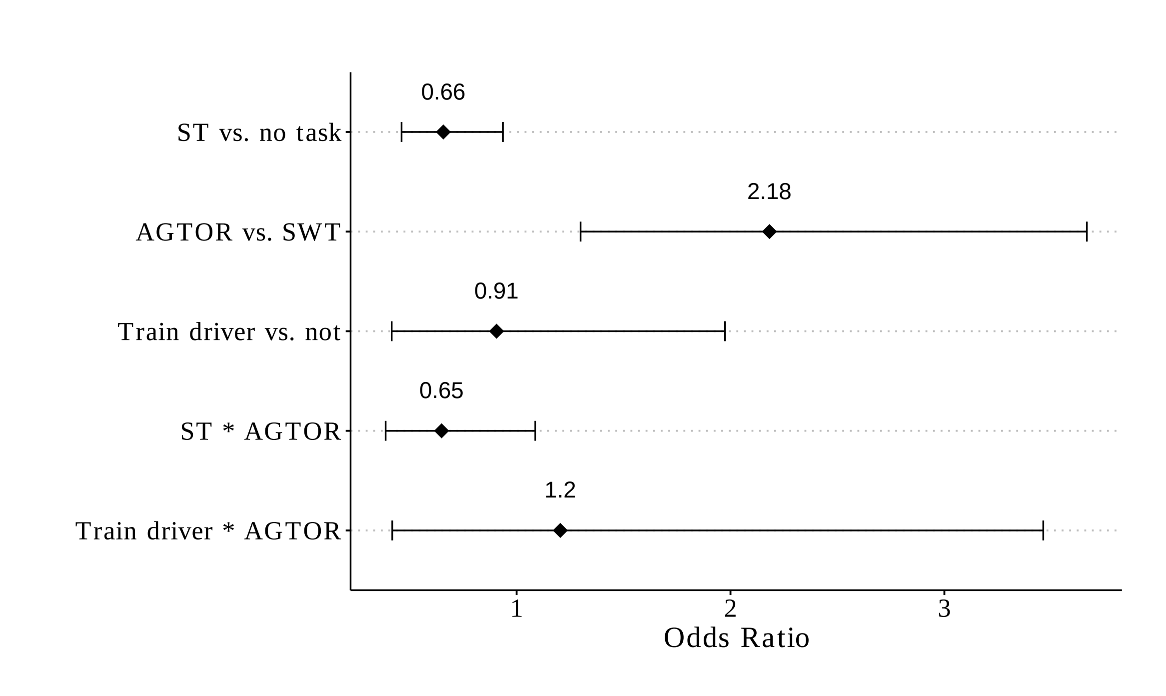


Figure 15. Odds ratios regarding takeover reason recognition. Error bars signify ± 95% *CI* limits.

Further, there was a significant main effect of the TOR condition on takeover reason recognition (*OR* = 2.18, *p* = .003, 95% CI [1.30, 3.67]). Irrespective of the ST condition, participants in the AGTOR group were 118% more likely to recognize takeover reasons (*M* = 5.70, *SD* = 1.84) compared to those in the SWT condition (*M* = 4.94, *SD* = 1.97). This result supports hypothesis 3.

There was no main effect for experience (*OR* = 0.91, *p* = .804, 95% CI [0.42, 1.98]), therefore hypothesis 5b is rejected. On a descriptive level, train drivers recognized slightly more of the takeover reasons (*M* = 5.44, *SD* = 2.14) than non-train drivers (*M* = 5.30, *SD* = 1.90).

There were no interaction effects between the ST condition and the TOR condition (*OR* = 0.65, *p* = .101, 95% CI [0.39, 1.09]), nor between experience and the TOR condition (*OR* = 1.20, *p* = .730, 95% CI [0.42, 3.46]). Therefore, hypotheses 4 and 6 are rejected.

Subjective SA

In addition to the two objective SA measures, subjective SA was assessed by asking participants to rate their confidence in having understood the scenario and the state of the train. This rating was collected three times per block following each takeover scenario, excluding those followed by the SAGAT questionnaire.

Participants reported higher subjective SA in the no-task (*Mdn* = 8.33) compared to the ST condition (*Mdn* = 6.67; $p < .001$). In both ST conditions, subjective SA was positively correlated with takeover reasons recognized (r_{ST} = .54, $p < .001$; $r_{no-task}$ = .58, $p < .001$). Participants in the AGTOR group reported marginally higher subjective SA (*M* = 7.45, *SD* = 1.79) than those in the SWT group (*M* = 7.22, *SD* = 1.76; *W* = 2898.5, *p* = .305). Only a slightly higher subjective SA was also found for train drivers (*M* = 7.72, *SD* = 1.43), compared to non-train drivers (*M* = 7.26, *SD* = 1.83; *W* = 1638.5, *p* = .348).

4.3 Workload

After each of the two blocks, workload was assessed using the DLR-WAT questionnaire (Grippenkoven et al., 2018). For the confirmatory analyses, a mean score across five subscales, each ranging from 1 to 200, was used. Across both ST conditions, participants had a mean workload score of 126.92 (*SD* = 29.03), which was above the optimum workload level of 100. Highest overall scores were obtained for the subscales *workload due to time pressure* (*M* = 136.81, *SD* = 37.43) and *workload due to information intake* (*M* = 134.93, *SD* = 33.58). The lowest scores were obtained for *workload due to decision-making* (*M* = 120.58, *SD* = 35.59) and *workload due to knowledge retrieval* (*M* = 118.21, *SD* = 41.87). Participants in the SWT group had slightly lower overall workload scores (*M* = 125.78, *SD* = 30.06) than those in the AGTOR group (*M* = 128.04, *SD* = 28.14; $t(156.7) = -0.48$, *p* = .625).

Confirmatory analyses of workload scores were conducted using a mixed-effects linear regression (model 3; detailed results in Appendix A.10), which yielded a marginal R^2 = .38 and a conditional R^2 = .54. This indicates that a substantial portion of the explained variance is attributable to participant-specific random intercepts. All main effects and distributions for model 3 can be found in Figure 16.

As expected in hypothesis 2, working on an ST had a significant effect on workload (*b* = 15.32, $p < .001$, *r* = .43), leading to higher workload (*M* = 134.49, *SD* = 32.07) than in the no-task condition (*M* = 119.35, *SD* = 32.84). This difference was most noticeable in the *workload due to information intake* ($M_{\Delta ST-NoTask}$ = 6.45) and least in the *workload due to decision-making* ($M_{\Delta ST-NoTask}$ = -0.35).

Train drivers had slightly lower workload scores (*M* = 123.92, *SD* = 28.55) than non train drivers (*M* = 127.56, *SD* = 29.20), but this effect was not significant (*b* = -3.14, *p* = .643, *r* = -.04). There was further no significant interaction between experience and the ST condition (*b* = -1.00, *p* = .869).

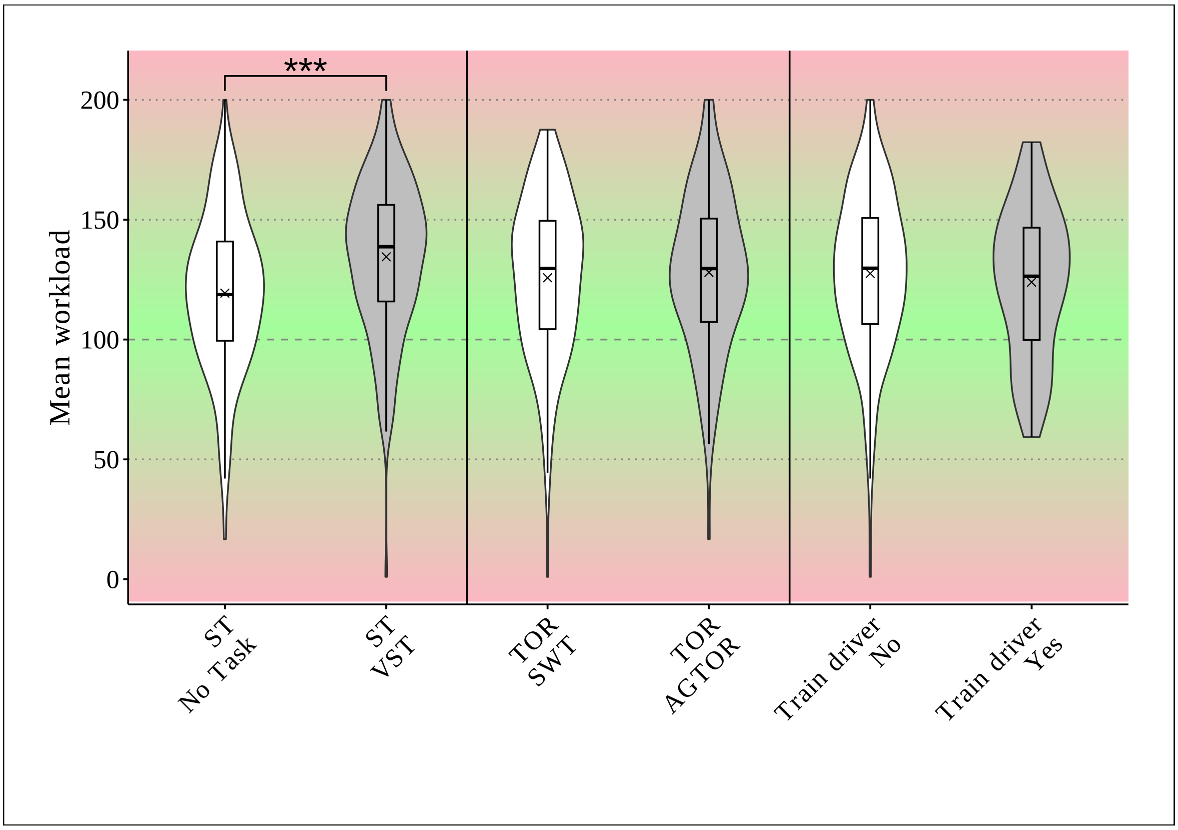


Figure 16. Main effects on workload scores and distributions. *** signifies $p < .001$. Green colour indicates the optimum workload level and gradients to red indicate deviations towards over- and underload.

Frustration and subjective performance

Due to differences in scales, the subscales *frustration with the task* and *subjective task accomplishment* were excluded from the mean workload score used in the confirmatory analyses (for further details, see Section 3.2). Frustration was higher in the ST (Mdn = 150.50) compared to the no-task condition (Mdn = 142.50; p = .018), with both scores around midpoint between optimal (100) and very frustrating (200). Regarding subjective task performance, participants self-rated their performance in the ST condition (Mdn = 43.00) as worse than in the no-task condition (Mdn = 60.5; $p < .001$), on a scale from very poor (1) to optimal (100).

Subjective performance was negatively associated with frustration (r = -.57), but positively correlated with SAGAT scores (r = .32), takeover reasons recognized (r = .46) and subjective SA (r = .66). Frustration on the other hand was negatively associated with takeover reasons recognized (r = -.42), subjective SA (r = -.53), usability (r = -.44) and fun working on the study (r = -.32). Additionally, it was positively correlated with overall workload (r = .41).

All correlations involving subjective performance and frustration reported in this paragraph were significant at $p < .001$.

4.4 Evaluation of the remote environment

Because of the novelty of the online remote environment, usability was assessed at the end of the study using the SUS (Brooke, 1996). Overall ratings for the remote environment were favourable (M = 76.92, SD = 23.50; aggregated score from 0 to 100). Those in the SWT group gave the environment a slightly higher rating (M = 78.73, SD = 23.49) than those in the AGTOR group (M = 75.15, SD = 23.52; W = 3501.5, p = .303). For a comparison of the subscale ratings for both TOR groups, see Figure 17. Train drivers (M = 77.05, SD = 21.80) did barely differ in their usability ratings from non-train drivers (M = 76.89, SD = 23.93; W = 1820.0, p = .902).

