Psychological factors influencing perceptions of autonomous vehicles and computerised systems

by

Owain Ritchie

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Department of Psychology

August 2020

Table of contents

List of tables

List of figures

Acknowledgments

First, I would like to thank my primary supervisor, Derrick Watson, for all of his advice and guidance throughout the whole process of the PhD. I could not have asked for a better supervisor and I am very grateful for his support. I would also like to thank the principal investigator of the Cooperative Car research group, Nathan Griffiths, for his invaluable advice and all of the hard work he has put into the project.

Next, I would like to thank all of my colleagues and collaborators from the departments of Computer Science, Psychology, Warwick Business School, and Jaguar Land Rover for their support and input into the work, as well as Jaguar Land Rover in particular for funding and supporting my research and making all of this possible.

I would also like to thank all the participants who have taken part in the research across the last few years – none of this research would be possible without your help. Seeing people's reactions the first time they see a driving simulator is also a very encouraging thing!

Finally, I would like to give my sincere thanks to all the friends and family who have been an invaluable support network for me every step of the way. At all of the most difficult and challenging parts of this process, your support has made all the difference.

Declaration

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree. The work presented (including data generated and data analysis) was carried out by the author.

Parts of this thesis have been published by the author. Specifically, Chapter 3 (*How should autonomous vehicles overtake other drivers?*) has been adapted from the following publication:

Ritchie, O. T., Watson, D. G., Griffiths, N., Misyak, J., Chater, N., Xu, Z., & Mouzakitis, A. (2019). How should autonomous vehicles overtake other drivers? *Transportation Research Part F: Traffic Psychology and Behaviour, 66,* 406 – 418.

This work was supported by Jaguar Land Rover and the UK-EPSRC grant EP/N012380/1 as part of the jointly funded Towards Autonomy: Smart and Connected Control (TASCC) Programme.

Abstract

Autonomous (self-driving) vehicles have been predicted to start arriving on roads in the very near future. Despite concerns over how self-driving vehicles will integrate into human-driven traffic, little empirical work has examined interactions between autonomous vehicles and human drivers. With this in mind, this thesis presents two complimentary streams of research which examined the psychological factors influencing evaluations of autonomous vehicles and computerised systems. The first stream consisted of two sets of driving simulator experiments which looked at perceptions of autonomous overtaking manoeuvres and occupant satisfaction during autonomous highway journeys. Overall, these experiments revealed that occupants and other road users are very sensitive to the driving style of the autonomous vehicle itself, with a smaller influence of traffic context, which sometimes resulted in autonomous vehicles being evaluated more negatively due to the behaviour of other drivers. The second stream consisted of two sets of laboratory-based experiments, which applied findings on the influence of relative ranking and framing on decision making to trust of computerised systems. Overall, this second stream found that users' evaluations were very sensitive to the accuracy of systems, slightly sensitive to some contextual factors, with little or no influence of framing or rank-based effects. Across both streams, the findings of this thesis suggest a limited degree of sensitivity to contextual factors when evaluating the behaviour of autonomous vehicles and computerised systems. The results are discussed in terms of implications for psychological theory and practical considerations for the introduction of autonomous vehicles.

Abbreviations

ACC	Adaptive Cruise Control
ANOVA	Analysis of Variance
BF	Bayes Factor
BPM	Beats per Minute
ECG	Electrocardiogram
EDA	Electrodermal Activity
FPS	Frames per Second
HGV	Heavy Goods Vehicle
HTML	Hypertext Markup Language
LCD	Liquid Crystal Display
µs	Microsiemens
MPH	Miles per Hour
PX	Pixels
RPM	Revolutions per Minute
SCL	Skin Conductance Level
SCR	Skin Conductance Response
SSQ	Simulator Sickness Questionnaire
SXGA	Super XGA
TRL	Transport Research Laboratory
VCQ	Vehicle Character Questionnaire
WMG	Warwick Manufacturing Group

1. General introduction

Abstract

This thesis presents two complimentary streams of research, focusing on the psychological factors that influence evaluations of autonomous vehicles, and trust of in-vehicle display systems and instrumentation. This general introduction covers previous research on factors influencing trust in these domains, leading into the research questions addressed by this thesis. Overall, previous work into evaluations of autonomous vehicles has found variables such as demographic variables, appearance of the vehicle, information provision and driving styles to influence trust. Despite concerns over interactions between human-driven and autonomous vehicles, there have been few empirical studies on the factors influencing perceptions of these interactions. These gaps are addressed by the first stream of the thesis (Chapters 3 & 4). In terms of trust of systems more generally, factors such as system appearance, accuracy and information provision have been shown to be influential. The second stream of this thesis extends previous research by investigating the influence of factors such as relative rank (Chapter 5) and framing effects (Chapter 6) on evaluations of context-neutral systems.

