UNIVERSITY OF TWENTE.

Context sensitive or rule compliant – the effect of a (semi-)autonomous system's behaviour on trust

Master thesis

University of Twente, The Netherlands

Author: Christian Peksen, s2239566

Programme: Psychology

Specialization: Human Factors and Engineering Psychology

First Supervisor: Prof. Dr. J.M.C. Schraagen

Second Supervisor: Dr. Simone Borsci

Abstract

It is well known that different fundamental trade-offs in (semi-)autonomous systems affect performance of those systems. This study aims to extend these findings by examining how balancing off fundamental trade-offs in (semi-)autonomous systems affects trust. We therefore conducted an experiment and compared two groups of participants operating a simulated (semi-)autonomous car. One group was driving a car that was strictly following traffic rules, the other group operated a car that acted upon additional contextual information and would act against traffic rules where it might appear reasonable. During the experiment, both groups repeatedly rated their confidence in the car, predictability of its actions, perceived safety and perceived effectiveness of the car. In addition, participants could regain control over their car if they wanted to overwrite a decision. We compared both groups in a between-subjects design and found that confidence and perceived safety was rated higher when the car exclusively considered traffic rules when making decisions, while perceived effectiveness was rated higher when the car considered contextual information in addition. There was no significant difference regarding predictability of the car's actions or the frequency with which participants regained control of the car. Based on additional exploratory analyses we performed, we hypothesize that whether participants followed an action proposed by the automated car or regained control of the car to perform another action was determined by a combination of the default action the car proposed and the presence or absence of risk, i.e., legal consequences for performing a certain action.

Keywords: trust in automated system, human computer interaction, trade-offs in automation, autonomous driving, explainable artificial intelligence

Table of contents

Introduction

Trust and the Role of Explainable Artificial Intelligence (XAI) in Automated Systems

With the continuous advancement of (semi-)autonomous systems, fuelled by technological advancements, questions about behavioural aspects of these systems arise, as they could play an important role when it comes to trust towards these systems. In general, a system is always autonomous up to a certain degree, meaning there are multiple levels of autonomy. For instance, the Society of Automotive Engineers (2014) defines five levels of autonomy for cars, ranging from rather simple driver assistance to full automation. Cars on the lowest level (1st level) provide single automated features, such as braking control, which can already be found in modern cars. Cars on the highest level (5th level) operate completely without the input of a human driver and are fully adaptable to various environments and situations. While even the lower levels possess great potential to make processes more efficient and effective, in comparison to level 5, they are still dependent on human-machine interactions in specific situations. A major determinant for how people interact with such systems, is the level of trust they put in the system (Tenhundfeld et al., 2019). Trust in human-machine interactions determines whether or not a (semi-)autonomous system will be accepted and relied on and to what extent users will utilize the possibilities introduced by the system. Trust is determined by how well we understand how a system works and arrives at its decisions and whether the decision rationale matches our own set of values and principles (Mayer, Davis & Schoorman, 1995). Difficulties often arise when the system acts in an unpredicted way, for instance, due to mode error, or what is aptly called 'automation surprise' (Sarter, Woods, Billings, 1997), and which translates into an observed deviation from expected system behaviour. With increasing complexity of these systems, fuelled by more and more advanced deep learning and machine learning algorithms, automated systems may become even more powerful, but at the same time turn to black boxes for the human agent. This is because in most situations it is almost impossible to grasp how algorithms arrive at their decisions and why they chose a specific option, as a result of which trust is likely to decrease rapidly (Endsley, 2017). Although one could argue that this downside would disappear if a black box's output is always as a human operator would expect it to be, this state of automation is not yet achieved.

Transforming those black-box-like systems into more transparent ones, by enhancing the requirements automated systems must meet by a call for explainability, could help to prevent such drastic shifts in trust levels that ultimately could lead to the rejection of this promising technology. Explainable artificial intelligence (XAI) takes into account the human factor when designing automated systems and aims on making them more transparent, thereby making them more predictable as well as comprehensible (Adadi & Berrada, 2018). A system designed under XAI principles will not only strive for high performance but will also explain its output, capabilities and limitations to the human

operator, e.g., via graphical illustrations and/or verbal explanations, in order to foster transparency and understandability. By doing so, the system supports human operators to build a precise mental model of what the system is capable of and how it works. However, it is important to mention that it is not enough to simply provide explanations of the system and trust the human agent to build a precise mental model. Rather, explanations must adapt to a person's understanding and take into account the fact that mental models are imprecise. For example, a very detailed and technical explanation could be perceived as being complete, but difficult to understand or even overwhelming at the same time. Instead of simply aiming for completeness, an explanation should address the question it aims to answer as well as the current state of the mental model of the human operator (see Miller, 2019). If this is the case, a loss of trust can be prevented as shown by Koo et al. (2015). In their study, they showed that explanations on why an automated car acted as it did, helped to maintain trust levels towards the system. Driverless cars are a great example of autonomous systems because most people have experience as drivers, therefore can relate to those systems and potentially already built mental models of how automated cars could function. However, these findings can also be relevant for other areas such as healthcare, legal, military, finance, and more (Adadi & Berrada, 2018).

Regardless of the domain, when systems are not designed to be transparent and XAI principles are not applied, the opacity of autonomous systems often leads to a maladjusted trust level. An important distinction here is to be made between mistrust and distrust (Lee & See, 2009). Mistrust refers to an over relying, while distrust refers to an inappropriate rejection. A famous real-life example of mistrust is drivers watching movies while driving (semi-)autonomous cars instead of monitoring the car and their environment. Distrust on the other hand can be exemplified by a strict rejection of automated features in cars, such as deactivating automated lane keeping, although using those features could improve aspects such as safety or effectiveness. To prevent these maladjustments, scientific literature gives plenty of design recommendations about how systems should be designed in order to enable the human agent to maintain common ground with the system, support coordination between the system and the human agent and foster appropriately calibrated trust in situations in which the system aids decision making (e.g., Banks & Stanton, 2015; Casner & Hutchins, 2019; Endsley, 2017; Körber, Baseler, Bengler, 2018) which include the XAI approach as well as behavioural aspects of the system, such as not initiating overtaking manoeuvres without manual interaction by the driver, in order to keep the driver in the loop.

Why would You trust Me?

To further understand the dynamics of trust in a human-machine relationship, it is helpful to specify the characteristics of trust in this context. In general, trust can be defined as "willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will

perform a particular action important to the trustor, irrespective of the ability to monitor or control that party" (Mayer et al., 1995, p. 712). In line with this, James (2002) emphasizes that trust is always linked to a certain level of risk, as a trustor is assuming that the trustee will act as expected but cannot be sure. As a result, trust is partly a function of the discrepancy between an expected behaviour and an observed behaviour, hence, predictability of behaviour is affecting trust levels.

Although this accounts for trust in a human-machine relationship as well as for trust in a human-human relationship, there is also a noteworthy difference, namely the nature of the trustee. In a human-machine relationship the trustee is a technical device, not a human being. It cannot feel or perceive trust, therefore the trust relationship described by James is one directional for human-machine interactions (Lippert, 2002, as cited in Lippert & Swiercz, 2005). Consequently, the system does not consciously choose to collaborate with the human or show a specific behaviour depending on how much it trusts the human. This one directional trust relationship means that human operators do not have to consider if the system is willing to perform a specific action or if the trust relationship is solid enough that one could expect a specific action, but rather is trust towards a system affected by the belief that the system is able to perform a certain task, referred to as a system's functionality, and the belief that the system will always act to the user's advantage, referred to as a system's helpfulness (Thatcher, McKnight, Baker, Arsal, & Roberts, 2011). It is noteworthy that these aspects do not exclusively play an important role in human-machine interactions. Mayer et al. (1995) described three similar trust dimensions identified in the context of organizational trust, namely integrity, ability and benevolence, which are widely accepted as trust determinants within the scientific community.

Lee and Moray (1992) embedded these concepts into their own trust model in which performance, process and purpose are seen as main factors influencing trust towards a system. In their view, performance is constituted by different aspects, such as the beliefs about the abilities of the system, about how reliable it is and how predictable its actions are. Performance mainly answers the "what-" questions a user could ask a system, that is, "what can the system do?". Process refers to the appropriateness of a system's actions in a specific environment or given a specific task. Closely linked to this, purpose refers to whether the system is used as intended by the designer/developer. By taking process and purpose into account, Lee and Moray (1992) pay respect to situational aspects of human-machine interactions and how they can affect trust. While one can trust a system to act as expected in one scenario, this trust level can be tremendously lower in a scenario with different characteristics.

Siau and Wang (2018) point out that not only characteristics of the situation and the system affect trust but also the trustor's disposition to trust, that is, his or her willingness to trust a system. Interestingly, Dzindolet, Peterson, Pomranky, Pierce, and Beck (2003) showed that the positivity bias, which refers to the tendency to assign positive attributes to unknown people when there is only little information about them available, also applies to automated systems. When asked how well an

automated system would perform in a detection task, participants expected it to perform better than themselves, although they had been told that the system is not free of err. At the same time Dzindolet et al. (2003) showed that the expectancy that the system would make only a few to zero errors – resulting from the positivity bias – led to a rapid decrease of trust when the system erred, ultimately leading to distrust and rejection of the system. However, this effect could be diminished by not only notifying participants beforehand that the system was prone to errors on some occasions but also providing an additional explanation why these errors could occur. In fact, people were more likely to rely on the system's decisions when this kind of explanation was provided, irrespective of whether or not the system performed better than themselves. This hints to mistrust, that is, a strong overreliance on the system, and is somewhat contradictory to assumptions that more transparency and predictability will enable operators to appropriately calibrate their trust (Adadi & Berrada, 2018). This could partly be because of a lower discrepancy between expected and observed behaviour. If I know why the system could err from time to time and expect it to do so, I might not be as surprised if it does err for this reason. The positivity bias seems to be corrected here in terms of expected reliability, but not in terms of trust. Hence, when evaluating whether to rely on a system's aid, trust appears to be the more important predictor, outweighing reliability. Although this appears to be rather non-rational and counterintuitive, this is also supported by the findings of Lee and Moray (1992), saying that the more precise the mental model of a system, the more people tend to trust it, even if this mental model includes information about possible flaws of the system. However, it is highly probable that there is a sweet spot for this balancing between trust factors. For example, a system not reliable at all, communicating all its flaws transparently, might still evoke feelings of trust, but is not relied on.