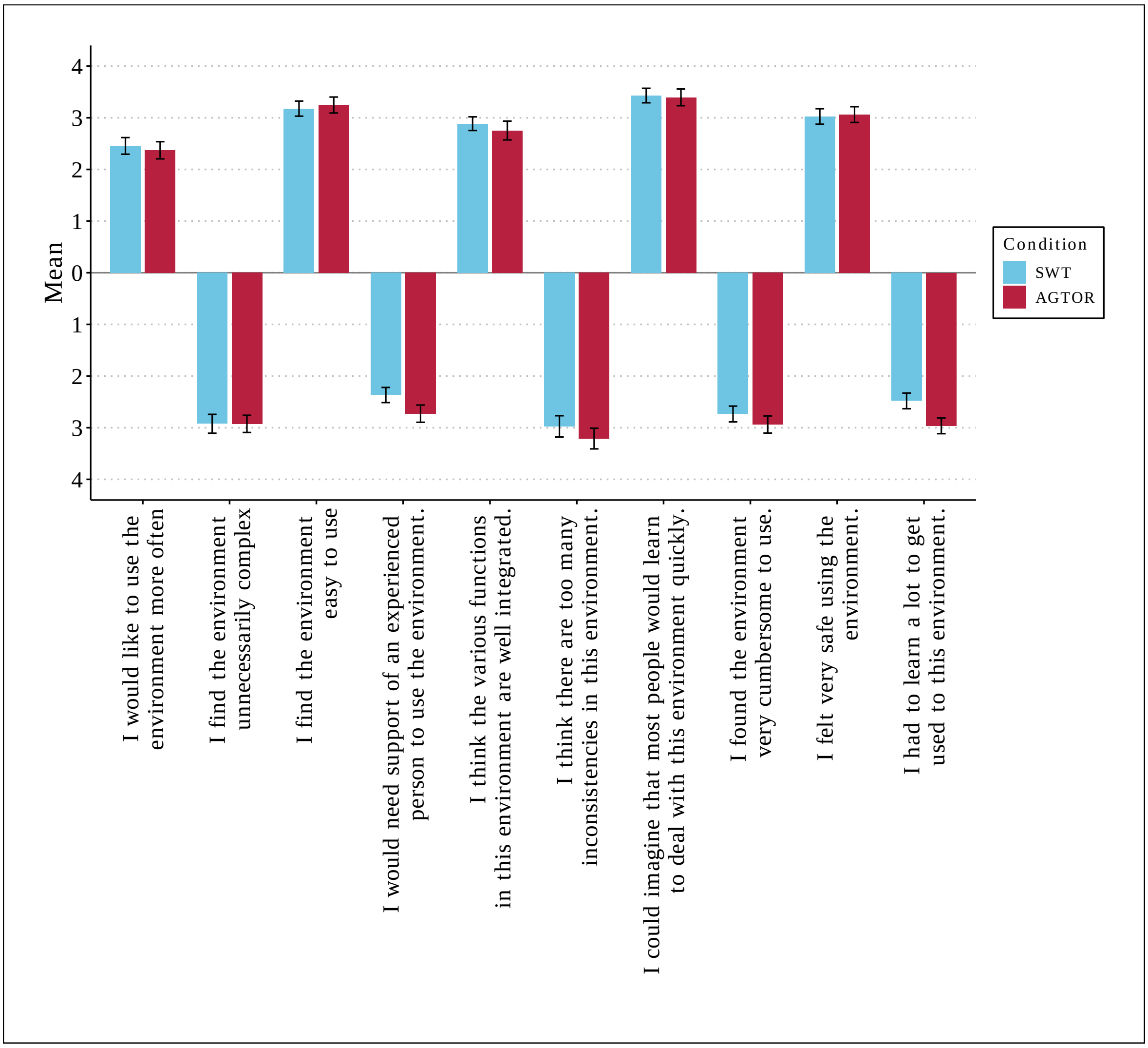


Figure 17. Usability ratings for single items between both TOR groups. Negatively worded items face downwards. Error bars signify ± SE.

Furthermore, to measure the extent to which people felt immersed in the remote environment, a telepresence questionnaire (Kim & Biocca, 1997) was included in the study. Participants rated the level of immersion in the environment midway between not at all and fully immersed (M = 2.28, SD = 0.86; mean score from 0 to 5). Ratings did barely differ between the AGTOR (M = 2.37, SD = 0.85) and SWT group (M = 2.19, SD = 0.86; $t(157.9)$ = -1.35, p = .180) and train drivers (M = 2.31, SD = 0.92) and non-train drivers (M = 2.28, SD = 0.85; $t(37.3)$ = -0.19, p = .854).

5 Discussion

This thesis aimed to measure how engagement in STs affects SA in takeover situations and workload in an RTO setting. Furthermore, it investigated whether AGTORs can help regain SA faster. To measure the extent to which work experience facilitates SA build-up, a train driver subsample was included.

Effects of the supplementary task

Engaging in an ST led to significantly fewer takeover reasons recognized, as was assumed in hypothesis 1b. Surprisingly, however, participants achieved significantly higher SAGAT scores in the ST condition compared to the no-task condition. This finding contradicts the expected direction of the effect proposed in hypothesis 1a.

These findings are particularly interesting as they reflect both potential downsides and benefits of introducing an ST. On one hand, they align with the assumption that an ST increases task-switching costs thus reducing SA (Shi & Bengler, 2022; Weaver & DeLucia, 2022; Xu et al., 2024), as indicated by lower recognition performance. On the other hand, they support previous research suggesting that an ST can enhance engagement, reduce fatigue, and mitigate out-of-the-loop effects (Lu et al., 2021; Miller et al., 2015; Mishler & Chen, 2023), which may explain the improved SAGAT scores observed in the study.

A possible explanation for these seemingly contradictory outcomes may lie in the distinct levels of SA targeted by each measure. SAGAT scores primarily reflect level one SA, namely, the perception of the elements in the environment. In contrast, recognition performance may better represent level two SA, which involves understanding the significance of those elements. Therefore, the ST may have hindered the development of level two SA by impairing deeper cognitive integration, while simultaneously supporting level one SA by maintaining perceptual engagement with the environment. This interpretation fits with theories of task switching, which suggest that switching between tasks with different cognitive demands can disrupt higher-level cognitive processing (Vandierendonck et al., 2010).

A possible approach to reducing task-switching costs could involve either incorporating a brief pause between the ST and the monitoring task or integrating a preview of the impending takeover within the ST itself (Vandierendonck et al., 2010; Xu et al., 2024). This could be achieved by using multistage TORs that gradually hand over control to the human rather than doing so immediately (Liu et al., 2023). Such measures may provide the TO with additional time for cognitive reconfiguration, thereby mitigating residual cognitive interference from the preceding task (Vandierendonck et al., 2010). However, given the time-critical nature of many takeover scenarios, implementing these strategies in practice may prove challenging. Nonetheless, future research could place greater emphasis on applying principles from task-switching theory to the RTO context.

Another explanation for these findings is that the visibility of the event that triggered the TOR was limited to a short five-second timeframe, whereas participants had more time (three additional seconds) to scan the entire environment after the takeover reason had disappeared. Thus, the ST might, as expected, have caused initial task-switching costs, impairing performance during the crucial first seconds post-TOR and reducing the ability to identify the cause of the takeover. However, once the initial switching cost was overcome, the ST may have promoted more holistic scene perception by increased engagement in contrast to the no-task condition, where performance remained more stable but possibly at a consistently lower level throughout the available timeframe.

Following this reasoning, participants who initially missed the takeover reason may have continued to search the environment more thoroughly, aware that identifying it was crucial. Consequently, they remained highly cognitively engaged throughout the timeframe and scanned the environment more effectively. Since more takeover reasons were missed in the ST condition, this effect might have led to the increased SAGAT scores compared to the no-task condition. Interestingly, participants subjectively rated their SA as higher in the no-task condition. This may suggest that recognizing the takeover reason was (correctly) perceived as the most critical aspect of the task and, consequently, was more strongly associated with subjective SA ratings than the SAGAT scores, which is supported by the observed correlations.

Both the assumption of increased task-switching costs and heightened engagement align with the finding that, as expected in hypothesis 2, working on an ST was associated with significantly higher workload values. However, in both conditions, workload levels exceeded the optimal range and leaned into the overload area of the DLR-WAT (Grippenkoven et al., 2018). Participants reported the highest workload for the dimensions of time pressure and information intake, while knowledge retrieval was rated as least demanding.

These results suggest that the experimental setup was successful in conveying the time-critical nature of identifying the takeover reason and building up SA. Nonetheless, it had been expected that workload levels in the no-task condition would fall closer to the under-load range, because of the monotonous nature of the monitoring task (Endsley, 2021). One possible explanation is that the novelty of the task and environment may have been somewhat overwhelming for participants. This interpretation is supported by usability ratings, which, while generally positive, indicated that participants found the environment to be somewhat complex and cumbersome to use.

Furthermore, the relatively short duration of the experiment (each block lasting approximately ten minutes) and the limited total number of eight trials may not have allowed participants to fully familiarize themselves with the environment and task, potentially inflating workload ratings. This is further reflected in the fatigue ratings, which remained relatively stable throughout the study. Future research should explore longer-term engagement with the system and the TO role to assess whether workload levels stabilize over time, potentially shifting toward more optimal or even underload conditions.

Nevertheless, even without extensive training or habituation to the environment and, in the case of non-train drivers, without prior domain-specific experience, participants in the study were able to recognize over 66% of takeover reasons overall, and 71% when assisted by AGTORs. While these recognition rates would likely fall short of the requirements for a real-world RTO setting, they suggest that the environment design already offers a solid foundation. With the integration of additional support measures, such as the ability to replay camera footage, extended detection timeframes, and structured training, recognition performance could effectively reach operationally acceptable levels.

Effects of AGTORs

Hypothesis 3 proposed that AGTORs would help participants direct their attention toward the most critical element of the scene, i.e. the takeover reason, first. This assumption was supported by the significantly higher number of takeover reasons recognized in the AGTOR condition compared to the SWT condition.

This finding supports the idea that AGTORs convey more information, namely guidance toward the location of the relevant event, compared to SWTs, which simply indicate the need to take over without offering contextual cues (Liu et al., 2023). However, while AGTORs improved recognition performance, they did not significantly affect SAGAT scores. In fact, participants in the AGTOR group exhibited slightly lower SAGAT scores than those in the SWT group, although this difference was not statistically significant. If understanding the takeover reason is indicative of level two SA, then it is possible that the AGTOR helped establish level two SA first, at the expense of slightly impaired level one SA. This fits the conceptualisation of SA being able to develop in a non-linear, top-down way (Endsley, 2015). It could also suggest a mild attentional tunnelling effect, where maintaining attention on a specific area of the environment may come at the cost of reduced overall SA (Endsley, 2021). Still, since the additional information outside the primary cause of the TOR was considered less crucial, this potential trade-off might not necessarily be a safety-critical detriment in practice.

This finding also raises a broader and perhaps more fundamental question, namely whether traditional SA, as defined by Endsley (1995), is the most appropriate model to apply to the RTO context. In conditionally automated car driving, regaining comprehensive SA should, for instance, include the vehicle's speed, surrounding traffic and road environment in order to understand the car's current and anticipated states and to determine what actions might be necessary for a safe takeover (Tan & Zhang, 2025). In contrast, the operational demands of GoA3 and GoA4 level train driving differ significantly. Lateral control is inherently managed by the train running on tracks and longitudinal control during emergencies is unlikely to be required, as the onboard systems will have already initiated an emergency brake (Naumann & Thomas-Friedrich, 2024).

Thus, RTO might align more closely with the concept of distributed SA (Stanton et al., 2006), which suggests that in complex and dynamic environments, SA is not the sole responsibility of an individual, but rather emerges from the interaction of all human and system components involved in the operation. In this framework, successful system performance relies on the quality of information sharing and integration, rather than the comprehensive SA of any single operator.

Consequently, a holistic and fast restoration of SA may be helpful but less crucial for TOs. Instead, it may be sufficient for the TO to quickly understand the cause of the TOR, particularly in scenarios where coordination with dispatchers or intervention in a wider system context is necessary (e.g., alerting other TOs or stopping all trains in the same sector). Manual control might then be resumed with more delay, meaning that a slower SA build-up might not be an issue. From this standpoint, AGTORs may be particularly well suited to the RTO context. If the primary goal of the TO is to quickly understand and communicate the reason for the TOR, it might be especially useful to orientate them rapidly to the most critical elements without overburdening them with unnecessary detail.