1.1. Overview

This thesis presents two interlinked streams of research, focusing on user evaluation of autonomous vehicles based on driving style ("Vehicle Character" – Chapters 3 & 4) and user evaluation of information presentation systems ("System Character" – Chapters 5 & 6). As such, this general introduction focuses on two key overlapping areas of background research: i) trust and evaluation of autonomous vehicles and driving styles (background for Vehicle Character stream), and ii) trust in information provision systems (background for System Character stream). The aim of this introduction is to provide a high-level overview of relevant existing work on the evaluation of autonomous vehicles and computerised systems, and to outline how this prior work ties into the novel contributions of this thesis. Within each chapter, the most relevant and comparable previous research is discussed in more detail as it relates to the focus of that specific chapter. Before surveying previous literature on trust of autonomous vehicles and computerised systems, this introduction will consider the wider context around the introduction of autonomous vehicles into human-driven traffic, outlining the importance of understanding the factors that influence trust of autonomous vehicles and in-vehicle systems.

Levels of automation have been defined as moving through 0 (no automation) to 5 (full autonomy in all driving modes), with intermediate levels in which the driver remains responsible for supervising the driving task and may be required to regain manual control (SAE J3016_201806, 2018). The work discussed in this literature review will cover multiple levels of autonomy, although much of the previous work has focused on lower levels of automation. As such, the empirical work presented in this thesis will focus on evaluations of fully autonomous vehicles (corresponding to Level 5 autonomy, Chapters 3 & 4) and in-vehicle systems (Chapters 5 & 6), into which the user does not have any direct manual input.

1.2. Introduction of autonomous vehicles

Autonomous (self-driving) vehicles have been predicted to begin arriving on roads as soon as 2020 ("Driverless car market watch", n.d.). One estimate has predicted that, by 2040, 75% of all vehicles on the road could be autonomous ("Look Ma, No Hands!", 05/09/2012), although a more recent prediction has indicated 2045 as a more realistic estimate for half of all new vehicles being autonomous (Litman, 2020). Autonomous vehicles have been argued to offer several important benefits for the future of transport and wider society, such as improving public health and road safety (Fleetwood, 2017; Morando, Tian, Truong & Vu, 2018; Pettigrew, Talati & Norman, 2018), positive environmental impacts (Igliński & Babiak, 2017; Jones & Leibowicz, 2019; Kopelias, Demiridi, Vogiatzis, Skabardonis & Zafiropoulou, 2020),

and opening up more transportation options to those who are unable to drive, due to factors such as age and medical conditions (Harper, Hendrickson, Mangones & Samaras, 2016; Penmetsa, Adanu, Wood, Wang & Jones, 2019).

While rapid technological progress is being achieved and the potential benefits are promising, it has been acknowledged that, at least in the initial stages, there will be a period of mixed traffic where autonomous and human-driven vehicles must share the road to some extent. In order for the transition to full autonomy to be successful, this transition period will require interactions between autonomous vehicles and other road users to be safe and acceptable (Hancock, 2019; Hancock, Nourbakhsh & Stewart, 2019). However, previous work has indicated that drivers are concerned that human-driven vehicles may take advantage of rule-based autonomous vehicles and behave aggressively towards them (Tennant, Howard, Franks, Bauer & Stares, 2016), and that autonomous vehicles may have trouble adapting implicit rules and conventions governing interactions between human drivers (Chater, Misyak, Watson, Griffiths & Mouzakitis, 2018; Tennant, 2015). One observational study found that, while road users overall behaved cautiously towards autonomous vehicles, riskier behaviours were adopted by other road users when the infrastructure did not support separate spaces for autonomous vehicles and other traffic (Madigan et al., 2019).