The Effect of Balancing fundamental Trade-Offs on Trust

While there is extensive literature and research on which factors influence trust in a human-machine-setting as well as on guidelines for designers on how to take these factors into account in automated systems, there is a knowledge gap regarding how systems should act in day-to-day situations in which – seen from the perspective of the human operator – more than one decision rationale is valid under given trust premises. To exemplify this, imagine an autonomous car driving down a highway with a speed limit of 100 km/h. The driver has an important meeting and will face serious consequences if he does not make it to the meeting in time. Therefore, he informed the car about his desired time of arrival. However, there was an unexpected traffic jam on the last few kilometers and with the current speed of 100 km/h he will not make it in time. Consequently, the car accelerates to 120 km/h in order to catch up for the lost time. While this is what a lot of people would do, this could be contradictory to what we would expect from an automated system, which one might expect to follow traffic rules. Nevertheless, it is, to some extent, helpful behaviour, as the system acts

in the driver's favor. Consequently, it should benefit trust levels (Thatcher, McKnight, Baker, Arsal, & Roberts, 2011). The situation is characterized by ambiguity which could make seem both behaviour patterns appropriate for the human operator and therefore theoretically trustworthy, but under different prerequisites.

At this point it must be mentioned that the Ethics Guidelines for Trustworthy Artificial Intelligence (European Commission, 2019) explicitly require that automated systems should prevent any possible harm. One could argue that the above-mentioned example deviates from this requirement and we do not want to argue against this or suggest that designers should deviate from those guidelines. However, although ethical guidelines provide a common ground for designers and therefore are significantly important for the future of those system, in our opinion it is interesting and relevant to know if human operators make the same requirements when it comes to trust. Although the car in our example takes a certain risk, it still could be perceived as trustworthy and effective, despite showing less safe behaviour. As automated systems should take into account an operator's mental model when cooperating with them, for designers it is important to know how different behaviours affect the operator's perception of the system.

A possible downside of a more context sensitive and human-like behaviour as shown in our example could be its unpredictability. For most drivers, in the above-mentioned example it would probably be easy to evaluate why the car is driving 100 km/h instead of going faster, if they pay attention to the speed limit signs. However, if the car is suddenly accelerating and exceeding the speed limit, this would probably be rather unexpected and could lead to a loss of trust.

While the strictly law-compliant behaviour might be easier to predict, it could lack perceived effectiveness, as it does not follow common-sense principles drivers usually show in day-to-day driving. An example here would be a (semi-)autonomous car approaching an intersection, with another car approaching the intersection from a lane with right of way. It is not unusual that the driver having the right of way waives it, for instance, because the limited space of the road would require a complicated manoeuvre to pass the other car. In this scenario, a strictly law compliant behaviour could also lead to a decrease of trust towards the system, induced by perceived ineffectiveness. To resolve the situation the driver must either regain control or accept the ineffectiveness with its consequence of a longer travelling duration. On the other hand, Portouli et al. (2006) conducted a survey to find out why car owners opted for a vehicle with an integrated cruise control system and found out that the two most important reasons were increased comfort and less worrying about getting police checks (which one could also interpret as an aspect of comfort). This could hint towards an acceptance of a possible ineffectiveness in a trade-off for increased comfort, although the question who is accountable for a (semi-) autonomous car's actions is not included in these considerations and might play an important role (e.g., if car manufacturers are accountable for actions of (semi-)autonomous cars, drivers could

potentially care less about whether or not their car is complying to traffic rules). What must be acknowledged here is that perceived effectiveness is highly dependent on how an individual defines effectiveness. Some drivers might include less worrying about police checks in their concept of effectiveness because they only consider a behaviour as effective if they do not have to fear legal consequences, while other drivers might separate those to concepts from each other.

As Alderson and Doyle (2010) stated, there are always trade-offs to be considered if a new technological feature is implemented into an automated system and optimization with respect to one demand could result in brittleness regarding another demand. In the examples given above, an optimization regarding predictability could result in less perceived effectiveness and vice versa. Another trade-off could be between those two factors and perceived safety. Obviously, it can be seen as less safe to exceed speed limits, however, there are also situations in which strictly conforming behaviour could result in negative consequences. Those situations often arise when another party shows non-conforming behaviour, such as speeding towards an intersection. In case of a green light switching to a yellow light it can be crucial for the vehicle in front to accelerate and risk crossing a red light instead of braking hard, as the braking distance of the following speeding car could be too long, potentially leading to an accident.

The trade-off in all the above-mentioned situations is best described as the optimality-resilience trade-off (Hoffmann & Woods, 2011), recounting the fact that “a work system can never match its environment completely; there are always gaps in fitness, and fitness itself is a moving target” (p. 68), which refers to the constantly changing and unforeseeable demands a real-life environment puts on the system. The more adaptive a system is designed, the less effective it will be in routine situations, while conversely, the more the system is suitable to perform well in routine situations, the less effective it will be in non-routine, surprising situations. This dynamic is also referred to as “robust, yet fragile” (Carlson and Doyle, 2002, p. 2539). The term describes a system that is “robust to what is common or anticipated but potentially fragile to what is rare or unanticipated” (p.2539). Hoffman and Woods (2011) describe this as a problem of bounded ecology. In routine situations a system strictly following traffic rules could potentially perform very well, however, it could have difficulties to adapt to situations where such behaviour collides with surprising and unforeseen events. On the other hand, a highly adaptive system making decisions by using the situational context only (e.g., using sensory information) could have difficulties to follow strictly defined routines and neglect sensory information where it is not needed to make decisions. A simple example here would be a red light on an empty intersection. A highly adaptive system programmed to be as effective as possible in terms of travel time could cross this red light and thereby risk legal consequences.

Nathanael, Tsagkas, and Marmaras (2017) shed more light on these kinds of trade-offs and point out that deviations of the behaviour shown in actual real-life experiences from defined and

compulsory processes occur multiple times a day in areas such as workplace environments. In theory, these deviations should result in risks and negatively affect safety, however, as situations reach a certain level of complexity, they are impossible to foresee and therefore prescribed behaviours are not suitable anymore. Rather, operators must adapt to the situation and make decisions based on the situational context they find themselves in. Nathanael et al. stress that even in highly formalized domains, operators are exposed to ambiguous situations with conflicting goals in which following standard procedures is just not applicable anymore. If we transfer this to the domain of automated systems, it appears to be necessary to enable systems to process situational context "as the humans involved would understand it" (Carpenter & Zachary, 2017, p. 1). In order to allow humans to understand the system and make appropriate judgements on factors such as effectiveness or safety, Carpenter and Zachary (2017) emphasize that reasoning about context would allow systems to adapt their behaviour appropriately. Note that this appears to be an approach which could also positively affect trust but has not been examined in the context of (semi-)autonomous systems yet. A system balancing trade-offs and prioritizing goals should always do this in a manner enabling the human to understand what was prioritized, why this was prioritized and what behaviour results from this. If we exemplify this in the context of (semi-)autonomous driving, a system should not only prioritize effectiveness and safety and then balance out these two factors. The aim is not going as fast as possible as safe as possible while ignoring all traffic rules, or in contrast bluntly follow traffic rules. Rather a system should apply a rationale, which shows acceptable and reasonable deviations from prescribed rules in situations in which they seem appropriate.

The previous section has shown that in situations with conflicting goals there is a trade-off between predictability, perceived effectiveness and perceived safety, which is associated with trust levels shown towards a system. Those factors are very likely influenced by which goals the system is prioritizing in such situations and which rationale it is following to pursue these goals. In addition, another significant aspect to increase trust is to make these decisions and the underlying rationale comprehensible by giving context specific explanations that follow human(-like) reasoning. In the next chapter we want to introduce contrastive explanations, which aim to match human reasoning and therefore could be well suited for making (semi-)autonomous systems more transparent and comprehensible.

Contrastive Explanations – Answering what Humans ask

Miller (2019) points out that when trying to comprehend why something happened, humans do not simply ask what caused a specific event but rather try to understand why this specific event happened instead of another, therefore, the relevant question to answer is "Why this outcome rather than that outcome?". An explanation answering this question is referred to as contrastive explanation

and highlights the differences between two options instead of giving complete explanations, including all information. As Miller states, real-life choices are obviously not always binary, but when questioning a decision or an event people mostly process them as if they were. This stems from the fact that in such situations a specific outcome was expected but a different outcome is observed. Hence, contrastive explanations are well suited in the context of XAI, because they are not only efficient in terms of information quantity but also have the potential to be highly effective, as they are well suited to how the human mind questions decisions. For example, driving an autonomous car we might expect that our car will strictly comply with the law, and if it does not, we want an explanation that informs us why it did so. "I increased our travelling speed to 120 km/h, because if I would comply with the speed limit of 100 km/h we would not reach our destination in time", would be an example. In contrast "I increased our travelling speed to 120 km/h in order to reach our destination in time" does not provide information about why the other option, namely complying with the law, was not chosen. The latter non-contrastive explanation leaves it unclear whether the speed limit was known and considered by the system - it is lacking completeness and missing important information, which is needed to build and maintain precise mental models (Kulesza et al., 2013).

Measuring Trust

After having discussed the role of trust in human-machine interactions, how trust is defined, which factors affect trust, how they interact with each other and how XAI can help those systems realize their full potential, the question remains how to measure trust and the associated factors.

Given the example of (semi-)autonomous cars, where drivers can turn off specific functions of the automated system or take over control completely, trust can be measured in two different ways. The first is through a questionnaire, assessing self-reported trust levels regarding specific aspects of the system. The second way to measure trust is through an observation of how the driver interacts with the system, that is, does the driver deactivate the automation in order to regain control over the car in a critical situation. As pointed out by Mayer et al. (1995), trust is the "willingness […] to be vulnerable to the actions of another party […] irrespective of the ability to monitor or control that party", hence, regaining control can be interpreted as a significant loss of trust.

There are numerous questionnaires that intend to assess trust (e.g., Adams et al., 2003; Cahour & Forzy, 2009; Jian et al., 2000; Madsen & Gregor, 2000; Merrit, 2011). One of the most basic scales was developed by Cahour and Forzy (2009) and assesses trust as well as three trust factors: safety, predictability and efficiency. It consists of four items:

1. Do you have the feeling of trust in the system?
2. Are the reactions of the system predictable?
3. Do you think that the system is safe?

4. Is the system efficient?

Although the Cahour-Forzy-Scale does not cover as many aspects of a system or factors of trust as for example the Madsen-Gregor-Scale (Madsen & Gregor, 2000), its simplicity suits research questions in which basic principles of automated systems are examined in isolation from other factors rather than during actual hands-on experiences with such systems, for which more complex interdependencies of various factors can be assumed. Furthermore, its conciseness allows time- and effort-efficient repeated measurements during the use of a system. As Hoffman et al. (2018) point out, trust should be seen as dynamic and affected by experience, consequently, trust measurements should not be done once but rather repeatedly.

From now on, to make it easier to differentiate between trust as a psychological concept and trust as measured by only the first item of the Cahour-Forzy-Scale, directly asking for trust in the system, the first one will be referred to with “trust” and the latter with “confidence”.