Contrary to hypothesis 4, the benefit of AGTORs in improving recognition performance was not influenced by the presence of an ST. Instead, participants in the AGTOR group outperformed those in the SWT group consistently across both ST conditions.

Since the ST condition led to fewer recognized takeover reasons overall, further refinement of AGTORs may help improve recognition to levels comparable to the no-task condition. This might be achieved by increasing the level of detail in the information the AGTOR conveys. For instance, instead of the AGTOR guiding attention to the camera view, it could specifically indicate the location within the video stream and the system's classification of the obstacle. Yet, since semantic information is specifically prone to impose information processing demands on the TO (Liu et al., 2023), a balance needs to be found between providing as much information as possible, while keeping workload levels close to optimum. It might also be helpful to inform the TO of the trains' states even in times of routine performance. This could either be done via short status updates that are presented to the TO either visually or acoustically or by having the TO resume manual control of a train occasionally for short periods of time, even when no TOR has been issued. This is already done in GoA2 in the Paris metro system, where operators are required to drive manually from time to time (Fourtune & Butruille, 2016).

Another approach to improve AGTORs might be to incorporate additional modalities, thus creating information redundancy to support adapting to varying TO states and ST demands (Yoon et al., 2019). For example, this could involve adding a visual element, such as a bounding box around an unidentifiable obstacle, or accentuating the relevant area within the remote environment to make it more salient (Grippenkoven et al., 2020). However, such enhancements must strike a careful balance between guiding the TO's attention effectively and avoiding attentional tunnelling.

It is important to note that this study deliberately employed a perfectly reliable system for the AGTORs to avoid overwhelming participants who were unfamiliar with the environment and task. However, real-world systems will inevitably face limitations due to the sheer number and variability of possible scenarios. In such conditions, there is a risk that TOs may become overly reliant on the AGTORs, leading to complacency (Endsley, 2021). As a result, they might place excessive trust in the ATO system's identification of the takeover reason and become insensible to possible errors. This issue may be aggravated by AGTORs due to their higher acceptance ratings and favourable assessments, as evidenced in previous studies (Bazilinskyy et al., 2018; Brandenburg & Epple, 2019). Therefore, future research should examine the effects of AGTOR unreliability on TO performance and SA, especially in relation to acceptance and trust.

Influence of driving experience

Because experience has been shown to influence the efficiency of SA development in other domains (Endsley, 2006), a subsample of train drivers was included in the study. Contrary to hypothesis 5a, train drivers did not exhibit significantly higher SAGAT scores compared to non-train drivers. However, even though not statistically significant, train drivers outperformed non-train drivers in the no-task condition, but achieved nearly identical scores in the ST condition. This may reflect train drivers' habituation to prolonged periods of low-demand monitoring, which was part of the no-task condition, but not present in the ST condition.

However, assuming that the integration of an ST into the TO role is inevitable, due to its benefits in enhancing job attractiveness, ergonomics, and overall RTO system efficiency (Grippenkoven et al., 2020; Thomas-Friedrich et al., 2024a), the observed advantage of train drivers under passive monitoring conditions may become less relevant in practice.

Furthermore, contradictory to hypothesis 5b, train drivers did not outperform non-train drivers in recognizing takeover reasons, with only marginal differences observed. Consistent with this, but contrary to hypothesis 6, AGTORs did not provide significantly greater benefits for non-train drivers compared to train drivers. A possible explanation might be that specific expertise of train drivers, such as familiarity with failure types or train displays, did not effectively transfer to the experimental remote environment. Previous research suggests that experience influences SA development primarily through existing schemata and mental models (Endsley, 2006, 2021). In this study, the remote environment, although somewhat similar to an actual train environment, may have been presented in a too spatially condensed and simplified format, and thus may have disabled the activation of these mental models. This dissimilarity may also be attributed to the simple scenarios and takeover reasons. This simplification was necessary to avoid systematic disadvantages for non-train drivers, who would have been unable to understand complex train-specific cues without acquiring prior domain knowledge.

Additional findings

A number of additional findings emerged from this study. Notably, increasing age was associated with lower recognition rates of takeover reasons and lower performance in the ST. The observed differences might reflect slower initial adaptation to the new environment and task compared to younger participants, who may be more familiar with technology-rich environments. Lower recognition rates for older participants are also consistent with previous research linking ageing to declines in SA, particularly due to declines in visual attention capacity (Caserta & Abrams, 2007). Thus, the combined demands of the ST and takeover task may have overwhelmed the attentional resources available to older individuals in the sample. Similarly, age-related reduced inhibition of task demands (Caserta & Abrams, 2007) when switching from the ST to monitoring the environment may have led to recognition deficits.

The influence of age might have been underestimated in this study and requires further investigation. In other safety-critical domains such as air traffic control, age restrictions are already in place to address similar age-related declines in key cognitive abilities such as vigilance and visual attention (Broach & Schroeder, 2005). This issue may become particularly relevant in the future, as a substantial proportion of TOs could be recruited from the current pool of train drivers. Among this group, a significant percentage are of higher age: specifically, 40% of German train drivers were aged 50 or over in 2022 (Statista, 2025). Solutions to mitigate these effects in RTO contexts might include more frequent breaks (Feldhütter et al., 2019), or tailoring support systems such as AGTORs to the specific needs of different age and experience groups (Liu et al., 2023).

Another notable finding was that performance in the ST showed a strong association with the number of takeover reasons recognized. This may reflect higher general engagement or attentional investment in the study. However, the correlation with SAGAT scores was minimal, suggesting that enhanced ST performance may be more indicative of superior visual attention and the ability to allocate attentional resources effectively. This interpretation is further supported by the substantial variance in each model explained by individual intercepts, indicating strong differences in baseline performance across participants. It highlights the importance of considering individual differences in cognitive capabilities in the selection and training of TOs.

Besides these findings, response patterns in the SAGAT answers, revealed concerns regarding safety culture. When participants answered incorrectly, they predominantly selected the option indicating that no deviation from the ideal system state was present. They rarely chose either the "I don't know" option or (falsely) indicated an anomaly. This pattern suggests that when operators report no issue, because they failed to recognize an anomaly, critical safety-relevant information may be overlooked. One potential countermeasure could be to restructure the questions or report forms so that operators are first asked whether they are confident in their knowledge of the system state before being allowed to give a definitive answer.

This would also require fostering a culture that normalizes uncertainty and encourages transparency about unknowns and errors, which may help reduce the likelihood of complacency (Wiegmann et al., 2004). Emphasizing critical evaluation of received information and promoting a judgment culture that values cautious reporting over confident but incorrect assessments could thus enhance operational safety.

5.1 Limitations

This study employed a novel approach to assess SA and workload in an online RTO environment through a dynamic and interactive online setup. While this design allowed for the inclusion of a relatively large and diverse sample, several limitations should be noted.

First, the ATO system was implemented with perfect reliability and did not produce any false alarms or, in the case of AGTORs, incorrect guidance. However, automation reliability has been shown to significantly impact operator performance and attention allocation (Onnasch, 2015). Future research should therefore examine how varying levels of reliability, including system errors or ambiguous cues, might affect operator behaviour and SA.

Second, the proportion of train drivers in the sample was lower than originally intended. Although efforts were made to achieve a balanced sample, the relative scarcity of train drivers and the difficulty of recruiting them posed significant challenges. As a result, the potential influence of previous experience may have been underestimated in this study and may not have allowed for statistically significant results. Future studies with larger samples of professional train drivers are needed to explore these effects in more depth.

Additionally, while the displays in the remote environment were designed to resemble those of conventional train systems, the simplified format used in the study may have limited the applicability of professional experience. Although this approach was necessary to avoid disadvantaging participants without railway-specific knowledge, the simplified representation may have reduced the extent to which train drivers' skills could be transferred to the task. Thus, future research could employ more detailed and realistic environments.

Conducting the study online also meant that full experimental control could not be ensured. Although this design was chosen to obtain a larger sample, it introduced variability in participants' technical setups and surroundings. Nonetheless, several measures were implemented to address this, including technical compatibility checks, extended training procedures, attention checks, and the exclusion of participants who showed low engagement or inattention. In addition, the overall duration of the study was relatively short to reduce participant dropouts. However, this short duration may have limited the ability to observe effects such as monotony or fatigue, both of which are likely to be relevant in real-world RTO scenarios. Additionally, experience and continuous exposure to AGTORs have been found to improve SA (Roche et al., 2018), suggesting this might be a promising focus for future studies, that employ longer time-on-task periods.

Finally, cultural differences between rail systems may limit the generalizability of the study's findings beyond the context of Germany. This applies both to features, such as infrastructure or signal design, and to differences in safety culture and train driver training. However, this limitation may be somewhat mitigated by the fact that a substantial proportion of participants lacked specific knowledge of railway systems in the first place.

Nevertheless, future research should include studies in other countries and cultural contexts to better understand the influence of varying safety cultures and operational demands on RTO.

6 Conclusion

As automation progresses toward GoA3 and GoA4, the train driver's role will transform from being physically present on the train to serving as a TO responsible for monitoring multiple trains within a control area (Naumann & Thomas-Friedrich, 2024). This transition introduces new challenges, including risks of monotony, fatigue, and a subsequent decline in SA (Thomas-Friedrich et al., 2024a).

This study extends existing literature by addressing key human factors challenges in the scarcely researched context of RTO. Using both objective and subjective SA measures in a dynamic online environment with a diverse sample, it offers insights into the TO's role as a crucial fallback layer.

Engaging in an ST led to fewer correctly identified takeover reasons, suggesting initial task-switching costs, but was also associated with higher SAGAT scores and increased workload. This pattern indicates that while the ST may impair visual attention and recognition in the first seconds following a TOR, it may simultaneously promote overall engagement and support a more comprehensive perception of the environment. Interestingly, AGTORs were effective in mitigating the detrimental effects of the ST by directing attention toward the takeover reason first, thereby improving recognition performance. These findings suggest that AGTORs hold promise as supportive tools in future RTO environments, as they may effectively help TOs regain situation understanding following a TOR quickly, even when a demanding ST has been performed beforehand.

Contrary to expectations, prior train driving experience did not lead to improved SA, except for a slight advantage in SAGAT scores during the passive monitoring (no-task) condition. This may be attributed to the simplified and abstract nature of the simulated environment, which may not have sufficiently resembled real-world conditions to activate train driving specific knowledge, as well as the lower than intended proportion of train drivers in the sample.

Taken together, these findings highlight the importance of integrating human factors into the design of automated railway systems. If TOs are to serve as a reliable fallback layer in unanticipated situations, system and task design must support rapid SA recovery and sustained engagement. Future research should focus on more realistic remote environments, effects of system unreliability, extended time-on-task, and more systemic perspectives on SA.