The potential for aggressive behaviours towards autonomous vehicles from human drivers presents the issue of potentially reduced journey satisfaction and comfort for occupants of autonomous vehicles. This highlights the need for smooth interactions between human-driven and autonomous vehicles in shared spaces. In order to ensure a successful transition through a mixed period of autonomous and human-driven traffic, research into the psychological factors behind what makes such interactions acceptable (from the perspective of both the occupant and other road users) is required. As such, Chapters 3 and 4 of this thesis focus directly on the issue of interactions between human drivers and autonomous vehicles.

1.3. Self-reported perceptions of autonomous vehicles

Previous research into perceptions of autonomous vehicles has used self-report survey-based methodologies to investigate which factors correlate with people's levels of *a priori* acceptance. For instance, factors such as personality traits (Choi & Ji, 2015; Kyriakidis, Happee & de Winter, 2015), daily driving behaviours (Howard & Dai, 2014) and experience (Bansal & Kockelman, 2018), cultural differences (Haboucha, Ishaq & Shiftan, 2017; Kyriakidis et al., 2015; Schoettle & Sivak, 2014), gender (Hohenberger, Spörrle & Welpe, 2016; Howard & Dai, 2014; Hulse, Xie & Galea, 2018; Schoettle & Sivak, 2014) and age (Bansal & Kockelman,

2018; Haboucha, et al., 2017; Hohenberger et al., 2016; Hulse et al., 2018; Kyriakidis et al., 2015; Rovira, McLaughlin, Pak & High, 2019) have been linked to differing attitudes towards autonomous vehicles. A multi-survey review found that younger people, males, people in urban environments, and drivers more familiar with driving aids and more up to date with technological news showed more positive attitudes towards autonomous vehicles (Becker & Axhausen, 2017).

Survey-based methodologies provide the advantage of large datasets on the influence of demographic variables on trust of autonomous vehicles. However, they are limited in that they cannot examine individuals' trust in autonomous vehicles based on the experience of being an occupant, or evaluations of specific driving behaviours. Previous reviews (Hoff & Bashir, 2015; Marsh & Dibben, 2003) have distinguished between three levels of system trust: *Dispositional* trust (based on user characteristics), *situational* trust (based on context and interaction with a system) and *learned* trust (built up over time with experience of using system). Although survey-based work allows examination of dispositional trust of autonomous vehicles (e.g. based on demographic variables), investigation of users' situational and learned trust requires methodologies in which users have direct experience viewing and interacting with autonomous vehicles and systems. As such, the next section will review previous work on the evaluation of the features and behaviour of automated and autonomous vehicles from an occupant/driver perspective.

1.4. Evaluations of autonomous vehicle behaviour and driving styles

In order to test more experience-based evaluations (Hoff & Bashir, 2015), previous research has utilised driving simulators and, more recently, real-world testing to examine road user perceptions of autonomous vehicles. Research in this area has focused on both overall levels of trust in the vehicle, and (to a lesser extent) evaluations of specific driving styles or manoeuvres.

Previous research has linked both the appearance of, and information provided by, in-vehicle interfaces to levels of trust in autonomous vehicles. For instance, Beller, Heesen and Vollrath (2013) found that trust in an automated vehicle was increased by providing occupants with feedback on the system's level of uncertainty, while Helldin, Falkman, Riveiro and Davidsson (2013) found a decrease in trust when participants were presented with uncertainty information. Helldin et al. (2013) interpreted this finding as evidence that participants did not place an excessive level trust in the vehicle, which could be problematic in the case of lower levels of automation where the driver may be required to regain manual control (SAE

J3016_201806, 2018). One study found that in-vehicle display systems that gave instructions on how to deal with traffic scenarios were rated more positively than those that simply provided information, but only in light traffic (Cramer, Evers, Kemper & Wielinga, 2008). Drivers presented with descriptions of automated systems have been shown to favor those that provide more information and align more closely with the users' driving goals (Verberne, Ham & Midden, 2012). Trust in autonomous vehicles has also been shown to be sensitive to the vehicle's features and appearance, with autonomous vehicles and agents that had more anthropomorphised (human-like) features being rated as more trustworthy, compared to those with more robot-like features (Lee, Kim, Lee & Shin, 2015; Waytz, Heafner & Epley, 2014).