Research Goals

In our experiment we want to examine how a system balancing off effectiveness, predictability, and safety in a context sensitive manner compares to a system not balancing off these factors but strictly following prescribed rules. Specifically, we are interested in how these two conditions will affect trust.

In the first condition, the system will strictly comply with rules, irrespective of external influences. In the second condition, the system will show a more context sensitive behaviour, including contextual information and deviating from prescribed rules in situations where it might seem appropriate to do so. From here on, we will use RC (rule compliant) and CS (context sensitive) as labels for these two conditions.

To examine this question, we chose to focus our research on the domain of (semi-) autonomous cars. As a great share of persons are in possession of a driving license, this can be used as criterion that allows us to ensure that all participants have a basic knowledge about the rule system in which they are operating, as well as having experience with situations in which deviations from prescribed behaviour might be more appropriate, i.e., more safe or more effective, than strictly following rules.

We hope to be able to derive conclusions on what exact behaviour is deemed to be trustworthy, predictable, safe and effective, as measured by an adapted version of Cahour-Forzy scale, in day-to-day scenarios marked by ambiguity.

As Nathanael et al. (2017) pointed out, in a complex environment it is impossible to foresee every possible event. Therefore, prescribed behaviours are not suitable anymore and operators must adapt to the environment by taking into account contextual information in order to handle a specific

situation effectively and safely. As road traffic can indeed be described as complex and partly unforeseeable, we hypothesize that perceived safety, perceived effectiveness and overall confidence in the system will be higher when the system is context sensitive, therefore:

H1: $Confidence_{RC} < Confidence_{CS}$

H2: $Perceived\ effectiveness_{RC} < Perceived\ effectiveness_{CS}$

H3: $Perceived\ safety_{RC} < Perceived\ safety_{CS}$

We further hypothesize that confidence within both conditions will increase over time, as participants will be able to gather experience, thereby building increasingly precise mental models, therefore:

H1a: $Confidence_{RC,\ initial} < Confidence_{RC,\ terminal}$

H1b: $Confidence_{CS,\ initial} < Confidence_{CS,\ terminal}$

We also hypothesize that perceived predictability will be higher in the strictly compliant condition, as traffic rules should provide a valid framework for unambiguous prediction of behaviour:

H4: $Predictability_{RC} > Predictability_{CS}$

Method

Participants

Participants were invited to take part in the experiment via distribution of an online link leading to the experiment we designed on Qualtrics. The link was distributed via social media as well as direct contact with the participants by the author and a kind request to share the link with friends, colleagues and family members (snowballing method). After exclusion of 115 participants either without a driving license or showing a suspiciously low duration for the experiment, the number of participants in our sample amounted to *N* = 75, consisting of 52 males, 21 females and 2 others with an average age of 27 (*MIN* = 20, *MAX* = 59, *SD* = 7.7).

Participants were assigned to one of two groups, by distributing two different weblinks (each leading to one of the two conditions). 27 participants (18 males, 9 females) with an average age of 26 (*MIN* = 20, *MAX* = 40, *SD* = 4.8) were assigned to the RC condition while 48 participants (34 males, 12 females, 2 others) with an average of 27 (*MIN* = 20, *MAX* = 59, *SD* = 9.0) were assigned to the CS condition. The difference in size evolves from the fact that participants were not manually assigned to one condition but rather automatically and randomly assigned to one condition, determined by which link participants found on social media or were provided by the author, family members, or friends. Although we distributed the same number of links for both groups, the distribution channel of the link

leading to the CS condition became way more effective than the link leading to the RC condition, resulting an uneven distribution of participants.

This research was approved by the BMS Ethics Committee of the University of Twente, The Netherlands.

Materials

Participants were shown different scenarios on digital slides, as shown in Figure 1.

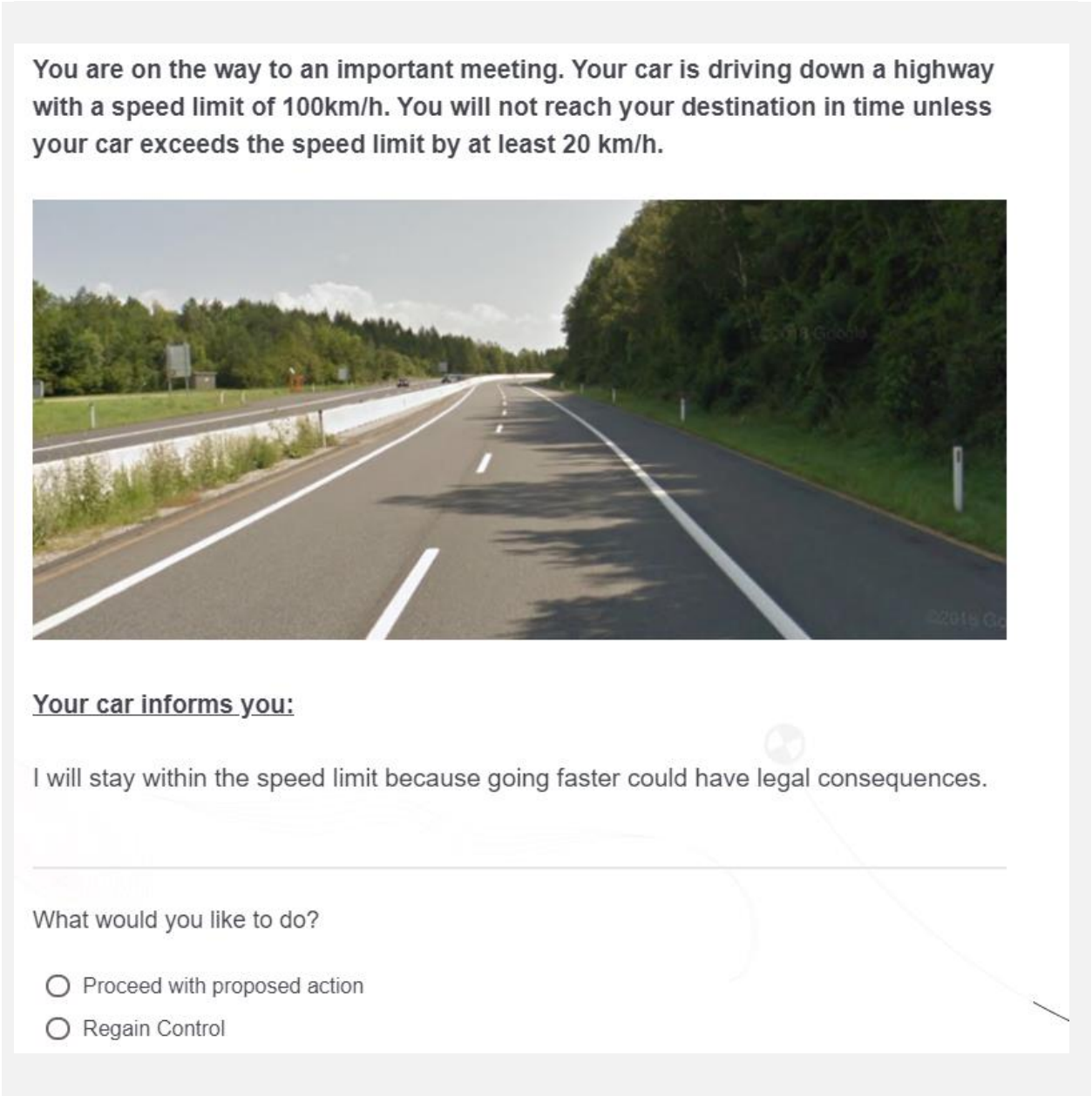


Figure 1. Materials used in the experiment.

They also had to fill in a questionnaire used to assess trust levels which was an adapted version of the Cahour-Forzy-Scale, taking into account considerations proposed by Hoffman et al. (2018) and can be seen in Figure 2 (see Appendix A for more information about adaptations we made):

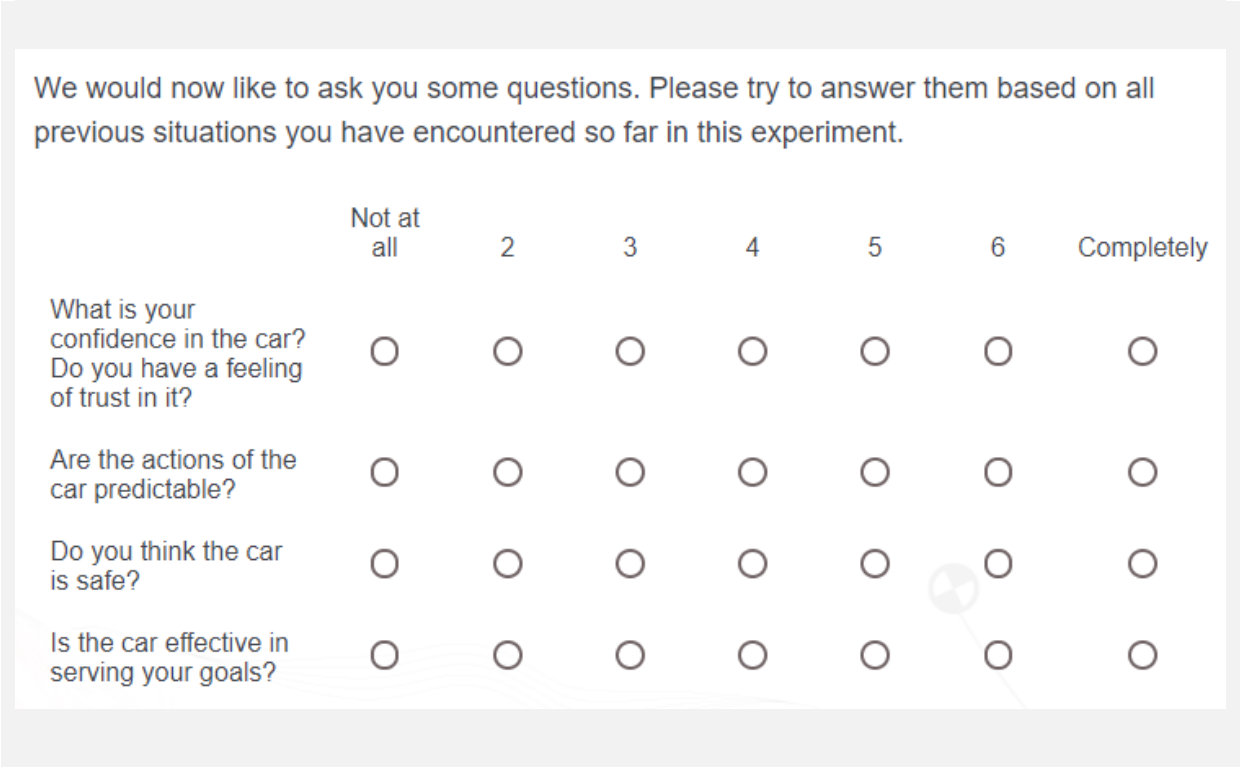


Figure 2. Questionnaire used in the experiment.