References

Åkerstedt, T., & Gillberg, M. (1990). Subjective and Objective Sleepiness in the Active Individual. *International Journal of Neuroscience*, *52*(1-2), 29–37. https://doi.org/10.3109/00207459008994241

Amre, S., & Sun, Y. (2021). Effects of Non-driving Task Related Workload and Situational Awareness in Semi-autonomous Vehicles. In J. Wright, D. Barber, S. Scataglini, & S. Rajulu (Eds.), *Advances in Simulation and Digital Human Modeling* (pp. 242–250, Vol. 264). Springer. https://doi.org/10.1007/978-3-030-79763-8_30

Bainbridge, L. (1983). Ironies of Automation. In G. Johannsen & J. Rijnsdorp (Eds.), *Analysis, Design and Evaluation of Man–Machine Systems* (pp. 129–135). Pergamon. https://doi.org/10.1016/B978-0-08-029348-6.50026-9

Bazilinskyy, P., Petermeijer, S. M., Petrovych, V., Dodou, D., & de Winter, J. (2018). Takeover requests in highly automated driving: A crowdsourcing survey on auditory, vibrotactile, and visual displays. *Transportation Research Part F: Traffic Psychology and Behaviour*, *56*, 82–98. https://doi.org/10.1016/j.trf.2018.04.001

Biondi, F. N., McDonnell, A. S., Mahmoodzadeh, M., Jajo, N., & Strayer, D. L. (2023). Vigilance Decrement During On-Road Partially Automated Driving Across Four Systems. *Human Factors*, *66*(9), 2179–2190. https : / / doi . org / 10 . 1177 / 00187208231189658

Bosch, E. J., Schackmann, D., Hoyer, S., Kilian, W., Schwanitz, S., & Hamann, A. (2025). Multi-Sensor Vigilance Detection: Sensor Reliability and Comfort in Automated Train Operations. *6th German Conference on Rail Human Factors*. https://elib.dlr.de/213399/

Brandenburg, S., & Epple, S. (2019). Drivers' Individual Design Preferences of Takeover Requests in Highly Automated Driving. *i-com*, *18*(2), 167–178. https://doi.org/10.1515/icom-2018-0028

Brandenburger, N., & Naumann, A. (2018). Towards remote supervision and recovery of automated railway systems: The staff's changing contribution to system resilience. *2018 International Conference on Intelligent Rail Transportation (ICIRT)*, 1–5. https://doi.org/10.1109/ICIRT.2018.8641576

Brandenburger, N., & Naumann, A. (2019a). Der Arbeitsplatz des Train Operator im Fokus. *Der Eisenbahningenieur EI*, (11), 13–17. https://elib.dlr.de/130565/

Brandenburger, N., & Naumann, A. (2019b). On Track: A Series of Research about the Effects of Increasing Railway Automation on the Train Driver. *IFAC-PapersOnLine*, *52*(19), 288–293. https://doi.org/10.1016/j.ifacol.2019.12.115

Brandenburger, N., Naumann, A., & Jipp, M. (2021). Task-induced fatigue when implementing high grades of railway automation. *Cognition, Technology & Work*, *23*, 273–283. https://doi.org/10.1007/s10111-019-00613-z

Brandenburger, N., Thomas- Friedrich, B., Naumann, A., & Grippenkoven, J. D. (2018). Automation in Railway Operations: Effects on Signaller and Train Driver Workload. In B. Milius & A. Naumann (Eds.), *Proceedings of the 3rd German Workshop on Rail Human Factors* (pp. 51–60). ITS Mobility. https://elib.dlr.de/119746/

Broach, D., & Schroeder, D. (2005). *Review of the Scientific Basis for the Mandatory Separation of an Air Traffic Control Specialist at Age 56* (Technical Report). Office of Aerospace Medicine, Federal Aviation Administration. https://apps.dtic.mil/sti/pdfs/ADA460056.pdf

Brooke, J. (1996). SUS: A 'Quick and Dirty' Usability Scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, & I. L. McClelland (Eds.), *Usability Evaluation in Industry* (pp. 189–194). Taylor & Francis. https://doi.org/10.1201/9781498710411-35

Broughton, A., Tanis, J., Brambilla, M., Voss, E., & Vitols, K. (2024, March). *Trends, challenges and opportunities in the EU transport labour market*. European Parliament, Directorate--General for Parliamentary Research Services. Brussels. https://www.europarl.europa.eu/RegData/etudes/STUD/2024/747266/IPOL_STU(2024)747266_EN.pdf

Cabrall, C. D. D., Eriksson, A., Dreger, F., Happee, R., & de Winter and, J. (2019). How to keep drivers engaged while supervising driving automation? A literature survey and categorisation of six solution areas. *Theoretical Issues in Ergonomics Science*, *20*(3), 332–365. https://doi.org/10.1080/1463922X.2018.1528484

Caserta, R. J., & Abrams, L. (2007). The relevance of situation awareness in older adults' cognitive functioning: A review. *European Review of Aging and Physical Activity*, *4*, 3–13. https://doi.org/10.1007/s11556-007-0018-x

Chai, C., Lei, Y., Wei, H., Wu, C., Zhang, W., Hansen, P., Fan, H., & Shi, J. (2024). The effects of various auditory takeover requests: A simulated driving study considering the modality of non-driving-related tasks. *Applied Ergonomics*, *118*, 104252. https://doi.org/10.1016/j.apergo.2024.104252

Chen, Q., Li, J., & Tei, K. (2023). Attention-guiding Takeover Requests for Situation Awareness in Semi-autonomous Driving. *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, 416–421. https://doi.org/10.1145/3568294.3580118

Choi, D., Sato, T., Ando, T., Abe, T., Akamatsu, M., & Kitazaki, S. (2020). Effects of cognitive and visual loads on driving performance after take-over request (TOR) in automated driving. *Applied ergonomics*, *85*. https://doi.org/10.1016/j.apergo.2020.103074

Colley, M., Gruler, L., Woide, M., & Rukzio, E. (2021). Investigating the Design of Information Presentation in Take-Over Requests in Automated Vehicles. *Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction*. https://doi.org/10.1145/3447526.3472025

DB InfraGO AG. (2009). Ril 483.0202A03: Darstellung von Beispielen der Entlassung aus der LZB-Führung [Guideline, valid from April 15, 2009]. *DB InfraGO AG*. https://www.dbinfrago.com/resource/blob/12596202/0429c1faa03a6b07465367c0bd417d87/Ril-483-0202A03-NBN-2025-data.pdf

Deng, M., Gluck, A., Zhao, Y., Da Li, Menassa, C. C., Kamat, V. R., & Brinkley, J. (2024). An analysis of physiological responses as indicators of driver takeover readiness in conditionally automated driving. *Accident Analysis & Prevention*, *195*. https://doi.org/10.1016/j.aap.2023.107372

Deutsche Bahn AG. (2024). Integrierter Bericht 2024. https://ibir.deutschebahn.com/2024/de/start/

Deutsche Bahn AG. (2025). *Digitale S-Bahn Hamburg*. Retrieved April 10, 2025, from https://digitale-schiene-deutschland.de/de/projekte/DigitaleS-Bahn-Hamburg

Digitale Schiene Deutschland. (2024, May). *Fully Automated Driving in Germany: World Premiere in Hamburg*. Retrieved May 20, 2025, from https://digitale-schiene-deutschland.de/en/news/2024/Fullyautomateddriving

Dorrian, J., Roach, G. D., Fletcher, A., & Dawson, D. (2007). Simulated train driving: Fatigue, self-awareness and cognitive disengagement. *Applied Ergonomics*, *38*(2), 155–166. https://doi.org/10.1016/j.apergo.2006.03.006

Dovetail Games. (2023). Train Sim World 4 [Software]. https://trainsimworld.com/

Du, Y., & Zhi, J. (2022). Impacts of Attention Level on Manual Take-Over Performance in Automatic Driving on High-Speed Railways. *International Journal of Human–*

Computer Interaction, *40*(3), 603–612. https://doi.org/10.1080/10447318.2022.2121455

Dunn, N., & Williamson, A. (2012). Driving monotonous routes in a train simulator: the effect of task demand on driving performance and subjective experience. *Ergonomics*, *55*(9), 997–1008. https://doi.org/10.1080/00140139.2012.691994

ElevenLabs. (2025). *Elevenlabs voice generation*. https://www.elevenlabs.io/

Endsley, M. R. (1988). Situation awareness global assessment technique (SAGAT). *Proceedings of the IEEE 1988 National Aerospace and Electronics Conference*, *3*, 789–795. https://doi.org/10.1109/NAECON.1988.195097

Endsley, M. R. (1995). Toward a Theory of Situation Awareness in Dynamic Systems. *Human Factors*, *37*(1), 32–64. https://doi.org/10.1518/001872095779049543

Endsley, M. R. (2006). Expertise and Situation Awareness. In K. A. Ericsson, N. Charness, P. J. Feltovich, & R. R. Hoffman (Eds.), *The Cambridge Handbook of Expertise and Expert Performance* (pp. 633–651). Cambridge University Press. https://doi.org/10.1017/CBO9780511816796.036

Endsley, M. R. (2015). Situation Awareness Misconceptions and Misunderstandings. *Journal of Cognitive Engineering and Decision Making*, *9*(1), 4–32. https://doi.org/10.1177/1555343415572631

Endsley, M. R. (2019). Situation Awareness in Future Autonomous Vehicles: Beware of the Unexpected. In S. Bagnara, R. Tartaglia, S. Albolino, T. Alexander, & Y. Fujita (Eds.), *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018)* (pp. 395–402, Vol. 824). Springer. https://doi.org/10.1007/978-3-319-96071-5_32

Endsley, M. R. (2021). Situation Awareness. In G. Salvendy & W. Karwowski (Eds.), *Handbook of Human Factors and Ergonomics* (pp. 434–455). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119636113.ch17

Epic Games. (2023). *Unreal engine* (Version 4.26) [Software]. https://www.unrealengine.com

Evans, A. W. (2011). Fatal train accidents on Europe's railways: 1980–2009. *Accident Analysis & Prevention*, *43*(1), 391–401. https://doi.org/10.1016/j.aap.2010.09.009

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191. https://doi.org/10.3758/BF03193146

Feldhütter, A., Hecht, T., Kalb, L., & Bengler, K. (2019). Effect of prolonged periods of conditionally automated driving on the development of fatigue: with and without non-driving-related activities. *Cognition, Technology & Work*, *21*(1), 33–40. https://doi.org/10.1007/s10111-018-0524-9

Ferdinand, A. O., & Menachemi, N. (2014). Associations Between Driving Performance and Engaging in Secondary Tasks: A Systematic Review. *American Journal of Public Health*, *104*(3), 39–48. https://doi.org/10.2105/AJPH.2013.301750

Filtness, A., & Naweed, A. (2017). Causes, consequences and countermeasures to driver fatigue in the rail industry: The train driver perspective. *Applied Ergonomics*, *60*, 12–21. https://doi.org/10.1016/j.apergo.2016.10.009

Fourtune, L., & Butruille, N. (2016). Paris Metro and Automatic Driving Systems. *Proceedings of the International Railway Safety Council (IRSC)*. https://international-railway-safety-council.com/wp-content/uploads/2017/09/fourtune-the-paris-metro-and-automatic-driving-systems.pdf

Greenlee, E. T., DeLucia, P. R., & Newton, D. C. (2022). Driver Vigilance Decrement is More Severe During Automated Driving than Manual Driving. *Human Factors*, *66*(2), 574–588. https://doi.org/10.1177/00187208221103922

Grippenkoven, J. D., Meirich, C., Roth, M. H., Caspar, M., & Hungar, H. (2020). Teleoperierte Triebfahrzeugführung als Rückfallebene der Hochautomation / Teleoperation as a fallback solution for highly automated rail traffic. *SIGNAL + DRAHT*, (6), 6–13. https://elib.dlr.de/135415/

Grippenkoven, J. D., Rodd, J., & Brandenburger, N. (2018, March). DLR-WAT: Ein Instrument zur Untersuchung des optimalen Beanspruchungsniveaus in hochautomatisierten Mensch-Maschine-Systemen. In ITS mobility e.V. (Ed.), *AAET Automatisiertes & Vernetztes Fahren* (pp. 199–213). https://elib.dlr.de/119443/

Gruden, T., Tomažič, S., Sodnik, J., & Jakus, G. (2022). A user study of directional tactile and auditory user interfaces for take-over requests in conditionally automated vehicles. *Accident Analysis & Prevention*, *174*, 106766. https://doi.org/10.1016/j.aap.2022.106766

Hancock, G. M., Longo, L., Young, M. S., & Hancock, P. A. (2021). Mental Workload. In G. Salvendy & W. Karwowski (Eds.), *Handbook of Human Factors and Ergonomics* (pp. 203–226). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119636113.ch7

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In P. A. Hancock & N. Meshkati (Eds.), *Human Mental Workload* (pp. 139–183, Vol. 52). North-Holland. https://doi.org/10.1016/S0166-4115(08)62386-9

Hely, M., Shardlow, T., Butt, B., Friswell, R., McIntosh, A., & Williamson, A. (2015). Effects of automatic train protection on human factors and driving behaviour. *Proceedings 19th Triennial Congress of the IEA*, *9*, 14. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=735896e2df9bff15877f93f22432bcf0d16927b6

Henninger, F., Shevchenko, Y., Mertens, U. K., Kieslich, P. J., & Hilbig, B. E. (2021). lab.js: A free, open, online study builder. *Behavior Research Methods*, *54*, 556–573. https://doi.org/10.3758/s13428-019-01283-5

Hitachi Rail. (2025). *Driverless Trains*. Retrieved June 1, 2025, from https://www.hitachirail.com/products-and-solutions/rolling-stock/driverless-trains/

Huang, G., & Pitts, B. J. (2022). Takeover requests for automated driving: The effects of signal direction, lead time, and modality on takeover performance. *Accident Analysis & Prevention*, *165*, 106534. https://doi.org/10.1016/j.aap.2021.106534

IEC. (2009). *IEC 62267:2009 – Railway applications – Automated urban guided transport (AUGT) – Safety requirements* (International Standard). International Electrotechnical Commission. Geneva, Switzerland. https://webstore.iec.ch/en/publication/6681

Janicki, J. (2018). *Railway system knowledge: How the German rail system works* (1st edition). Bahn Fachverlag.