Users' prior expectations and the nature of their introduction to autonomous vehicles has been shown to influence overall evaluations of trust. For instance, Hartwich, Witzlack, Beggiato and Krems (2018) found that occupant's initial evaluations of autonomous vehicles become more positive with increasing experience with a simulated autonomous vehicle. In the domain of recovering control from automated vehicles, being provided with more information on a vehicle's limitations has been shown to reduce the amount of attention towards non-driving tasks (Körber, Baseler & Bengler, 2018), in line with Helldin et al.'s (2013) finding that information prevented over-trust in the vehicle's automation. Drivers of automated vehicles have been shown to recover manual control more quickly if they are expecting automation to be switched off at regular intervals (Merat, Jamson, Lai, Daly & Carsten, 2014), and also recover control more slowly if the level of automation decreases through gradual steps rather than progressing directly to manual control (Abe, Sato, Uchida & Itoh, 2019). Drivers of automated vehicles have also been shown to recover manual control more slowly in the event of automation failure if they had higher levels of trust in the system, but not when they had been given more extensive practise with the vehicle (Payre, Cestac & Delhomme, 2016).

In addition to focusing on general trust of autonomous vehicles, some previous work on driving styles has examined the effect of factors such as participants' manual driving styles and traffic context on choice/evaluation of autonomous driving styles. Users of an autonomous vehicle in one test track study rated more cautious driving styles as more trustworthy than more aggressive driving styles (Ekman, Johansson, Bligård, Karlsson & Strömberg, 2019), with driving simulator-based studies showing that occupants tend to prefer autonomous driving styles that are more cautious than their own driving style (Abe, Sato & Itoh, 2017; Basu, Yang, Hungerman & Dragan, 2017). A simulator-based study by TRL found that participants preferred to change lane after two vehicles had passed (rather than move between them), regardless of whether the second vehicle was autonomous or human-driven, and irrespective

of the vehicles' appearance (TRL PPR807, 2017). An on-track study in which occupants experienced their vehicle overtaking a parked car found that participants gave higher ratings of trust when there was oncoming traffic (compared to no oncoming traffic (Venturer Trial 2, 2017)). This suggests some influence of traffic context on the degree to which certain manoeuvres are acceptable.

Several studies on evaluation of autonomous driving styles have focused on following distance (or time headway) between vehicles. For instance, some work into the use of driving assistance systems that regulate following distance has shown that drivers are often accepting of time headways of less than 1 second (Nowakowski, O'Connell, Shladover & Cody, 2010), although are more likely to hover over the brake at shorter time headways, possibly indicating greater indication to regain manual control (Balk, Jackson & Philips, 2017). Previous research on driving styles has also examined interactions between time headway and speed on drivers/occupants' experience of journeys. Simulator-based experiments in this domain have shown reductions in self-reported levels of comfort below a 2-second time headway, independent of speed (Lewis-Evans, De Waard & Brookhuis, 2010; Siebert, Oehl & Pfister, 2014). When provided with ascending and descending sequences of time headways between 0.5 and 4s, one simulator study found that drivers started to feel more comfortable at around 1.5s when the time headway was increasing but reported becoming less comfortable at around 2-2.5s when it was decreasing (Siebert, Oehl, Bersch & Pfister, 2017). This suggests that there is an asymmetry in how time headway is evaluated, based on whether the driver is catching up to or pulling away from the vehicle ahead.

Relatedly, previous research has examined effects of autonomous 'platooning' (very small distances between a series of autonomous vehicles) on driving comfort and manual driving styles. For example, simulator-based work has shown that drivers are more likely to use shorter time headways both after driving alongside tightly-spaced platoons of lorries (Gouy, Wiedemann, Stevens, Brunett & Reed, 2014), and after being the occupant of an autonomous vehicle in a tightly-spaced platoon (Skottke, Debus, Wang & Huestegge, 2014). These findings suggest that manual driving behaviour can be influenced both by the behaviour of other road users and recent experience of autonomous driving, highlighting the importance of understanding knock-on effects of mixed traffic environments in the introduction of autonomy. In terms of trust and comfort, other simulator-based work has found that occupants of vehicles in autonomous platoons are most comfortable when they are able to freely choose a non-driving task to complete (Heikoop, de Winter, van Arem & Stanton, 2017). Occupants have been shown to trust systems with higher levels of automation less than basic driver assistance

systems when part of a platoon, although report higher levels of subjective workload when using lower levels of automation (Hjälmdahl, Krupenia & Thorslund, 2017).