Design

In a between-subjects design, each group was exposed to one of two conditions and then compared with each other. Participants in the RC condition were shown scenarios in which a semi-autonomous car strictly complies with traffic rules while participants in the CS condition were shown the exact same scenarios, but with a semi-autonomous car showing a more context sensitive behaviour, not fully complying to traffic rules in some scenarios.

The behaviour of the car, i.e., the condition a participant was assigned to, was treated as independent variable. The dependent variables of interest were number of control-regains by the participant (as objective measure for confidence), self-reported confidence in the system, perceived predictability of the system, perceived safety of the system and perceived effectiveness of the system.

In addition, we gathered qualitative data in cases in which participants decided to regain control over the system, in order to explore what led them to this decision.

Procedure

As trust is partly a function of experience and preciseness of one's mental model, participants were shown a series 32 scenarios to have sufficient time to gather experience and build a mental model. Eleven of those 32 scenarios incorporated critical situations with ambiguity (see Appendix B for descriptions of all critical scenarios). The remaining 21 scenarios incorporated non-critical situations in which the car's behaviour did not differ between conditions, such as stopping at a red light. The scenarios appeared in randomized order.

Before starting, participants received a short briefing about the system, explaining its capabilities. However, they were not made aware of the fact that the car was either strictly following traffic rules or not always following traffic rules, as this could provoke effects of social desirability. They were made aware that they could regain control of the car in every scenario. Participants were then asked to fill out the Cahour-Forzy-Scale, in order to assess their initial trust, prior to actually experiencing the car in concrete scenarios. After every third scenario they were asked to fill out the Cahour-Forzy-Scale again, ultimately leading to 12 times of measurement.

During the experiment, as shown in Figure 1, participants were shown traffic situations via slides and received a written description of the situation, including their goals as a driver when this was useful, e.g.:

"You are on the way to an important meeting. Your car is going down a highway with a speed limit of 100km/h. You will not reach your destination in time unless your car exceeds the speed limit by at least 20 km/h."

They then received a message from the system, explaining its intended behaviour in a contrastive manner, e.g.:

"In order to reach your destination in time, I will accelerate to 120 km/h because sticking to the speed limit would mean you will not reach your destination in time."

While the above-mentioned explanation refers to a more context sensitive behaviour shown in the CS condition, explanations in the RC condition read as follows:

"I will stay within the speed limit because going faster could have legal consequences."

After every explanation, participants were asked if they wanted to proceed with the proposed action or wanted to regain control over the car. When participants decided to regain control, they were asked what action they intended to perform instead of following the action the car proposed. This also applied to the 21 scenarios not marked by ambiguity. The median duration of the experiment was 23 minutes. After the experiment, all participants were provided with contact information of the author in case of open questions.

Data analyses

Control measurements

To ensure that there were no initial differences caused by specifics of the respective group, we performed independent samples t-tests comparing how participants of both groups rated confidence, predictability, perceived safety and perceived effectiveness of semi-autonomous systems prior to the actual experiment, based on their previous experiences and expectations.

Hypothesis Testing

Confidence. To evaluate our first hypothesis, we accumulated confidence ratings of all but the initial questionnaire for each group separately and performed an independent samples t-test in order

to compare mean values averaged across time and participants. We excluded the initial questionnaire for our between-subjects analyses because we wanted to make sure to not underestimate the differences by including data unaffected by the independent variable, thereby making group means look more similar than they are.

As a second and objective measurement for confidence, we examined whether there was a relationship between the behaviour of the car and the frequency with which participants regained control. We therefore selected all eleven critical scenarios and counted how often participants chose to regain control and how often they chose to proceed with the action proposed by the car for each condition separately. We then performed a Chi-Square test in order to examine if the condition had an effect on how often participants chose to regain control.

To examine whether confidence increased over time, as per hypotheses 1a and 1b, we performed a one-way ANOVA with repeated measures and performed tests of within-subjects effects (with time of measurement as within subject-factor) for each condition.

Perceived effectiveness, perceived safety and predictability. In order to examine differences between both conditions regarding perceived effectiveness, perceived safety and predictability, we accumulated ratings of all questionnaires (except for the initial questionnaire) for both groups and each of the three factors and performed independent samples t-tests in order to compare group means for each factor averaged across time and participants.

Exploratory Analyses

In order to explore possible explanations for our results, we also performed explorative analyses on our data.

First, we started by having a look at how ratings of confidence, perceived effectiveness, perceived safety, predictability as well as regains of control correlate with each other.

Second, we analyzed the qualitative data obtained from requesting participants to explain which action they would perform after regaining control. We therefore examined all answers given to this question for all eleven critical scenarios and clustered them in three categories:

1. **not interpretable**: all unclear comments went into this category (e. g., “the car is a rebel”, “come on…”)
2. **what the other car would do**: all answers that suggested that the participant would perform the same action as the car would perform in the opposite condition (i.e., all answers of participants in the RC condition which describe the corresponding action of the car in the CS condition and vice versa)
3. **something else**: all answers that suggested that the participant would perform an action that neither resembled the action performed by the car in the RC nor in the CS condition in the corresponding situation

In the following we excluded all non-interpretable answers for our further analyses.

Third, to further examine the importance of the presence or absence of risks of facing legal consequences as a critical aspect of a situation for trust, we grouped all eleven critical scenarios into two categories: those scenarios where traffic rules were a critical aspect of the situation (e.g., scenarios in which the system could either stop or not stop at a stop sign, go faster than the speed limit or stay within the speed limit, etc.), and those where traffic rules played a minor role (e.g., allowing or not allowing a pedestrian to cross the street although there was no crosswalk). We then counted how often participants chose to regain control in those scenarios where traffic rules were obviously important (i.e., where a non-compliant behavior could result in legal consequences) compared to those where traffic rules played a minor role, for the CS and RC condition individually. Lastly, we performed a Chi-Square test for each condition separately to examine if the presence or absence of risks of facing legal consequences had an effect on how often participants chose to regain control.

Results

Control Measurements

Descriptive statistics obtained from the initial questionnaire for both groups are displayed in Table 1. Table 2 shows that there were no significant initial differences between participants in the RC and the CS condition regarding confidence, predictability, perceived safety or effectiveness.

Table 1

Initial means for confidence, predictability, perceived safety and perceived effectiveness.

Factor	Group	*N*	*M*	*SD*
Confidence	RC	27	4.19	1.44
	CS	48	4.08	1.56
Predictability	RC	27	4.56	1.89
	CS	48	4.42	1.25
Perceived safety	RC	27	4.48	1.42
	CS	48	4.40	1.38
Perceived	RC	27	4.52	1.55
effectiveness	CS	48	4.46	1.35

Table 2

Mean baseline differences between both conditions for confidence, predictability, perceived safety and perceived effectiveness.

Factor	*t*	*Df*	*p*	*d*	Mean difference	95% CI of the mean difference	
						Lower	Upper
Confidence	.28	73	.781	.07	.11	-.63	.83
Predictability	.47	73	.640	.11	.14	-.45	.72
Perceived safety	.26	73	.799	.06	.08	-.58	.76
Perceived effectiveness	.18	73	.861	.04	.06	-.62	.74

Hypothesis testing

Confidence

Our results suggest a significant difference between the RC condition (*M* = 4.77, *SD* = 1.27) and CS condition (*M* = 4.46, *SD* = 1.63) with a mean difference of .31 (*95% CI* = .12, .52); *t*(742) = 3.2, *p* = .001, *d* = .20, meaning that confidence was higher in the RC condition.

We found no significant relation between the behaviour of the car, i.e., the condition in which participants found themselves, and the frequency with which participants regained control, X^2 (1, *N* = 823) = 3.14, p = .077, φ = .06. Participants in the RC condition chose to regain control of the car in 35% of all cases, while participants in the CS condition chose to regain control in 42% of all cases.

Consequently, contrary to our expectations (H1), confidence ratings were higher in the RC condition as compared with the CS condition and there was no significant difference regarding how often participants regained control.

Confidence over time changed with a maximum difference of .74 for the RC condition and .53 for the CS condition. The means obtained from the confidence scales of each questionnaire/measurement point are shown in Figure 3. For the RC condition, Mauchly's test indicated that the assumption of sphericity was violated, $X^2(65) = 170.82$, $p < .001$, consequently the degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity. The results for the RC condition indicate that there was a significant effect of time - meaning opportunities of participants to gather experience with the car - on confidence, $F(4.88, 126.87) = 3.19$, $p = .01$, $d = .11$. Tests of within-subjects contrasts showed that this effect stems from a linear growth of trust over time in the RC condition, $F(1) = 7.45$, $p = .01$, $d = .22$. For the CS condition, Mauchly's test also indicated that the assumption of sphericity was violated, $X^2(65) = 172.64$, $p < .001$, consequently the degrees of

freedom were corrected using Greenhouse-Geisser estimates of sphericity as well. In contrast to the RC condition, the results indicate that there was no significant effect of time - meaning opportunities of participants to gather experience with the car - on confidence observed in the CS condition, $F(6.07, 267.00) = 1.20$, $p < .287$, $d = .026$.

Consequently, H1b must be rejected while H1a is accepted.

Figure 3. Average trust levels across all twelve times of measurement.

Perceived effectiveness, perceived safety and predictability

Perceived effectiveness. The test results show that there is a significant difference between the RC condition ($M = 4.59$, $SD = 1.36$) and the CS condition ($M = 4.77$, $SD = 1.63$) with a mean difference of -.18 (*95% CI* = -.39, .03); $t(812) = -1.61$, $p = .046$, $d = .11$. Consequently, H2 is accepted. Effectiveness was perceived higher in the CS condition as compared to the RC condition.

Perceived Safety. The test results show that there is a significant difference between the RC condition ($M = 4.74$, $SD = 1.25$) and the CS condition ($M = 4.57$, $SD = 1.60$) with a mean difference of .175 (*95% CI* = -.02, .37); $t(738) = 1.74$, $p = .041$, $d = .11$. Consequently, H3 must be rejected. Safety was perceived higher in the RC condition as compared to the CS condition.

Predictability. The test results show that there is no significant difference between the RC condition ($M = 4.96$, $SD = 1.22$) and the CS condition ($M = 4.83$, $SD = 1.52$) with a mean difference of .13 (*95% CI* = -.07, .31); $t(726) = 1.26$, $p = .11$, $d = .09$. Consequently, H4 must be rejected.

Exploratory Analyses

Correlations of confidence, predictability, perceived safety, perceived effectiveness and regains of control

As a first step, we inspected how strong the relationships between the measured factors are. Table 3 shows all correlations:

Table 3

Correlations of confidence, predictability, perceived safety, perceived effectiveness and regains of control.