Kim, T., & Biocca, F. (1997). Telepresence via Television: Two Dimensions of Telepresence May Have Different Connections to Memory and Persuasion. *Journal of Computer-Mediated Communication*, *3*(2), 325. https://doi.org/10.1111/j.1083-6101.1997.tb00073.x

Leiner, D. J. (2019). Too Fast, too Straight, too Weird: Non-Reactive Indicators for Meaningless Data in Internet Surveys. *Survey Research Methods*, *13*(3), 229–248. https://doi.org/10.18148/srm/2019.v13i3.7403

Liu, W., Li, Q., Wang, Z., Wang, W., Zeng, C., & Cheng, B. (2023). A Literature Review on Additional Semantic Information Conveyed from Driving Automation Systems to Drivers through Advanced In-Vehicle HMI Just Before, During, and Right After

Takeover Request. *International Journal of Human–Computer Interaction*, *39*(10), 1995–2015. https://doi.org/10.1080/10447318.2022.2074669

Lu, G., Zhai, J., Li, P., Chen, F., & Liang, L. (2021). Measuring drivers' takeover performance in varying levels of automation: Considering the influence of cognitive secondary task. *Transportation Research Part F: Traffic Psychology and Behaviour*, *82*, 96–110. https://doi.org/10.1016/j.trf.2021.08.005

Masson, É., Richard, P., Boussif, A., Gadmer, Q., David, P., Robert, E., & Vitry, C. (2022). TC-Rail: Railways Remote Driving further experiments. *13th World Congress on Railway Research (WCRR)*. https://hal.science/hal-04662695

Masson, É., Richard, P., Garcia-Guillen, S., & Morral, A. G. (2019). TC-Rail: Railways Remote Driving. *Proceedings of the 12th World Congress on Railway Research*, 1–5. https://hal.science/hal-04662611

Microsoft Corporation. (2024). *Microsoft PowerPoint* (Version 16.0). https://www.microsoft.com/en-us/microsoft-365/powerpoint

Miller, D., Sun, A., Johns, M., Ive, H., Sirkin, D., Aich, S., & Ju, W. (2015). Distraction Becomes Engagement in Automated Driving. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *59*(1), 1676–1680. https://doi.org/10.1177/1541931215591362

Mishler, S., & Chen, J. (2023). Boring But Demanding: Using Secondary Tasks to Counter the Driver Vigilance Decrement for Partially Automated Driving. *Human Factors*, *66*(6). https://doi.org/10.1177/00187208231168697

Müller, R., & Drax, C. (2014). Aircraft Accident Investigations. In R. Müller, A. Wittmer, & C. Drax (Eds.), *Aviation Risk and Safety Management: Methods and Applications in Aviation Organizations* (pp. 139–148). Springer International Publishing. https://doi.org/10.1007/978-3-319-02780-7_12

Naumann, A., & Thomas-Friedrich, B. (2024). Automation in Railway: How tasks and roles of operational staff change. *Advances in Human Factors of Transportation*, *148*, 670–677. https://doi.org/10.54941/ahfe1005259

Onnasch, L. (2015). Crossing the boundaries of automation—Function allocation and reliability. *International Journal of Human-Computer Studies*, *76*, 12–21. https://doi.org/10.1016/j.ijhcs.2014.12.004

Onnasch, L., Wickens, C. D., Li, H., & Manzey, D. (2014). Human Performance Consequences of Stages and Levels of Automation: An Integrated Meta-Analysis. *Human Factors*, *56*(3), 476–488. https://doi.org/10.1177/0018720813501549

Parasuraman, R., Sheridan, T., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, *30*(3), 286–297. https://doi.org/10.1109/3468.844354

R Core Team. (2022). *R: A Language and Environment for Statistical Computing* (Version 4.2.2). Vienna, Austria. https://www.R-project.org/

Roche, F., Somieski, A., & Brandenburg, S. (2018). Behavioral Changes to Repeated Takeovers in Highly Automated Driving: Effects of the Takeover-Request Design and the Nondriving-Related Task Modality. *Human Factors*, *61*(5), 839–849. https://doi.org/10.1177/0018720818814963

Salmon, P. M., Regan, M. A., & Johnston, I. R. (2005). *Human Error and Road Transport: Phase One – Literature Review* (Report No. 256). Monash University. Clayton, Victoria, Australia. https://research.usc.edu.au/esploro/outputs/report/Human-error-in-road-transport-Phase/99450051702621

Schackmann, D. (2025a, January). Online Study on Remote Operation in Teleoperated Railways [Study Repository on OSF]. https://doi.org/10.17605/OSF.IO/2MXTR

Schackmann, D. (2025b, January). Remote Operation in Teleoperated Railways: Impact of a Non-Driving-Related Task, Attention-Guiding Take-Over Requests and Work Experience on Situational Awareness and Workload [Preregistration on OSF]. https://doi.org/10.17605/OSF.IO/AVQJB

Schloerb, D. W. (1995). A Quantitative Measure of Telepresence. *Presence: Teleoperators & Virtual Environments*, *4*(1), 64–80. https://doi.org/10.1162/pres.1995.4.1.64

Schöne, S., Mönsters, M., Käthner, D., & Brandenburger, N. (2023). ATO-Cargo: Operating procedures for the fallback levels of highly automated railway operation. *SIGNAL + DRAHT*, *115*(10), 18–25. https://elib.dlr.de/195223/

Sharp Grant, J. (1971). Concepts of Fatigue and Vigilance in Relation to Railway Operation. *Ergonomics*, *14*(1), 111–118. https://doi.org/10.1080/00140137108931229

Shi, E., & Bengler, K. (2022). Non-driving related tasks' effects on takeover and manual driving behavior in a real driving setting: A differentiation approach based on task switching and modality shifting. *Accident Analysis & Prevention*, *178*, 106844. https://doi.org/10.1016/j.aap.2022.106844

Spring, P., McIntosh, A., Caponecchia, C., & Baysari, M. T. (2009). Level of automation: Effects on train driver vigilance. In J. R. Wilson, A. Mills, T. Clarke, J. Rajan, & N. Dadashi (Eds.), *Rail human factors around the world* (pp. 792–801). CRC Press/Balkema.

Stanton, N. A., Stewart, R., Harris, D., Houghton, R. J., Baber, C., McMaster, R., Salmon, P., Hoyle, G., Walker, G., Young, M. S., Linsell, M., Dymott, R., & Green, D. (2006). Distributed situation awareness in dynamic systems: Theoretical development and application of an ergonomics methodology. *Ergonomics*, *49*(12-13), 1288–1311. https://doi.org/10.1080/00140130600612762

Statista. (2025). *Anzahl und Altersstruktur der Lokführerinnen in Deutschland*. Retrieved May 20, 2025, from https://de.statista.com/infografik/32000/anzahl-und-altersstruktur-der-lokfuehrerinnen-in-deutschland/

Stene, T. M. (2018). Automation of the Rail – Removing the Human Factor? In S.-E. Haugen, A. Barros, C. van Gulijk, T. Kongsvik, & J. E. Vinnem (Eds.), *Safety and Reliability – Safe Societies in a Changing World: Proceedings of ESREL 2018, June 17–21, 2018, Trondheim, Norway* (1st ed., pp. 1947–1955). CRC Press. https://doi.org/10.1201/9781351174664-244

Tan, X., & Zhang, Y. (2025). Driver Situation Awareness for Regaining Control from Conditionally Automated Vehicles: A Systematic Review of Empirical Studies. *Human Factors*, *67*(4), 367–403. https://doi.org/10.1177/00187208241272071

Thomas-Friedrich, B., Bassin, N., Brosch, F., Huber, A., & Michaelsen, R. (2024a). Vermeidung von Monotonie am Tele-Tf-Arbeitsplatz. *Der Eisenbahningenieur EI*, 134–137. https://elib.dlr.de/206872/

Thomas-Friedrich, B., Schöne, S., & Mühl, K. (2024b). Tele-Tf: Ein neues Berufsbild entsteht. *ETR-Eisenbahntechnische Rundschau*, (11), 37–41. https://elib.dlr.de/210122/

UNIFE. (2021). Factsheet #28 - ERTMS Advantages. Retrieved June 1, 2025, from https://www.ertms.net/wp-content/uploads/2021/06/28.-ERTMS-Advantages-factsheet_final.pdf

Vandierendonck, A., Liefooghe, B., & Verbruggen, F. (2010). Task switching: Interplay of reconfiguration and interference control. *Psychological Bulletin*, *136*(4), 601–626. https://doi.org/10.1037/a0019791

Vogelpohl, T., Kühn, M., Hummel, T., Gehlert, T., & Vollrath, M. (2018). Transitioning to manual driving requires additional time after automation deactivation. *Transportation Research Part F: Traffic Psychology and Behaviour*, *55*, 464–482. https://doi.org/10.1016/j.trf.2018.03.019

Walch, M., Lange, K., Baumann, M., & Weber, M. (2015). Autonomous driving: Investigating the feasibility of car-driver handover assistance. *Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 11–18. https://doi.org/10.1145/2799250.2799268

Wang, A., Guo, B., Yi, Z., & Fang, W. (2023). Research on Enhanced Situation Awareness Model with DMI Visualization Cues for High-Speed Train Driving. *International Journal of Human-Computer Interaction*, *40*(20), 6185–6199. https://doi.org/10.1080/10447318.2023.2247613

Wardrop, A. W. (2019). Autonomous freight trains in australia. *8th International Conference on Railway Operations Modelling and Analysis (ICROMA), Norrköping, Sweden, June 17th – 20th, 2019*, *69*, 1120–1130.