While there has been a significant amount of work into trust of autonomous vehicles and evaluations of autonomous driving styles, there has been less work specifically focusing on interactions between human-driven and autonomous vehicles in shared spaces. Previous survey-based work has indicated that drivers are concerned that human-driven vehicles may take advantage of rule-based autonomous vehicles and behave aggressively towards them (Tennant, et al., 2016), and there are concerns that autonomous vehicles may have trouble adapting to implicit rules and conventions governing interactions between human drivers (Chater et al., 2018; Tennant, 2015). The potential for aggressive behaviours towards autonomous vehicles from human drivers presents the issue of potentially reduced journey satisfaction and comfort for occupants of autonomous vehicles, highlighting the need for smooth interactions between human-driven and autonomous vehicles in shared spaces. While some of the driving styles work covered above has covered this area (e.g. Abe et al., 2017; Basu et al., 2017; Gouy et al., 2014; TRL PPR807, 2017; Venturer Trial 2, 2017), there has been little focus on evaluation of specific autonomous driving behaviours from multiple road user perspectives. In addition, much of the work covered above has focused on time headway/following distance, with manoeuvres such as autonomous overtaking receiving less focus.

With these gaps in mind, Chapter 3 presents a simulator-based investigation of autonomous overtaking manoeuvres from both occupant and driver perspectives, and also investigates the role of traffic context and immersion level. Immersion was defined here as level of realism, with the main comparison being between a highly realistic driving simulator setting, and a less realistic setup in which driver's-eye footage from the simulator was presented on computer screens. Chapter 4 also investigates the influence of position in a stream of traffic (overtaking vs being overtaken) on occupant experiences of autonomous highway journeys. In addition to focusing on the effects of traffic context and autonomous driving styles, Chapter 4 examines the influence of information presentation from in-vehicle systems on perceptions of autonomous journeys. Chapter 4 therefore represents a bridge between the more applied driving style-based work (Vehicle Character) and the work on information presentation systems covered in Chapters 5 and 6 (System Character). As such, previous background work on information presentation (both within in-vehicle settings and more laboratory-based research) is discussed in the section on System Character below, as a general background to Chapters 5 and 6, and to some extent Chapter 4.

1.5. Trust in computerised systems

While the first stream of this thesis focuses on the evaluations of autonomous vehicles, in terms of specific driving behaviours (Chapter 3) and general evaluations of journeys (Chapter 4), the second stream takes a broader, more theoretical approach, focusing on trust of systems in general. System trust has been previously operationalised as *"a positive expectation regarding the behavior of somebody or something in a situation that entails risk to the trusting party"* (Marsh & Dibben, 2003, pp. 470). System trust can be influenced by factors relating to the human user (e.g. the occupant of an autonomous vehicle), the features/behaviour of the system itself, and the context in which the interaction occurs: Previous work suggests that factors relating to the automation itself have the largest influence on trust (Hancock et al., 2011; Schaefer, Chen, Szalma & Hancock, 2016). Trust can also be divided into *dispositional* (based on characteristics of the user, e.g. demographics), *situational* (based on the system's behaviour and the context of the interaction) and *learned* trust that is built up over time and experience (Hoff & Bashir, 2015; Marsh & Dibben, 2003). As with the experiments presented in Chapters 3 and 4, the focus of this second stream is on the *situational* factors influencing system trust, such as the performance of the systems being evaluated.

Previous research has shown that the visual appearance of computerised systems can influence user trust. For instance, users show higher levels of trust in autonomous agents that have a more human-like appearance (Lee et al., 2015), and autonomous vehicles with more anthropomorphised features have been shown to be more trustworthy than those with a more robotic appearance, and are also blamed to a lesser extent in the event of an automation failure (Waytz et al., 2014). A study in the medical domain found that decision aids with a more anthropomorphised appearance were rated as more trustworthy, but only for younger users, with no influence of anthropomorphism on trust reported by older adults (Pak, Fink, Price, Bass & Sturre, 2012). In-vehicle displays with a more realistic level of detail have been shown to be more trustworthy than those that represent the road in a more simplified manner (Barthou, Kemeny, Reymond, Mérienne & Berthoz, 2010).