Factor	Confidence	Predictability	Perceived safety	Perceived effectiveness	Regains of control
Confidence	--	.81*	.89*	.68*	-.12
Predictability	.81*	--	.82*	.60*	-.08
Perceived safety	.89*	.82*	--	.73*	-.11
Perceived effectiveness	.68*	.60*	.73*	--	-.15
Regains of control	-.12	-.08	-.11	-.15	--

Note. * indicates $p < .001$

Analyses of qualitative data

During our analyses of the qualitative data we have gathered we excluded 13% of all answers (36 in total) because they were not interpretable. Of the remaining answers across both conditions, 93% (240 in total) could be categorized as “what the other car would do”, while 7% were categorized as “something else” (19 in total). The distribution within each condition was quite similar for both conditions with 94% of all answers in the RC condition describing “what the other car would do” and 90% in the CS condition.

Analyses of frequencies of regains of control in two different types of scenarios

Figure 4 shows that participants in the RC condition decided to proceed with the action proposed by the car (namely a rule compliant action) relatively more often when traffic rules were an important aspect of the situation, i.e., where non-compliant behaviour could result in legal consequences, compared to situations where traffic rules played a minor role.

A Chi-Square test showed that there was a significant difference between those two types of scenarios for the RC condition, X^2 (1, N = 297) = 4.95, p = .026, φ = .13. This means the likelihood that participants would regain control of the car was significantly lower when traffic rules were a critical aspect of the situation.

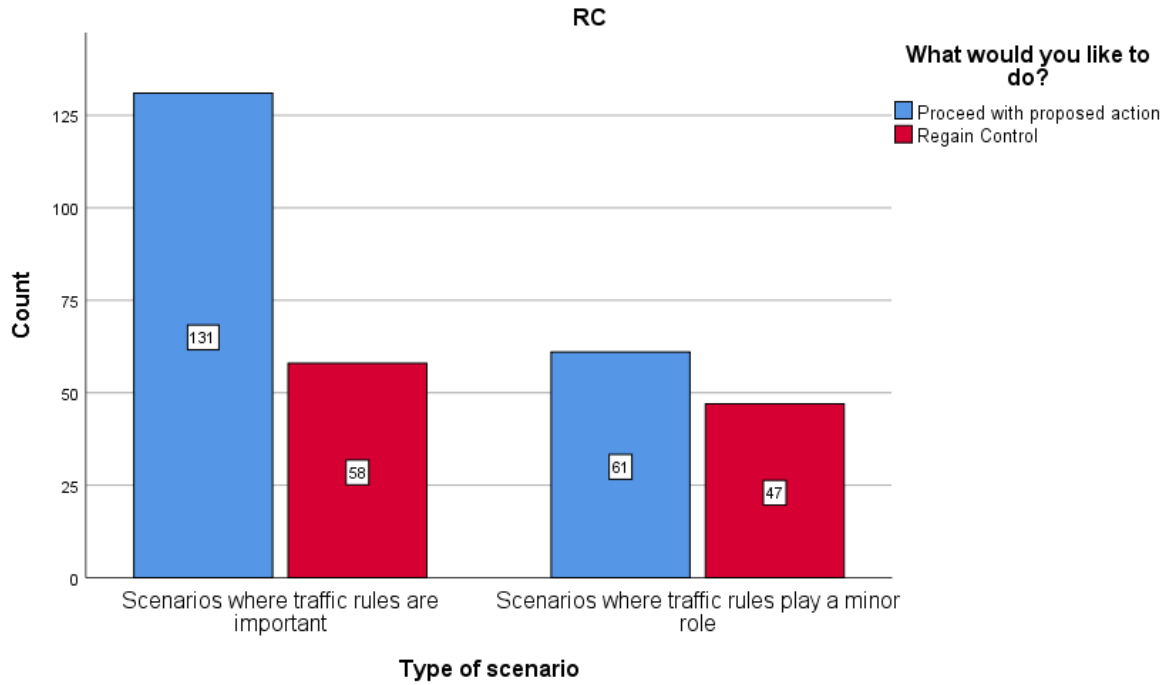


Figure 4. Count of regains of control in the RC condition for scenarios where traffic rules are important versus scenarios where traffic rules play a minor role.

For the CS condition we observed the exact opposite. Figure 5 shows that participants in the CS condition decided to proceed with the action proposed by the car (namely a non-compliant action) relatively less often when traffic rules were an important aspect of the situation compared to situations where traffic rules played a minor rule. A Chi Square test showed that there was a significant difference between the two types of scenarios for the CS condition, X^2 (1, N = 526) = 64.10, p < .001, φ = .35, meaning the likelihood that participants in the CS condition would regain control was significantly lower when traffic rules played a minor role compared to when traffic rules were an important aspect of the situation.

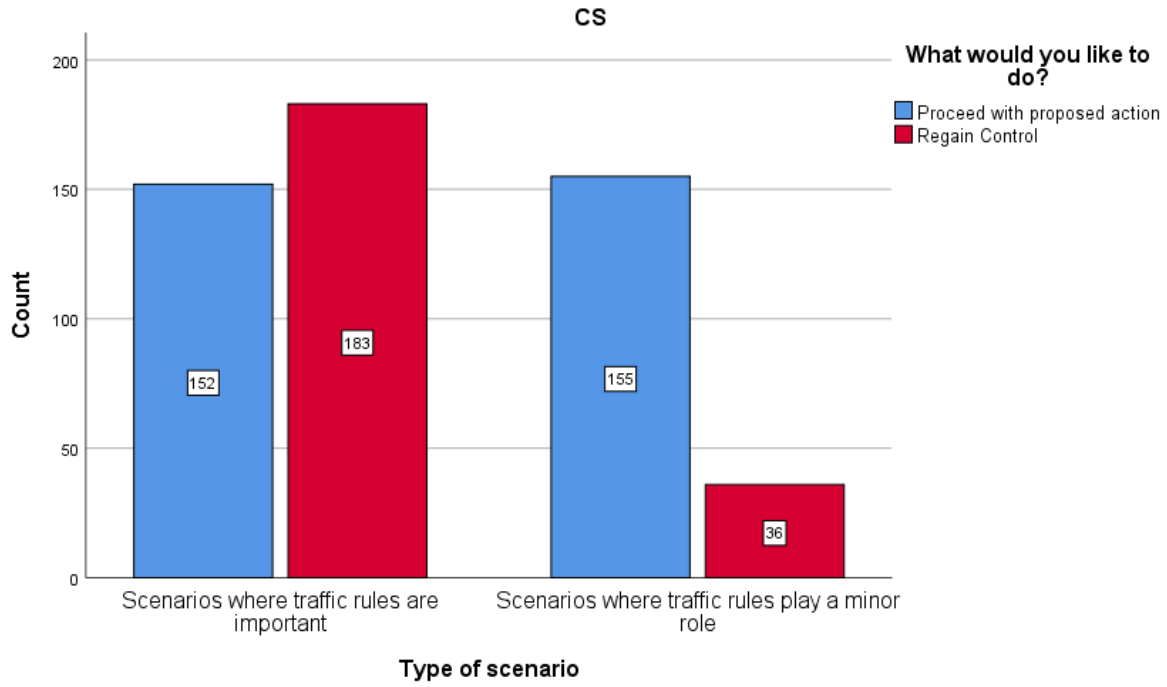


Figure 5. Count of regains of control in the CS condition for scenarios where traffic rules are important versus scenarios where traffic rules play a minor role.

Discussion

The goal of our research was to examine how different behaviour patterns of a (semi-)autonomous system affect trust, as measured by confidence, predictability, perceived effectiveness and perceived safety as well as by the frequency with which participants regained control over the system. We therefore simulated two (semi-)autonomous cars with disparate modus operandi, i.e., differently balanced unavoidable trade-offs in (semi-)automated systems. The first car (RC) strictly followed prescribed traffic rules, irrespective of additional contextual information. The second car (CS) acted upon additional contextual information in order to increase effectiveness and safety and was willing to act against traffic rules where it might appear reasonable.

Results

Confidence. While we expected to find higher confidence ratings and fewer regains of control in the CS condition, the results showed a different picture. Confidence ratings were significantly higher in the RC condition, supported by fewer regains of control compared to the CS condition, although the latter finding was not significant. If we look at how confidence levels developed over time, we can see that in the RC condition confidence increased, as we had expected for a well-designed system designed to avoid automation surprises by consistently acting according to traffic rules.

In contrast, confidence in the CS condition did not significantly increase but rather followed an almost quadratic function and levelled off after around two-thirds of the experiment. This resembles what we would expect after an automation surprise as described by Sarter, Woods, and

Billings (1997). At this point it is important to recall that participants encountered all scenarios in a fully randomized order, meaning that we cannot attribute the sudden decrease to a specific scenario or single action of the car. However, this could be explained by some sort of accumulation of "mild" automation surprises which at some point reached a certain threshold after which participants progressively lost confidence in the system. What speaks against this explanation and the occurrence of an automation surprise is the observation that predictability did not differ significantly between both conditions. In addition, our exploratory analyses showed that participants actually perceived critical scenarios as binary, just as we had expected based on the work of Miller (2019). This means they could at the same time predict how a context sensitive action as well as a rule compliant action would look like, irrespective of the condition they were in.

By its nature, an automation surprise implies that one is not expecting it, i.e., cannot predict it – ultimately, this means that there could be another reason for the sudden loss of confidence we observed.

We assume that one reason could be that participants in the RC condition did not have to worry about legal consequences of the car's actions, while participants in the CS condition did. As mentioned earlier, Portouli et al. (2006) showed that people using automation features in (semi-)autonomous cars reported that increased comfort and less worrying about getting police checks are the most important reasons for using those features. And indeed, our analyses showed that participants in our experiment often reported that they took legal consequences into account when deciding whether or not they should regain control over the car ("It's not my fault if we crash", "I would brake since running a red light is highly illegal […] if [the following car] jumps into me though, I will receive some nice insurance check!", "follow the rules", "the action […] is not allowed by law"). Based on this information, one could infer that participants in the CS condition felt they had to monitor the car more closely in order to be sure that they were not in danger of facing legal consequences, potentially resulting in less confidence compared to the RC condition (from here on all aspects associated with "less worrying" will be referred to as "comfort"). But why did confidence in the CS condition increase before falling off again?