Weaver, B. W., & DeLucia, P. R. (2022). A Systematic Review and Meta-Analysis of Takeover Performance During Conditionally Automated Driving. *Human Factors*, *64*(7), 1227–1260. https://doi.org/10.1177/0018720820976476

Wickens, C. D. (1980). The Structure of Attentional Resources. In R. S. Nickerson (Ed.), *Attention and Performance VIII* (pp. 239–257). Psychology Press. https://doi.org/10.4324/9781315802961

Wickens, C. D. (2008). Multiple Resources and Mental Workload. *Human Factors*, *50*(3), 449–455. https://doi.org/10.1518/001872008X288394

Wickens, C. D., Li, H., Santamaria, A., Sebok, A., & Sarter, N. (2010). Stages and Levels of Automation: An Integrated Meta-analysis. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *54*(4), 389–393. https://doi.org/10.1177/154193121005400425

Wiegmann, D. A., Zhang, H., Von Thaden, T. L., Sharma, G., & Gibbons, A. M. (2004). Safety Culture: An Integrative Review. *The International Journal of Aviation Psychology*, *14*(2), 117–134. https://doi.org/10.1207/s15327108ijap1402_1

Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, *1*(6), 80–83. https://doi.org/10.2307/3001968

Xu, J., Fard, M., Zhang, N., Davy, J. L., & Robinson, S. R. (2024). Cognitive load and task switching in drivers: Implications for road safety in semi-autonomous vehicles. *Transportation Research Part F: Traffic Psychology and Behaviour*, *107*, 1175–1197. https://doi.org/10.1016/j.trf.2024.11.005

Yoon, S. H., Kim, Y. W., & Ji, Y. G. (2019). The effects of takeover request modalities on highly automated car control transitions. *Accident Analysis & Prevention*, *123*, 150–158. https://doi.org/10.1016/j.aap.2018.11.018

Yun, H., & Yang, J. H. (2020). Multimodal warning design for take-over request in conditionally automated driving. *European Transport Research Review*, *12*(34). https://doi.org/10.1186/s12544-020-00427-5

Zhang, B., De Winter, J., Varotto, S., Happee, R., & Martens, M. (2019). Determinants of take-over time from automated driving: A meta-analysis of 129 studies. *Transportation Research Part F: Traffic Psychology and Behaviour*, *64*, 285–307. https://doi.org/10.1016/j.trf.2019.04.020

Zhang, X., Li, J., Liu, Y., Zhang, Z., Wang, Z., Luo, D., Zhou, X., Zhu, M., Salman, W., Hu, G., & Wang, C. (2017). Design of a Fatigue Detection System for High-Speed Trains Based on Driver Vigilance Using a Wireless Wearable EEG. *Sensors*, *17*(3), 86. https://doi.org/10.3390/s17030486

Zhangyu, W., Guizhen, Y., Xinkai, W., Haoran, L., & Da, L. (2021). A Camera and LiDAR Data Fusion Method for Railway Object Detection. *IEEE Sensors Journal*, *21*(12), 13442–13454. https://doi.org/10.1109/JSEN.2021.3066714

Usage of AI Tools

The following AI tools have been used in the work on this study and thesis, and for the following purposes:

DeepL Translate[2]

- Translation of text from German to English
- Language refinement and vocabulary enhancement

DeepL Write[3]

- Sentence restructuring for improved clarity and flow
- Grammar and style checks

OpenAI's ChatGPT[4]

- Support with JavaScript and HTML code for the setup of the experimental framework
- Assistance with Python scripting to preprocess experimental data exported from lab.js
- Support with R code to preprocess and merge experimental and survey data
- Language refinement and coherence improvement, including rewording existing content (no original content was generated)

[2]https://www.deepl.com/de/translator
[3]https://www.deepl.com/de/write
[4]https://chatgpt.com/

A Appendix

A.1 Karolinska Sleepiness Scale (Åkerstedt & Gillberg, 1990)

Translated to German

Wie schläfrig fühlen Sie sich in diesem Moment? Kreuzen Sie bitte Zutreffendes an.

1. äußerst wach
2. sehr wach
3. normal wach
4. ziemlich wach
5. weder wach noch schläfrig
6. schläfrig
7. schläfrig, ohne Mühe wach zu bleiben
8. schläfrig, etwas Mühe wach zu bleiben
9. sehr schläfrig, große Mühe wach zu bleiben
10. äußerst schläfrig, kann nicht wach bleiben

A.2 DLR-WAT (Grippenkoven et al., 2018)

Bitte beurteilen Sie Ihre Beanspruchung infolge der aktuellen Aufgabe anhand der folgenden Kriterien. Die aktuelle Aufgabe umfasst alles, was Sie gerade im Vollbildmodus bearbeitet haben.

1. Wie sehr waren Sie während der Gesamtaufgabe durch Informationssuche- und Aufnahme beansprucht?
2. Wie sehr waren Sie durch den Abruf von relevantem Wissen während der Bearbeitung der Gesamtaufgabe beansprucht?
3. Wie sehr wurden Sie durch Entscheidungsfindung und Handlungsauswahl während der Gesamtaufgabe beansprucht?
4. Wie sehr waren Sie durch die Gesamtaufgabe motorisch und körperlich beansprucht?
5. Wie sehr haben Sie sich durch die Gesamtaufgabe unter Zeitdruck gefühlt?
6. Wie sehr mussten Sie sich anstrengen (geistig und körperlich) um Ihre Leistung in der Gesamtaufgabe zu erbringen?
7. Wie frustriert haben Sie sich während der Bearbeitung der Gesamtaufgabe gefühlt?
8. Wie beurteilen Sie Ihre Leistung in der Bewältigung der Gesamtaufgabe?

A.3 Telepresence questionnaire (Kim & Biocca, 1997)

Adapted and translated to German

Bitte beantworten Sie die folgenden Fragen:

1. Als die Aufgabe zu Ende war, hatte ich das Gefühl, nach einer Reise in die „reale Welt“ zurückzukehren.
2. Die virtuelle Umgebung kam zu mir und schuf eine neue Welt für mich, und diese Welt verschwand plötzlich, als die Aufgabe beendet war.
3. Während der Aufgabe hatte ich das Gefühl, mich in der von der Umgebung geschaffenen Welt zu befinden.
4. Während der Aufgabe habe ich niemals vergessen, dass ich mich mitten in einem Experiment befand.
5. Während der Aufgabe befand sich mein Körper im Raum, aber gedanklich war ich in der von der virtuellen Umgebung geschaffenen Welt.
6. Während der Aufgabe war die virtuelle Umgebung für mich realer oder präsenter als die „reale Welt“.
7. Die virtuelle Umgebung erschien mir nur als „etwas, das ich gesehen habe“, und nicht „etwas, wovon ich ein Teil war“.
8. Während der Aufgabe war ich mit meinen Gedanken in der realen Welt und nicht in der von der virtuellen Umgebung geschaffenen Welt.

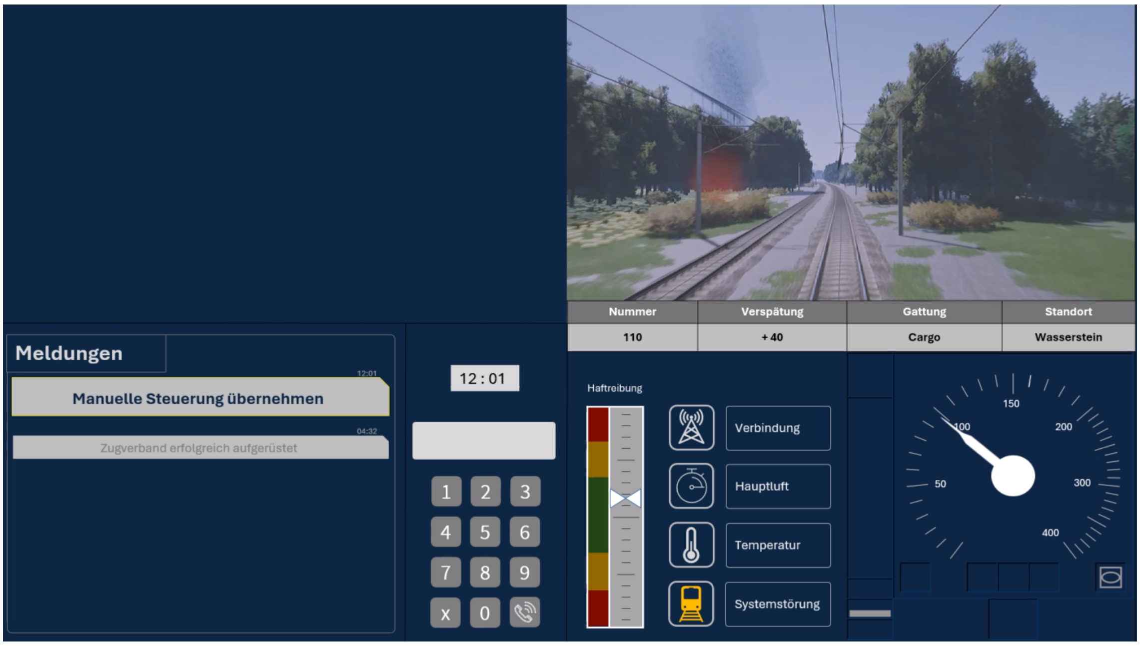
A.4 System Usability Scale (Brooke, 1996)

Translated to German

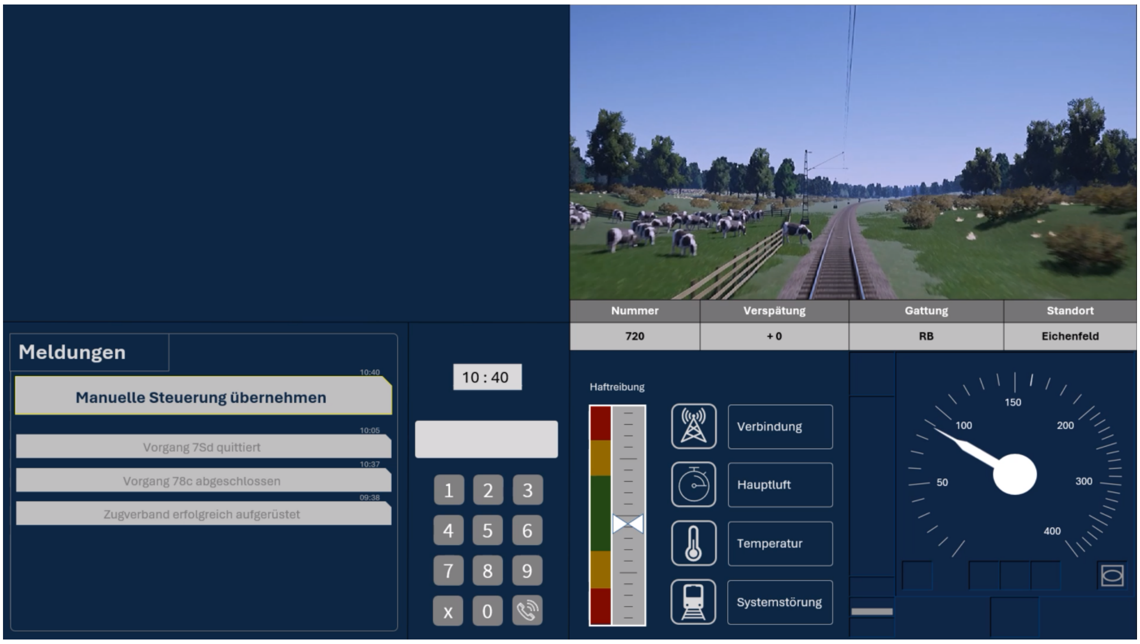
Bitte beurteilen Sie noch die gerade genutzte Arbeitsumgebung anhand der folgenden Aussagen. Wichtig: Die Nebenaufgabe ist NICHT Teil der Arbeitsumgebung und sollte nicht in die Bewertung miteinfließen.

1. Ich denke, ich würde die Arbeitsumgebung gerne häufiger benutzen.
2. Ich finde die Arbeitsumgebung unnötig komplex.
3. Ich finde, die Arbeitsumgebung ist einfach zu benutzen.
4. Ich denke, ich würde die Unterstützung einer erfahreneren Person brauchen, um in der Lage zu sein, die Arbeitsumgebung zu benutzen.
5. Ich finde, die verschiedenen Funktionen in dieser Arbeitsumgebung sind gut integriert.
6. Ich denke, es gibt zu viele Inkonsistenzen in dieser Arbeitsumgebung.
7. Ich könnte mir vorstellen, dass die meisten Leute sehr schnell lernen würden mit dieser Arbeitsumgebung umzugehen.
8. Ich fand die Arbeitsumgebung sehr schwerfällig im Gebrauch.
9. Ich fühlte mich sehr sicher bei der Benutzung der Arbeitsumgebung.
10. Ich musste eine Menge lernen, bevor ich mit dieser Arbeitsumgebung zurechtkam.