The type of information provided to users has also been shown to influence system trust. In the automotive domain, trust in autonomous vehicles has been shown to be sensitive to information on the system's level of uncertainty (Beller et al., 2013; Helldin et al., 2013), with some evidence that displaying system confidence information increases trust in decision aid systems (Verame, Constanza & Ramchurn, 2016). Users of information display systems have also been shown to give more positive evaluations of systems that provide frequently updated

information (McGuirl & Sarter, 2006), as well as systems that provide more detailed information (Barthou et al., 2010; Cramer et al., 2008; Verberne et al., 2012). Providing users with more information on the inner workings of a system has been shown to decrease trust if the systems' performance is poor (Kaltenbach & Dolgov, 2017), and providing users with information on the reliability of a system reduces trust in that system if the information itself is inaccurate (Barg-Walkow & Rogers, 2016). A study on system explanations found that users reported having a more complete understanding of systems that provided "*why*" explanations (justifying why the system made a particular decision), compared to those that provided "*why not*" explanations (justifying why the system *did not* make a different decision: Lim, Dey & Avrahami, 2009).

The performance of computerised systems has also been linked to user trust. Systems with higher levels of accuracy tend to be rated as more trustworthy (Chancey, Proaps & Bliss, 2013; Madhavan & Phillips, 2010; Yu et al., 2017). Trust in computerised systems has also been shown to drop rapidly with decreasing system accuracy (Chavaillaz & Sauer, 2017; Chavaillaz, Wastell & Sauer, 2016; Yu et al., 2017). Kaltenbach and Dolgov (2017) found that less accurate systems were rated as less trustworthy, but only when increased levels of system transparency revealed the inner workings of the system, suggesting some interplay between information provision and system performance. Further to this interaction, trust and system interaction research has shown that framing information about levels of system accuracy negatively (i.e. "20% incorrect" rather than "80% correct") can lead to reduced purchase intentions (Cheng & Wu, 2010), reduced intentions to follow up on system advice (Huerta, Glandon & Petrides, 2012), and increased likelihood of missing targets in signal detection tasks (Lacson, Wiedemann & Madhavan, 2005). In addition to how often a system makes the incorrect decision, the nature of the incorrect decision has also been shown to influence users' performance when assisted by the system. For instance, false alarms (indicating that a target is present when it is absent) have been found to have a larger negative impact on performance compared to misses (indicating that a target is absent when it is present (Chancey, Yamani, Brill & Bliss, 2017; Dixon, Wickens & McCarley, 2007; Rice & McCarley, 2011; Wiczorek & Meyer, 2016)).

The second stream of this thesis presents two sets of experiments, covered in Chapters 5 and 6, which aimed to extend previous research by investigating whether psychological biases involving ranking and value judgment apply to the domain of system trust. Chapter 5 investigated the combined influence of relative ranking and objective accuracy on trust of information display systems, following on from findings in behavioural science which show

that decision making is highly sensitive to the ranking of options rather than their objective value (Stewart, 2009; Stewart, Chater & Brown, 2006; Vlaev, Chater, Stewart & Brown, 2011; Walasek & Stewart, 2015; Walasek & Stewart, 2019). Chapter 5 also made the contribution of using a relatively context-neutral methodology (rather than asking participants to evaluate analogues of real-world systems), as there is some evidence that knowledge of existing systems can bias evaluations of novel systems in trust research (Lim et al., 2009), and that trust of automation can vary dependent on the domain being studied (Pak, Rovira, McLaughlin & Baldwin, 2017). Chapter 6 also investigated evaluation of information display systems (using a more applied setting of battery estimation/charging for electric vehicles) but focused on the effects of system variability and framing biases on users' evaluation. Previous research into value integration has revealed that, when asked to choose between two streams of digits based on value, people's judgments can be biased by the framing of the question – leading to paradoxical decision making (Glickman, Tsetsos & Usher, 2018; Tsetsos, Chater & Usher, 2012; Usher, Tsetsos, Glickman & Chater, 2019). Chapter 6 tested whether these paradoxical choices still apply in the domain of trust of in-vehicle systems, presenting five experiments on the influence of system variability and framing effects on users' choices.