From Dzindolet et al. (2003) we know that the positivity bias which can be observed in human-human interaction is also observable in human-machine interactions. If we do not encounter automation surprise confidence is likely to increase with experience. The turnaround of this dynamic was probably induced by an increasing disagreement with the car's actions. Participants could indeed predict what the car would do in a critical scenario, but it seems that not all participants did favour the actions performed by the car in the CS condition. This violates a basic principle of trust, namely that the trustee will always act in favour of the trustor. While the majority of situations in our experiment were neither critical nor provoking disagreement, increasing experience came with more and more

critical situations where participants could have disagreed with the car, at some point overstretching the range of tolerance participants had for such behaviour, leading to decreasing confidence. This is also in line with previous findings of Moray et al. (2000) who showed that trust towards an automated system is not only determined by the occurrence of automation surprises or its reliability but also by how often the human operator does not agree with the system's decisions.

Predictability. Although predictability of the car's actions was rated higher in the RC condition, the difference was not significant, contrary to our expectations. We find that remarkable given that traffic rules provide a solid and clear framework for precise predictions of behaviour. But given our data, a more human-like and context sensitive decision rationale is perceived almost as predictable as a strictly rule-following decision rationale. To some extent this confirms Carpenter and Zachary's (2017) claim that a human-like reasoning of a (semi-)autonomous system in combination with XAI allows the human operator to predict the system's behaviour correctly, however, at least in terms of predictability, it does not show advantages over systems acting strictly in a rule compliant manner. In fact, in our experiment we saw that context sensitive reasoning even led to disadvantages regarding confidence and perceived safety.

Effectiveness. As expected, effectiveness was perceived significantly higher in the CS condition. The fact that this was the only factor where the car in the CS condition scored significantly higher than the car in the RC condition makes it likely that human operators treated effectiveness of a system as a separate construct and were willing to trade-off effectiveness against other aspects, such as comfort or perceived safety.

Perceived safety. The car's actions in the RC condition were perceived as significantly safer than in the CS condition. This is contradictory to the assumptions we made based on the finding of Nathanael et al. (2017), who stated that in complex environments necessary deviations from prescribed rules increase safety instead of decreasing it. However, if comfort is as important as we hypothesize based on our findings, then the urge to monitor the actions of the car more closely in order to prevent legal consequences might have raised suspicion and been harmful to perceived safety in the CS condition.

In conclusion, confidence and perceived safety were higher in the RC condition, while perceived effectiveness was higher in the CS condition, meaning that balancing fundamental trade-offs in automated systems does not only affect performance of those systems but also affects trust shown towards the system. There were no significant differences regarding the frequency of how often participants regained control in critical scenarios, although they regained control less often in the RC condition. Predictability was slightly higher in the RC condition, although not significantly so. An analyses of how confidence developed over time showed an increase in both conditions. However, in

the CS condition confidence started to decrease after an initial increase, potentially induced by increasing disagreement with the car's actions.

Our results lead us to the conclusion that fear of legal consequences may be an important determinant for whether participants trust a (semi-)autonomous car. We therefore performed exploratory analyses of our data in order to generate new hypotheses, which we will discuss in the following.

Exploratory Analyses

Correlations. An examination of the correlations between confidence, predictability, perceived safety and perceived effectiveness and regains of control yielded two interesting findings.

First, as can be seen in Table 3, confidence, predictability and perceived safety are highly intercorrelated, while effectiveness shows the lowest correlations with each of those three factors. This underscores the claim that drivers could treat effectiveness as a separate construct and in fact are willing to trade-off between effectiveness and other aspects, such as safety or potentially comfort. It also provides hints as to why overall confidence was rated higher in the RC condition. While effectiveness was the only factor where the car of the CS condition performed better, perceived safety was higher in the RC condition and, in comparison, appears to have a higher association with confidence.

Second, the frequency of regains of control showed relatively small correlations with the four assessed trust factors. This led us to the conclusion that there might be another reason explaining why and when participants regained control and when they did not.

Deeper analyses of qualitative data and regains of control. To explore possible explanations, we had a look at the qualitative data. What was interesting to see here is that when participants regained control, in 93% of all cases across the critical scenarios they reported they intended to perform the exact same action as the system would show in the other condition, making a strong point for Miller's (2019) claim that humans process situations in which they have to make a decision in a binary way as well as for the clarity of ambiguity created in our scenarios, meaning that the selected situations mainly provoked no more than two different behaviours (a strictly compliant behaviour and a more context sensitive behaviour) in participants, which was in line with the corresponding behaviour our simulated systems showed. This claim is essential for our further exploration, because it allows us to infer that participants mostly perceived two possible actions simultaneously in each critical scenario – either a strictly rule compliant action or a more context sensitive action.

Based on the previously mentioned claim that drivers put high emphasis on avoiding legal consequences when operating a car, we separated all critical scenarios into two groups, those where traffic rules played an important role and those where they did not. As a result, Figure 4 and Figure 5

display the strongest argument for the claim that human drivers put high emphasis on rule compliance and avoiding legal consequences.

Recall that participants perceived all decisions as binary (they had one RC action and one CS action in mind). In scenarios where traffic rules are important, participants across both conditions tended to prefer the RC action rather than the CS action – meaning they did not want to be involved in actions that potentially could have legal consequences. In scenarios where traffic rules play a minor role, participants in the RC condition were almost indifferent about which action to perform while participants in the CS condition preferred the CS action.

Moreover, the relative difference between the number of participants regaining control and those proceeding with the proposed action in the RC condition was higher in situations where traffic rules were important compared to situations where traffic rules played a minor role, while it was the other way around in the CS condition. But why is that? Did priorities of participants in the CS condition differ from those of participants in the RC condition?

A more likely explanation for this could be that participants were affected by the so-called automation bias (Mosier & Skitka, 1996; Dzindolet, Beck & Pierce; 2006). This refers to the tendency to rely on cues provided and decisions made by automated systems in a heuristic manner. In general, a human operator who is affected by an automation bias is more likely to use and rely on an automated aid. What is interesting though is that the strength of the effect of the automation bias was not exactly consistent across our two conditions. In the RC condition, participants were more likely to overcome this bias and overrule the system's intended action when traffic rules played a minor role while it was the other way round in the CS condition – participants in the CS condition were more likely to overrule the system when traffic rules played an important role.

Dzindolet, Beck, Pierce and Dawe's (2001) work provides a framework from which an explanation for our observations can be derived. We want to highlight two processes included in this framework.

First, Dzindolet et al. (2001) state that the likelihood whether or not a human operator follows the decision of an automated system (they call it "level of automation use") is determined by the relative trust shown towards the system, which is a direct product of the perceived utility of the automated system. The perceived utility is the difference between the perceived reliability of the automated system and the perceived reliability of the human operators themselves. When the perceived reliability of the automated system is higher than that of the human operator, this will result in a positive utility of the automated system, leading to higher trust towards the system and a higher chance that the human operator will follow the decision of the system (see Figure 6).

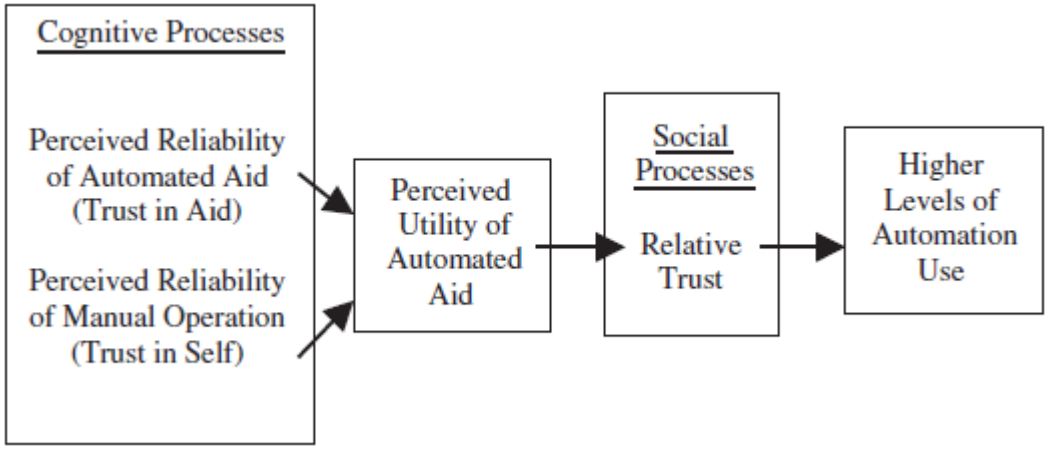


Figure 6. Relative Trust (Dzindolet et al., 2001).

However, when reliability of the human operator is perceived higher, utility of the automated system is negative, leading to higher trust in oneself and a higher chance that the human operator will overrule the decision of the system. Dzindolet et al. stress that it is not enough to increase the actual reliability of an automated system in order to increase the likelihood that a human operator follows its decisions but rather the perceived reliability of the system must be greater than that of the human operator.

In the context of our research, this helps to understand why participants were willing to trade off effectiveness against other aspects. Although effectiveness was perceived higher in the CS conditions (e.g., shorter travel times), participants were willing to trade effectiveness against overall utility, in particular with respect to a reduction of the risk of facing legal consequences. This leads us to the second process we want to highlight, which provides an explanation how participants could have evaluated the utility of the system.

In their framework of automation use, Dzindolet et al. (2001) point out that whether or not human operators rely on automation is also a function of the rewards and penalties expected and is linked to motivational processes. These processes start with human operators determining the importance of a specific outcome, before evaluating the value of this outcome (i.e., positive or negative) and the effort it takes to produce this outcome. If getting a specific reward or avoiding a specific penalty is judged important enough that a certain amount of effort is justified and requires to overrule the automated system, human operators will overrule the decision of the automated system in order to get that reward or avoid that penalty. If the outcome is not important enough and/or the effort to produce this outcome is deemed too high, human operators will follow the decision of the automated system, even though they might disagree with it. The key message here is that overruling the default decision of the automated system requires overcoming a motivational threshold, because it is not only the outcome but also the effort it takes to reach this outcome which determines our

behaviour. As a result, this increases the likelihood that the default action of the car is accepted by the human operator, as this is associated with the least effort.

In our research increased effectiveness could be seen as reward while legal consequences resemble penalties. From the results shown in Figure 4 we can deduce that for most participants in the RC condition, the outcome was either not important enough or the effort to reach this reward was too high to overrule the car's decisions in situations where traffic rules were important. However, this changed in scenarios were traffic rules played a minor role. Almost half of all participants were willing to invest the effort to overrule the system in order to get the reward of increased effectiveness. The same principle applied to participants in the CS condition, as we can see from Figure 5. Participants were more willing to invest effort when traffic rules were important than when traffic rules played a minor role. This potentially results from the fact that participants could avoid penalties by overruling the system when traffic rules were important but could actually not increase effectiveness or avoid any penalties when traffic rules played a minor role, so there was no reason to invest the effort to overrule the system.