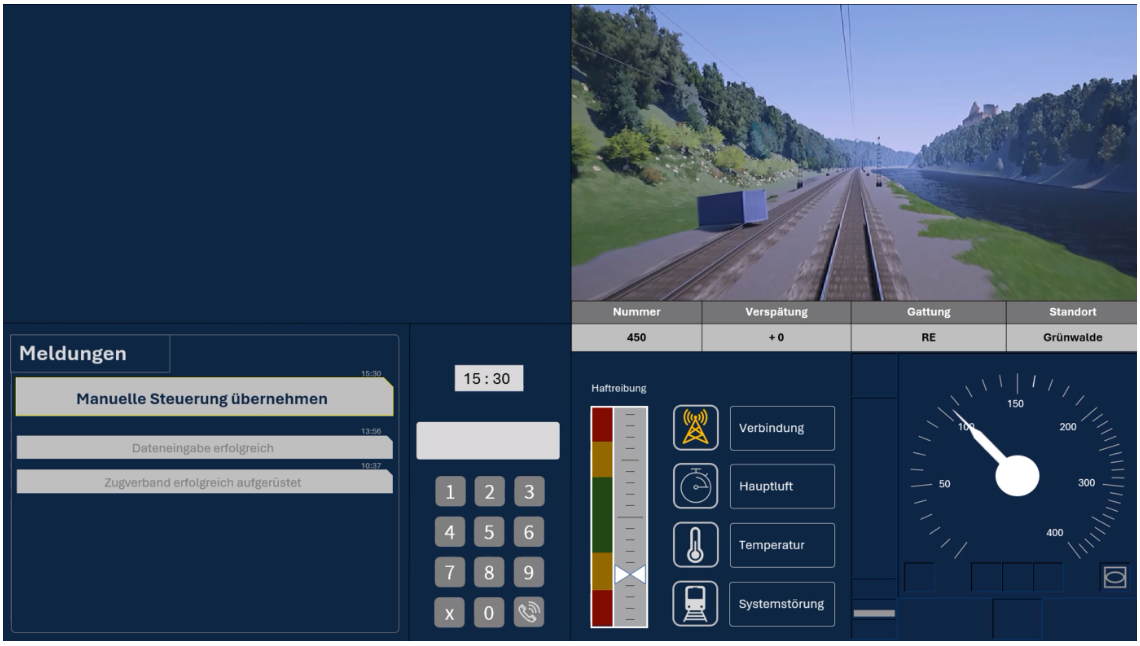
A.5 Overview of all scenarios with corresponding reasons for the TOR.