In conclusion, the measurement of how often participants regained control of the car in each critical situation in combination with the qualitative data we have gathered, allowed us to explore which role traffic rules and associated legal consequences of breaking them play when participants chose to regain control. The results suggest that the behaviour of participants was influenced not only by the presence of risk (i.e., legal consequences) but also by the default behaviour of the car. Participants appeared to be more likely to perform an action when it was the default action of the car.

Limitations

As we know, age and gender are highly influential on driving behaviour. Especially younger male drivers are more likely to take risks and show less compliance with traffic rules (Deery, 1999). As those tendencies could affect expectancies regarding the behaviour of (semi-)autonomous cars and discrepancies between expected behaviour and observed behaviour affect trust, it would have been preferred if our sample had shown representativeness for those two variables. However, most of our participants were male and in their twenties. The high share of participants in their twenties could be an artefact of the sampling strategy. Due to the COVID-19 pandemic, participants were recruited exclusively via a weblink leading to our experiment which was shared on social media platforms such as Facebook – where most active users tend to be teenagers and young adults – or sent directly via messenger to potential participants, who were asked to further distribute the link they received. This snowballing approach is known to reduce diversity of samples (Kirchherr & Charles, 2018), because the starting point of the way of distribution mostly is a single person inviting peers who then invite further peers and so on, ultimately leading to a relatively homogeneous group of participants. To counter the

expected imbalance of age groups and gender all participants were asked to distribute the links within their families. Although this helped us to widen the age range of our sample (our oldest participant was 59) we unfortunately still ended up with a skewed distribution, not representing the actual age or gender distribution found in European countries such as Germany, Austria or the Netherlands (where participants were mainly recruited). It is quite likely that the direct invitation of participants by the author led to the circumstance that participants resemble the author's age and gender.

A further limitation stems from effects of reactivity, meaning that participants potentially showed a behaviour different to the behaviour they would show in real life, because they knew they were tested. Some participants we talked to after the experiment, expressed that they felt as if they were back in driving school and felt tested. They reported that often they regained control as soon as they noticed the car was not acting in accordance with traffic rules. This could potentially have led to biases of our results, expressing themselves in overweighing the importance of traffic rules.

In hindsight, we believe it would have been helpful to have ratings of confidence, predictability, perceived effectiveness and perceived safety for every critical scenario individually. Our exploratory analyses showed that scenarios with different characteristics provoked different behaviour in participants (i.e., regaining control or not). By further exploring these characteristics we believe it could be possible to identify and characterize situations where context sensitive decisions of the (semi-)autonomous systems would positively affect confidence, perceived safety and perceived effectiveness. Our fully randomized order with a questionnaire after every third scenario unfortunately did not allow this.

The fourth limitation is induced by the length of our experiment. As the experiment was performed online, we wanted to find the sweet spot between keeping participants motivated to finish the experiment and expose them to a sufficient number of scenarios to be able to build up trust. As we now see, against our expectations confidence levelled off after some time in the CS condition instead of showing a linear increase. We do not know if this decrease would have continued, come to a halt at a certain level, turn around and become an increase again or even surpass confidence levels in the RC condition after more time to gather experience.

Finally, it must be mentioned that the effect sizes we observed were rather small, ranging from .11 to .35.

Future Research and Implications for Designers

For (semi-)autonomous cars there appears to be a trade-off between perceived effectiveness on one hand and confidence, perceived safety and potentially comfort on the other. Risks such as legal consequences could be a major determinant for this trade-off. Future research should examine if this

trade-off is generalizable to other domains and put emphasis on preventing effects of social desirability.

We also want to encourage researchers to further examine characteristics of situations where context sensitive decisions of (semi-)autonomous systems would have stronger positives effect on confidence, perceived safety, predictability and perceived effectiveness than those of a strictly rule compliant system and vice versa. Knowing which of both rationales is perceived more appropriate for a specific situation could be helpful for designers. An adaptive system that can switch between a RC rationale and a CS rationale, depending on the characteristics of the situations, could balance trustworthiness and performance at the sweet spot.

In addition, we want to highlight three important implications for designers, based on our findings.

First, in context of (semi-)autonomous driving, there might be a trade-off between perceived effectiveness and trust, which should be considered when developing decision rationales of automated cars. The most effective car might not automatically be the most trusted car.

Second, drivers are more likely to perform a certain action when the car proposes this action as default. Our research showed that drivers were more likely to act against traffic rules when the car proposed to do so, compared to when it did not. It should always be borne in mind, that the responsibility of designers does not end with giving human operators the authority to overrule decisions of the system.

Third, different situations require different decision rationales and adaptive automated system should be able to apply the appropriate rationale. As we have learned from our research, contextual information processing could result in a decrease of trust when it is associated with legal consequences. In such situations an adaptive automated car applying a rule compliant behaviour would potentially be less effectives, but more trustworthy. However, in situations where there is no risk of legal consequences, a more context sensitive behaviour could have positive effects on trust as well as effectiveness.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access, 6,* 52138-52160.

Adams, B.D., Bruyn, L.E., & Houde, S. (2003). Trust in automated systems. Report, Ministry of National Defence, United Kingdom.

Alderson, D. L., & Doyle, J. C. (2010). Contrasting views of complexity and their implications for network-centric infrastructures. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and humans, 40*(4), 839-852.

Banks, V. A., & Stanton, N. A. (2015). Keep the driver in control: Automating automobiles of the future. *Applied Ergonomics, 53*, 389-395.

Cahour, B., & Forzy, J. F. (2009). Does projection into use improve trust and exploration? An example with a cruise control system. *Safety Science, 47*(9), 1260-1270.

Carlson, J. M., & Doyle, J. (2002). Complexity and robustness. *Proceedings of the National Academy of Sciences, 99*(1), 2538-2545.

Carpenter, T. J., & Zachary, W. W. (2017, March). Using context and robot-human communication to resolve unexpected situational conflicts. In *2017 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)* (pp. 1-7). IEEE.

Casner, S. M., & Hutchins, E. L. (2019). What Do We Tell the Drivers? Toward Minimum Driver Training Standards for Partially Automated Cars. *Journal of Cognitive Engineering and Decision Making, 13*(2), 55-66.

Deery, H. A. (1999). Hazard and risk perception among young novice drivers. *Journal of Safety Research, 30*(4), 225-236.

Dzindolet, M. T., Beck, H. P., Pierce, L. G., & Dawe, L. A. (2001). *A framework of automation use*. Rep. No. ARL-TR-2412. Army Research Laboratory, Aberdeen Proving Ground, MD.

Dzindolet, M. T., Beck, H. P., & Pierce, L. G. (2006). Adaptive automation: Building flexibility into human-machine systems. In S. Burke, G. Pierce, & E. Salas (Eds.), *Understanding Adaptability: A Prerequisite for Effective Performance Within Complex Environments*. Emerald Group Publishing Limited, Bingley, 213-245.

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies, 58*(6), 697-718.

Endsley, M. R. (2017). Autonomous driving systems: A preliminary naturalistic study of the Tesla Model S. *Journal of Cognitive Engineering and Decision Making, 11*(3), 225-238.

European Commission (2019). *Ethics Guidelines for Trustworthy AI*. Retrieved from https://ec.europa.eu/futurium/en/ai-alliance-consultation

James, H. S. (2002). The trust paradox: a survey of economic inquiries into the nature of trust and trustworthiness. *Journal of Economic Behaviour, Organization, 47*(3), 291-307.

Jian, J. Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, *4*(1), 53-71.

Hoffman, R. R., & Woods, D. D. (2011). Beyond Simon's slice: five fundamental trade-offs that bound the performance of macrocognitive work systems. *IEEE Intelligent Systems, 26*(6), 67-71

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608.*

Kirchherr, J., & Charles, K. (2018). Enhancing the sample diversity of snowball samples: Recommendations from a research project on anti-dam movements in Southeast Asia. *PloS One, 13(8),* e0201710.

Koo, J., Kwac, J., Ju, W., Steinert, M., Leifer, L., & Nass, C. (2015). Why did my car just do that? Explaining semi-autonomous driving actions to improve driver understanding, trust, and performance. *International Journal on Interactive Design and Manufacturing (IJIDeM), 9*(4), 269-275.

Kulesza, T., Stumpf, S., Burnett, M., Yang, S., Kwan, I., & Wong, W. K. (2013). Too much, too little, or just right? Ways explanations impact end users' mental models. *In 2013 IEEE Symposium on Visual Languages and Human Centric Computing,* 3-10. IEEE.

Körber, M., Baseler, E., & Bengler, K. (2018). Introduction matters: Manipulating trust in automation and reliance in automated driving. *Applied Ergonomics, 66*, 18-31.

Lippert, S. K., & Michael Swiercz, P. (2005). Human resource information systems (HRIS) and technology trust. *Journal of Information Science, 31*(5), 340-353.

Lee, J., & Moray, N. (1992). Trust, control strategies and allocation of function in human–machine systems*. Ergonomics, 35*, 1243–1270.

Lee, J., & See, K. (2009). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50-80.

Madsen, M., & Gregor, S. (2000). Measuring human-computer trust. *In 11th Australasian Conference on Information Systems, 53*, 6-8.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, *20*(3), 709-734.

Merritt, S. M. (2011). Affective processes in human–automation interactions. *Human Factors, 53*(4), 356-370.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1-38.

Moray, N., Inagaki, T., & Itoh, M. (2000). Adaptive automation, trust, and self-confidence in fault management of time-critical tasks*. Journal of Experimental Psychology: Applied, 6*, 44–58.

Mosier, K. L., & Skitka, L. J. (1996). Human decision-makers and automated decision aids: Made for each other? In R. Parasuraman & M. Mouloua (Eds.), *Automation and human performance: Theory and applications. Human Factors in Transportation* (201-220). Lawrence Erlbaum Associates.

Nathanael, D., Tsagkas, V., & Marmaras, N. (2016). Trade-offs among factors shaping operators decision-making: the case of aircraft maintenance technicians. *Cognition, Technology & Work, 18*(4), 807-820.

Portouli, E., Papakostopoulos, V., Lai, F., Chorlton, K., Hjälmdahl, M., Wiklund, M., Lheureux, F. (2006). Long-term phase test and results. *Final Deliverable (Del. 1.2. 4) of the AIDE IP research project, Brussels, Belgium, Commission of the European Union-DG Information Society and Media (DG INFSO)(contract no. 507674).*

SAE On-Road Automated Vehicle Standards Committee. (2014). Taxonomy and definitions for terms related to on-road motor vehicle automated driving systems. *SAE Standard J, 3016*, 1-16.

Sarter, N.B., Woods, D.D., Billings, C. (1997). Automation surprises. In G. Salvendy (Ed.), *Handbook of human factors/ergonomics* (2nd ed., pp. 1926-1943). Wiley.

Siau, K., & Wang, W. (2018). Building trust in artificial intelligence, machine learning, and robotics. *Cutter Business Technology Journal, 31*(2), 47-53.