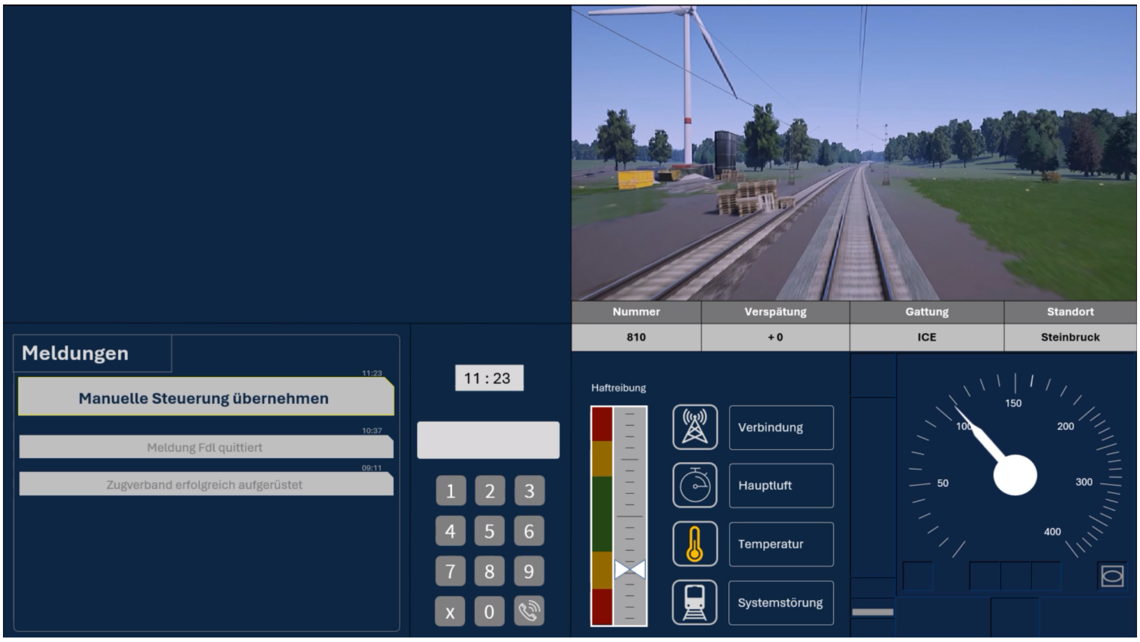
(1) Line-side fire



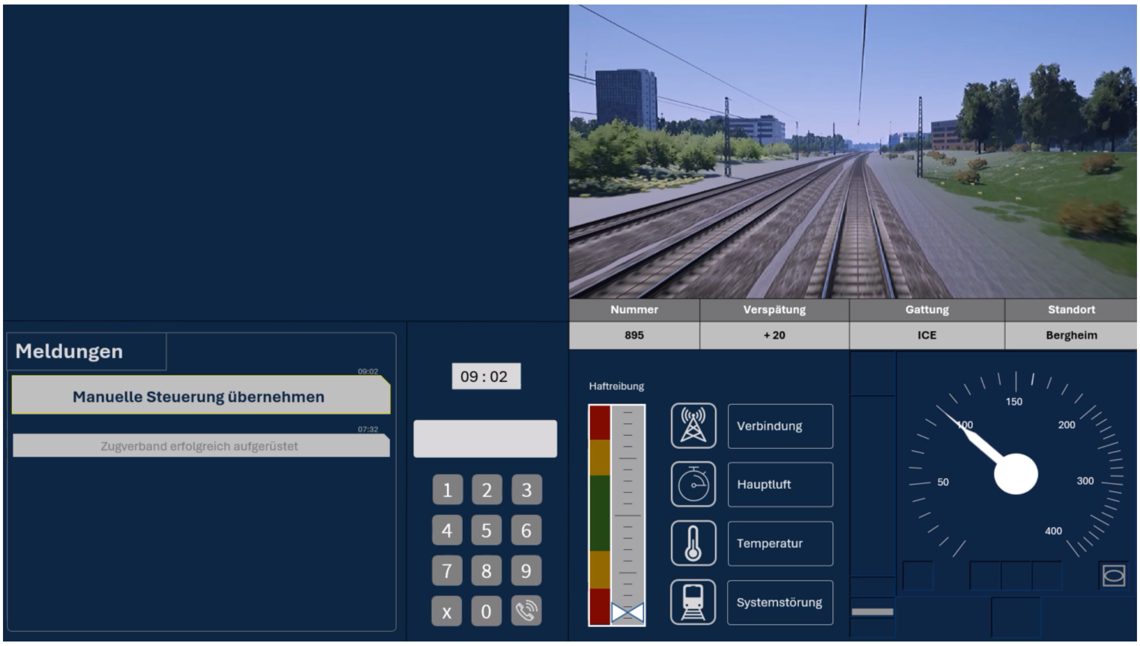
(2) Cow close to track



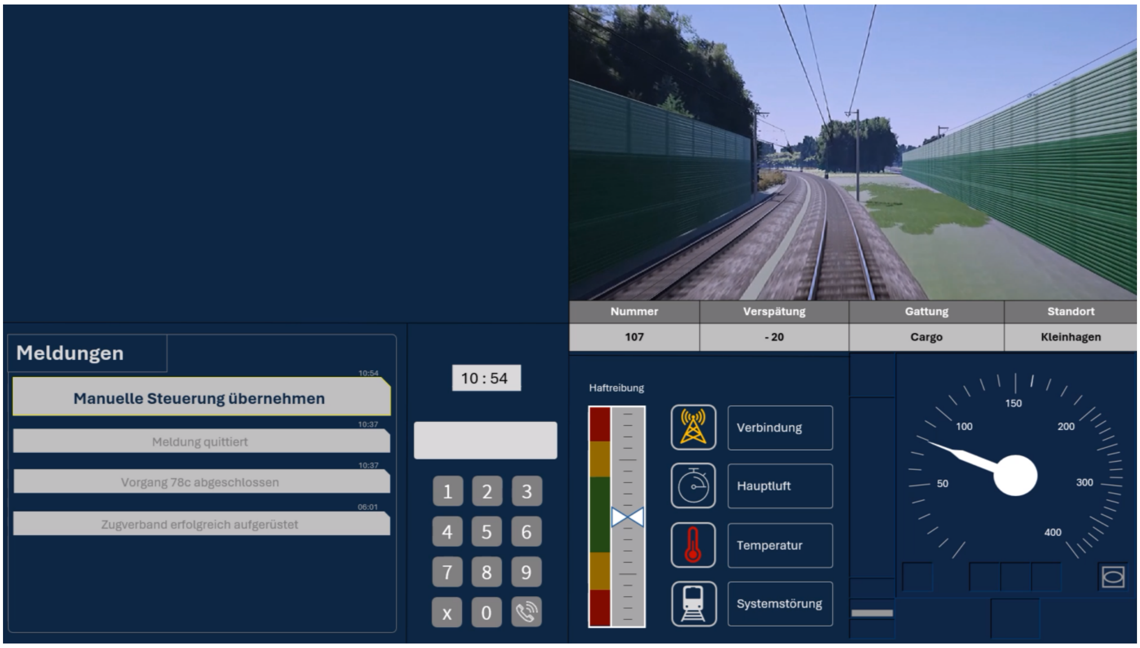
(3) Container on the opposite track



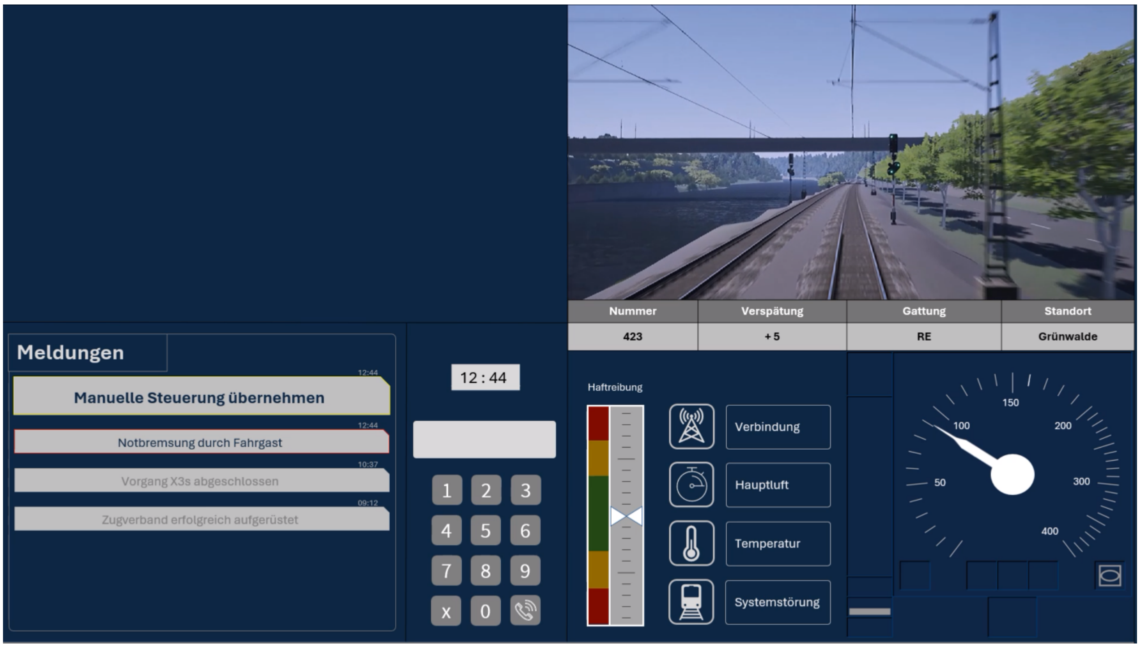
(4) Pallets on the opposite track



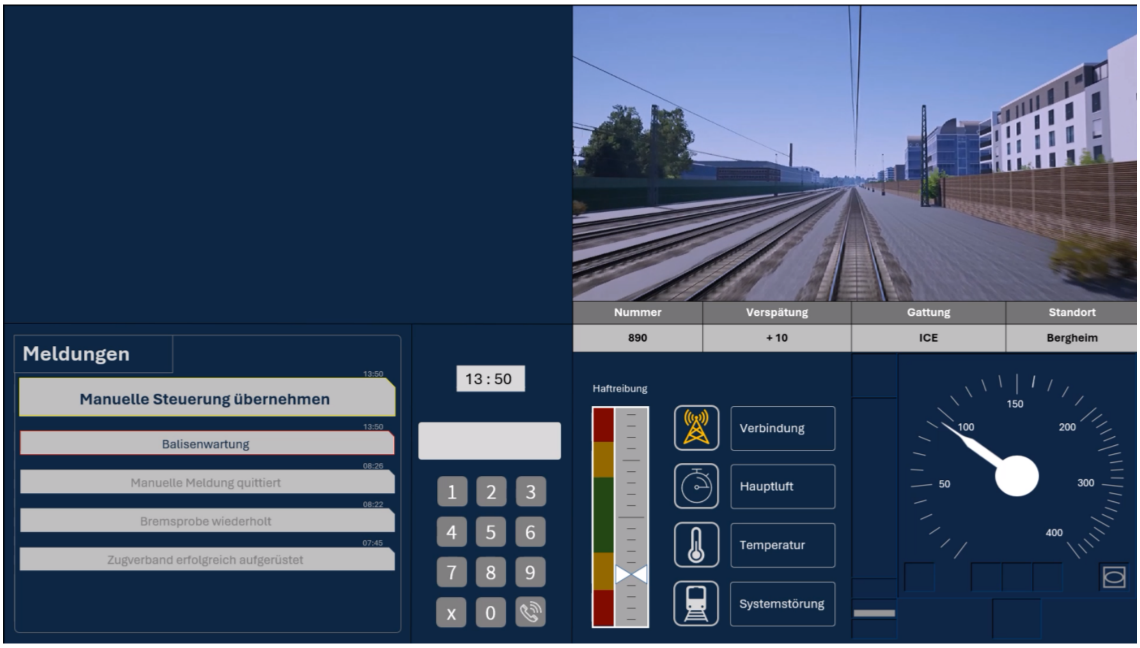
(5) Friction in critical range



(6) Temperature in critical range



(7) Passenger-initiated emergency brake



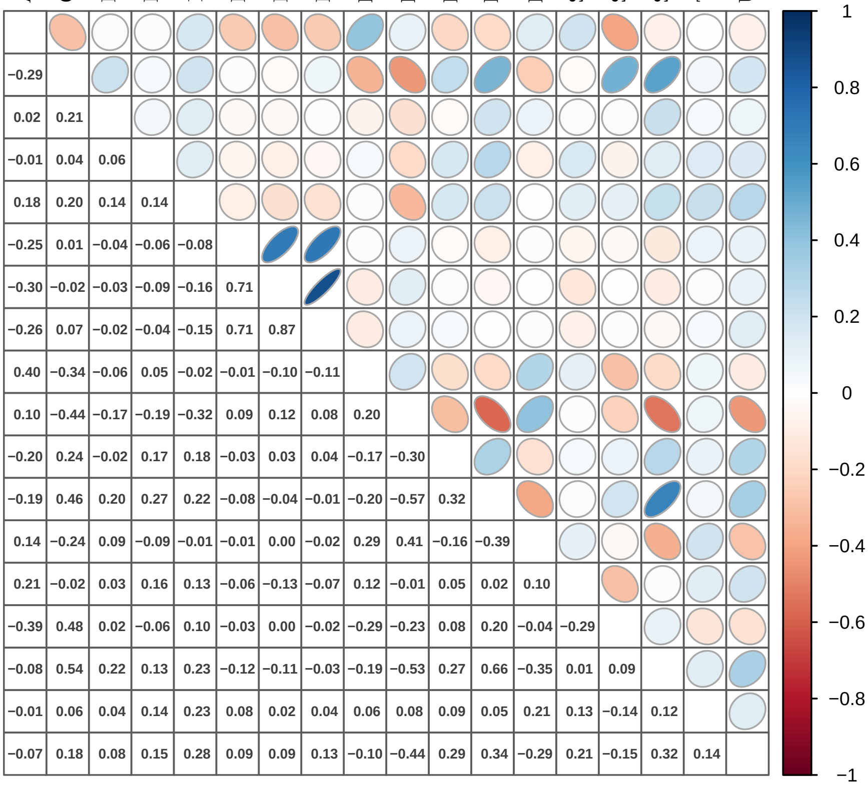
(8) Balise maintenance

A.6 Form to fill in the reason for the takeover

A.7 Participants' *n* for key variables in the study.

Variable	**Group**	***n* included**
ST condition	ST	157
	no-task	160
TOR condition	AGTOR	81
	SWT	79
Train driver (yes)	AGTOR	16
	SWT	12
	ST	27
	no-task	28
Workload	ST	160
	no-task	160
Subjective SA	ST	157
	no-task	160
Usability	ST	160
Telepresence	ST	160
Fun & Diligence	ST	160

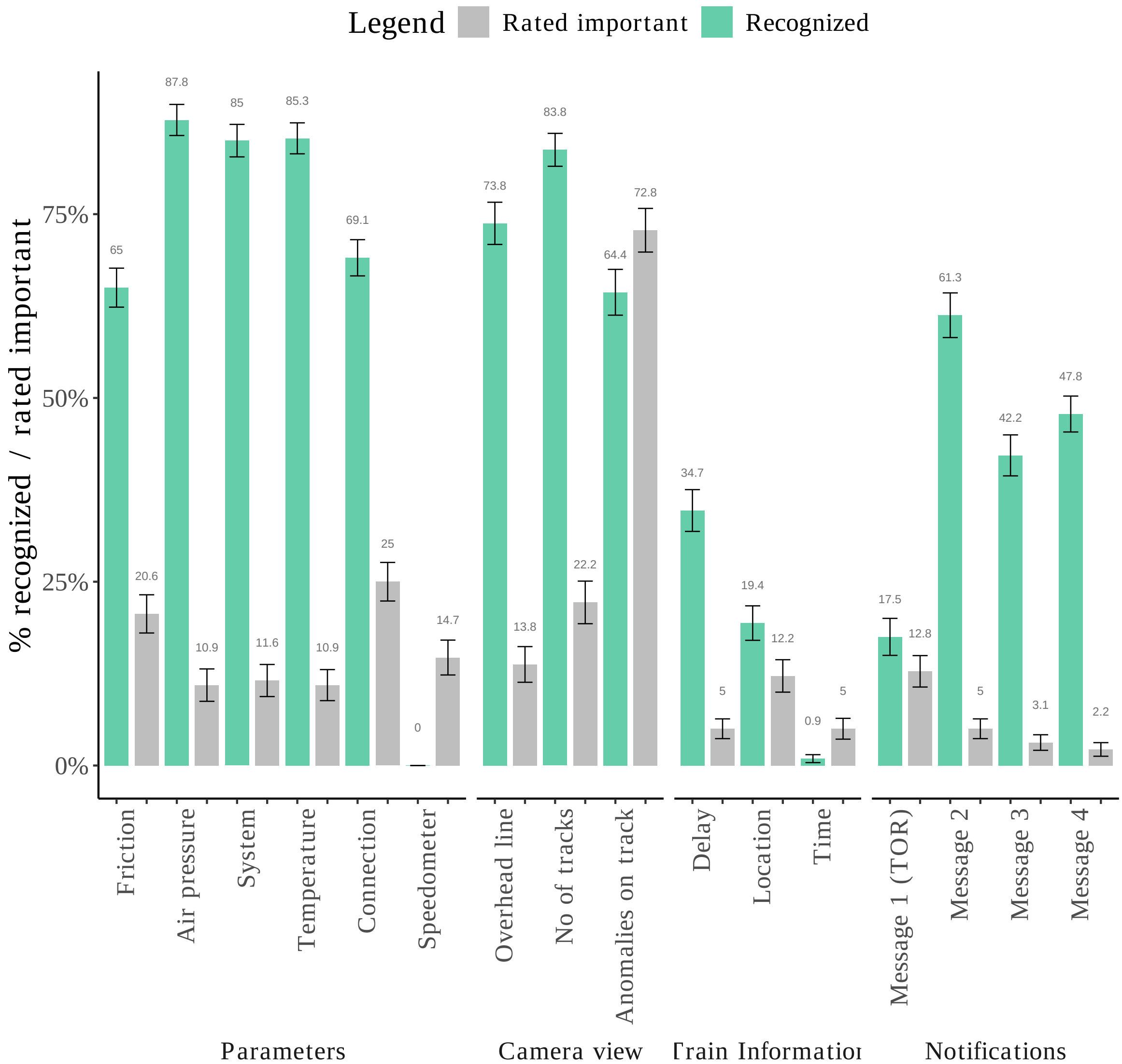
A.8 Correlation table of all variables used in the study.



	Age	Critical reasons recognised	Diligence	Experience	Fun	KSS before	KSS end	KSS middle	Mdn RT to attention check	Mean frustration	Mean SAGAT score	Mean subjective performance	Mean workload	ST difficulty	ST performance	Subjective SA	Telepresence score	Usability score
Age																		
Critical reasons recognised	−0.29																	
Diligence	0.02	0.21																
Experience	−0.01	0.04	0.06															
Fun	0.18	0.20	0.14	0.14														
KSS before	−0.25	0.01	−0.04	−0.06	−0.08													
KSS end	−0.30	−0.02	−0.03	−0.09	−0.16	0.71												
KSS middle	−0.26	0.07	−0.02	−0.04	−0.15	0.71	0.87											
Mdn RT to attention check	0.40	−0.34	−0.06	0.05	−0.02	−0.01	−0.10	−0.11										
Mean frustration	0.10	−0.44	−0.17	−0.19	−0.32	0.09	0.12	0.08	0.20									
Mean SAGAT score	−0.20	0.24	−0.02	0.17	0.18	−0.03	0.03	0.04	−0.17	−0.30								
Mean subjective performance	−0.19	0.46	0.20	0.27	0.22	−0.08	−0.04	−0.01	−0.20	−0.57	0.32							
Mean workload	0.14	−0.24	0.09	−0.09	−0.01	−0.01	0.00	−0.02	0.29	0.41	−0.16	−0.39						
ST difficulty	0.21	−0.02	0.03	0.16	0.13	−0.06	−0.13	−0.07	0.12	−0.01	0.05	0.02	0.10					
ST performance	−0.39	0.48	0.02	−0.06	0.10	−0.03	0.00	−0.02	−0.29	−0.23	0.08	0.20	−0.04	−0.29				
Subjective SA	−0.08	0.54	0.22	0.13	0.23	−0.12	−0.11	−0.03	−0.19	−0.53	0.27	0.66	−0.35	0.01	0.09			
Telepresence score	−0.01	0.06	0.04	0.14	0.23	0.08	0.02	0.04	0.06	0.08	0.09	0.05	0.21	0.13	−0.14	0.12		
Usability score	−0.07	0.18	0.08	0.15	0.28	0.09	0.09	0.13	−0.10	−0.44	0.29	0.34	−0.29	0.21	−0.15	0.32	0.14	

Note. Ellipses indicate direction and strength of association. Numbers are Pearson's *r* correlations.

A.9 Single SAGAT items' recognition rates and importance ratings



A.10 Results of the analyses of all statistical models

Results of the analysis of model 1.

	b	*95% CI*	*t*	*p*
Intercept	50.158	47.146 – 53.170	32.681	$< .001^{**}$
ST	5.461	2.848 – 8.075	4.101	$< .001^{**}$
AGTOR	-2.219	-5.912 – 1.475	-1.179	.240
Train Driver	3.505	-1.365 – 8.375	1.413	.160
Marginal R^2 = .283 *Conditional* R^2 = .512				

Results of the analysis of model 2.

	Odds Ratios	*95% CI*	*z*	*p*
Intercept	2.121	1.493 – 3.015	4.192	$<$ .001
NDRT	0.658	0.462 – 0.936	-2.329	.020
AGTOR	2.182	1.299 – 3.666	2.948	.003
Train Driver	0.906	0.416 – 1.975	-0.248	.804
NDRT $\times$ AGTOR	0.649	0.388 – 1.088	-1.641	.101
AGTOR $\times$ Train Driver	1.204	0.419 – 3.462	0.345	.730
Marginal R^2 = .046 *Conditional* R^2 = .255				

Results of the analysis of model 3.

	b	*95% CI*	*t*	*p*
Intercept	119.902	114.345 – 125.460	42.345	$<$ .001
NDRT	15.319	10.342 – 20.296	6.041	$<$ .001
Train Driver	-3.143	-16.428 – 10.142	-0.464	.643
NDRT $\times$ Train Driver	-1.003	-12.900 – 10.894	-0.165	.869
Marginal R^2 = .377 *Conditional* R^2 = .536				