Tenhundfeld, N. L., de Visser, E. J., Haring, K. S., Ries, A. J., Finomore, V. S., & Tossell, C. C. (2019). Calibrating trust in automation through familiarity with the autoparking feature of a tesla Model X. *Journal of Cognitive Engineering and Decision Making, 13*(4), 1-16.

Thatcher, J. B., McKnight, D., Baker, E. W., Arsal, R. E., & Roberts, N.H. (2011). The role of trust in postadoption it exploration: An empirical examination of knowledge management systems. *IEEE Transactions on Engineering Management, 58,* 56–70.

Appendix A

Cahour-Forzy-Scale (2009) as proposed by Hoffmann et al. (2018)

In its original version the questionnaire did not incorporate a Likert scale. Hoffmann et al. (2018) added the scales in order to meet psychometric standards, we follow this recommendation:

1. What is your confidence in the system? Do you have a feeling of trust in it?

1	2	3	4	5	6	7
I do not trust it at all						I trust it completely

2. Are the actions of the system predictable?

1	2	3	4	5	6	7
It is not predictable at all						It is completely predictable.

3. Is the tool reliable? Do you think the system is safe?

1	2	3	4	5	6	7
It is not at all safe.						It is completely safe.

4. Is the system efficient at what it does?

1	2	3	4	5	6	7
It is not at all efficient.						It is completely efficient.

We will alter the third question for two reasons. First, reliability can not be judged in the simulation approach we are pursuing. Second, in the proposed questionnaire reliability and safety are not synonymous words for the same concept but measured on the same scale, what is not exactly in line with psychometric standards:

Do you think the system is safe?

1	2	3	4	5	6	7
It is not at all safe.						It is completely safe.

We also altered the fourth question in order to assess perceived effectiveness:

Is the system effective in serving your goals?

1	2	3	4	5	6	7
It is not at all effective.						It is completely effective.

Appendix B

List of all eleven critical scenarios, proposed actions in each scenario and categorizations of actions participants intended after regaining control

Scenario 2

You are on the way to an important meeting. Your car is driving down a highway with a speed limit of 100km/h. You will not reach your destination in time unless your car exceeds the speed limit by at least 20 km/h.

Proposed action in RC: I will stay within the speed limit because going faster could have legal consequences.

Proposed action in CS: In order to reach your destination in time, I will accelerate to 120 km/h, because sticking to the speed limit would mean you will not reach your destination in time.

Distribution of answers obtained from qualitative measurements:

What the other car would do: 87.50%

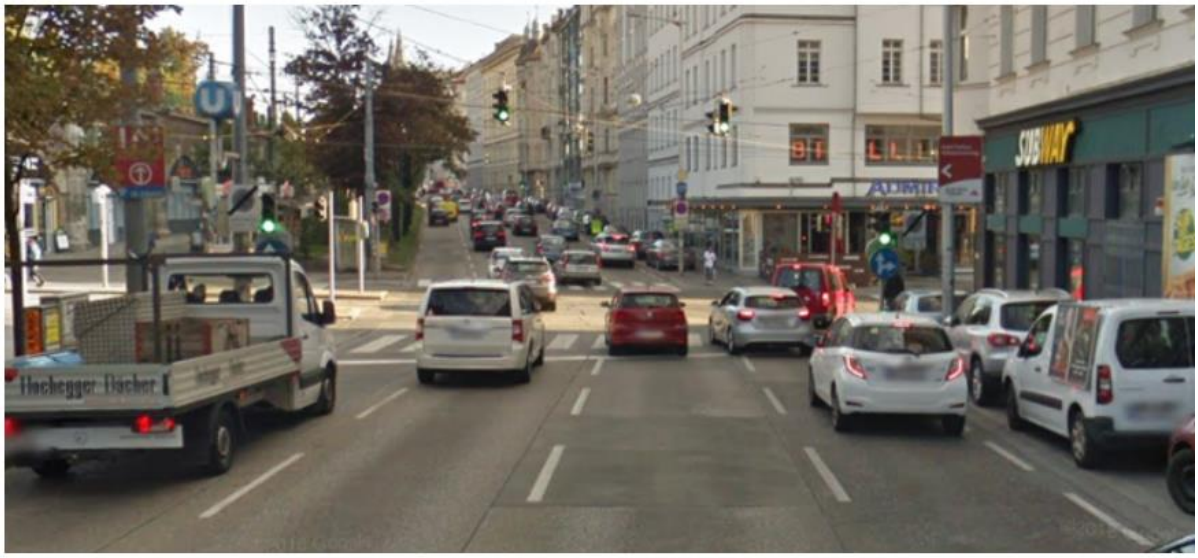
Something Else: 12.50%

Examples of the „something else" category:

"Drive too fast on my own" (CS), "I would gain speed not more than 10km/h" (CS), "I will be driving faster but not 20kmh faster" (CS)

Scenario 3

You are driving on a road with four lanes, approaching a green light. Some cars on the lane to your right are braking. One of the cars appears to cross your lane.



Proposed action in RC: I will not reduce speed, because the car switching lane has to wait until I passed.

Proposed action in CS: I will reduce speed, because otherwise, a crash could occur if one of the braking cars intends to change lanes.

Distribution of answers obtained from qualitative measurements:

What the other car would do: 94.12%

Something Else: 5.88%

Examples of the „something else" category:

"To stop, when it is necessary" (CS)

Scenario 5

You are approaching a yellow light that is turning red. In the rear-view mirror, you can see a car approaching with relatively high speed. You are unsure whether its braking distance is enough for not bumping into you.

Proposed action in RC: I will not cross the red light but stop, because crossing a red light is illegal and could be dangerous.

Proposed action in CS: I will not stop but cross the red light, because otherwise, the car approaching from behind could bump into us.

Distribution of answers obtained from qualitative measurements:

What the other car would do: 78.57%

Something Else: 21.43%

Examples of the „something else" category:

"Instead of stopping or crossing the street i would slightly turn right or left" (CS), "Stop and move to the side if necessary" (CS), "driving to the side" (RC), "brake and switch lanes quickly" (RC)

Scenario 7

You are approaching an intersection. Another car approaches the intersection from the right-hand side at a relatively slow pace, continuously decelerating. It appears as if the second car is stopping for some reason.

Proposed action in RC: I will stop and wait for the other car to make use of its right of way.

Proposed action in CS: I will decelerate but not stop, because the other car did not make use of its right of way.

Distribution of answers obtained from qualitative measurements:

What the other car would do: 95.65%

Something Else: 4.35%

Examples of the „something else" category:

"Stop and give him a sign to go an. He could still change his decision and I would be responsible for an accident" (CS)

Scenario 9

You are driving down a narrow street with cars parking on both sides. There's a speed limit of 30 km/h.

Proposed action in RC: I will not drive slower than 30 km/h because on this road it is allowed to keep this speed.

Proposed action in CS: I reduce my speed to a slower pace because driving 30 km/h could be dangerous with that little overview.

Distribution of answers obtained from qualitative measurements:

What the other car would do: 100.00%

Something Else: 0.00%

Scenario 10

You are driving down a small inner-city road. A pedestrian is standing on the right side of the street. He's not looking in your direction and there is no crosswalk. For some reason the car in front of you is decelerating.

Proposed action in RC: I will not stop to check if the pedestrian wants to cross the street because there is no crosswalk.

Proposed action in CS: I will stop to check if the pedestrian wants to cross the street because going on would not save that much time, as the car in front is stopping anyways.

Distribution of answers obtained from qualitative measurements:

What the other car would do: 61.11%

Something Else: 38.89%

Examples of the „something else" category:

"Slow down" (CS), "reduce the speed instead of stopping" (CS), "Reduce my speed" (CS), "Also slow down but in that situation i'd like to have the control" (CS), "Decelerating because the car in front of me is slowing down" (RC), "Regardless the pedestrians action- the minivan in front of me decelerates. So I do also to avoid an [rest is missing]" (RC), "Stop because of the car jn front of me" (RC),

Scenario 15

You are searching for a parking spot for quite a while now. To your right, there is a free spot in a one-way street.

Proposed action in RC: I will drive around the block to enter the one-way street in the right direction because entering it in the wrong direction is forbidden.

Proposed action in CS: I will enter the one-way street in the wrong direction instead of driving around the block because this would take a while and someone else could take the free spot.

Distribution of answers obtained from qualitative measurements:

What the other car would do: 94.12%

Something Else: 5.88%

Examples of the „something else" category:

"Drive Backwards into the oneway" (CS), "search on my own, sometimes you can park where there's no offial parking but still good" (CS)

Scenario 17

You are approaching an intersection with a stop sign in front of you.

Proposed action in RC: Although I have a good view to the left, I am stopping instead of only reducing speed, because ignoring the stop sign would be illegal.

Proposed action in CS: Although there is a stop sign, I am only reducing speed instead of stopping, because the view to the left is unobstructed.

Distribution of answers obtained from qualitative measurements:

What the other car would do: 100.00%

Something Else: 0.00%

Scenario 18

While waiting in front of a red light your car notifies you that you are low on gas. The next gas station is to your left. You are currently on the lane leading straight ahead.

Proposed action in RC: I will follow my lane straight ahead because there is a solid line preventing us from turning left.

Proposed action in CS: I will cross the solid line and turn left to drive to the gas station as soon as lights turn green as there is no other car on the left turn lane at the moment.

Distribution of answers obtained from qualitative measurements:

What the other car would do: 88.46%

Something Else: 11.54%

Examples of the „something else" category:

"Switch to the left lane, while there is no car and the traffic light is red" (CS), "I would only turn right if i cant't see any car behind me or behind the car on the right lane" (CS), "The car should turn on the left lane immediately and not wait until it is green. Until then another car could be on the left lane" (CS)

Scenario 19

You are approaching an intersection, the light in front of you just turned yellow. At your current speed you are within braking distance.

Proposed action in RC: I am stopping instead of crossing the intersection because I am within braking distance, hence, I am obliged to stop.

Proposed action in CS: I am crossing the intersection instead of stopping because I can make it before the light turns red.

Distribution of answers obtained from qualitative measurements:

What the other car would do: 100.00%

Something Else: 0.00%

Scenario 23

You realized that you forgot your wallet at home. Therefore, you changed your destination to get back home. To reach your new destination you left th highway you were on and now you have to re-enter the highway in the opposite direction.

Crossing the dashed lines in the middle of the road is only allowed for cars entering the street from the left access road.

Proposed action in RC: I will drive straight on and turn back to the highway at the next possibility because crossing the dashed lines in the middle of the road to turn back is not allowed.

Proposed action in CS: I will cross the dashed lines in the middle of the road to turn back to the highway because going straight on would take more time.

Distribution of answers obtained from qualitative measurements:

What the other car would do: 100.00%

Something Else: 0.00%