1.6. Summary and link to the present work

This thesis focuses on the psychological factors influencing evaluations of autonomous vehicles and information display systems. These two overlapping topics are reflected in two complimentary streams of research, which consisted of four sets of experiments detailed in Chapters 3 – 6. In terms of background, research into evaluations of autonomous vehicles has revealed factors such as demographics, information presentation and driving style as influential. However, despite concerns over how autonomous and human-driven vehicles will interact in shared road environmental, little empirical work has been done on this topic. Accordingly, the first – more applied – stream presents several driving simulator-based experiments on factors influencing perceptions of autonomous overtaking manoeuvres (Chapter 3) and highway journeys (Chapter 4). Research into trust of computerised systems and instrumentation more widely has revealed factors such as the type and framing of information provided and level of objective accuracy as important. The second – more theoretical – stream of this thesis aims to extend this literature by examining whether evaluation of context-neutral systems is sensitive to relative ranking (Chapter 5) and whether users of computerised systems are sensitive to framing biases and paradoxical choice in value integration (Chapter 6). This general introduction has presented an overview of key findings

and background, whereas the introduction sections for each chapter will provide more detail on the most relevant research for that chapter. The next chapter provides an overview of the two streams and associated experiments that are presented across the remainder of this thesis.

2. Overview of the thesis

The thesis is divided into two main streams. One stream focuses on the evaluation of autonomous vehicles and driving styles (Chapters 3 & 4), and the other on the evaluation of computerised systems (Chapters 5 & 6). The experiments presented in Chapters 3 and 4 mostly utilise driving simulator-based methodologies, in order to increase levels of immersion and to reflect the more real-world focus of the work, whereas the experiments in Chapters 5 and 6 use more abstract laboratory-based methodologies, due to the work being more theoretical in nature. Across four chapters of experiments relating to these two streams, the thesis presents twenty-three experiments on psychological factors influencing perceptions of autonomous vehicles and computerised systems. Chapter 7 integrates findings across all four empirical chapters, discussing theoretical and practical implications as well as suggestions for future research. The research questions and methodologies of each of the four empirical chapters are outlined below.

2.1. Vehicle character stream

The first stream of the thesis focuses on perceptions of the behaviour and driving styles of autonomous vehicles, from both an occupant perspective (Chapters 3 & 4) and the perspective of other road users (Chapter 3). The work in Chapter 3 investigated evaluations of autonomous overtaking manoeuvres, as a function of pull-in distance, vehicle perspective, and the presence and following distance of a third vehicle. Chapter 4 investigated occupants' evaluations of highway journeys in an autonomous vehicle, as a function of lane position, vehicle speed, information presented by the vehicle and traffic context. The three experiments presented in Chapter 4 were all driving simulator-based, however Chapter 3 also presented three driving simulator experiments with three video-based replications, in order to examine potential effects of the level of immersion. While the focus of Chapter 3 is entirely on the evaluation of autonomous vehicles based on their behaviour and the surrounding traffic context, the manipulation of information presented to occupants in Chapter 4 provides links to the second stream of the thesis, which focuses on trust of instrumentation designed to present users with information.

2.2. System trust stream

The second stream of the thesis focuses on users' trust of computerised systems that attempt to reach a target state (e.g., estimate remaining vehicle range or charge a battery to capacity) while providing participants with different types of information. Chapter 5 presents

nine experiments which investigated the influence of accuracy, relative rank position, task complexity and closeness to solution on system trust. Chapter 6 presents five experiments which tested for effects of framing, system variability and the level of information provided on participants' evaluations of systems. Although both Chapters 5 and 6 use relatively abstracted and context-neutral methodologies, the psychological factors investigated in this more theoretical stream can be applied to evaluations of in-vehicle display systems.

3. How should autonomous vehicles overtake other drivers?

Abstract

Previous research that has examined trust of autonomous vehicles has largely focused on holistic trust, with less work on evaluation of specific behaviours and interactions with human-controlled vehicles. Six experiments examined the influence of pull-in distance, vehicle perspective (overtaking/being overtaken), following distance and immersion on self-reported evaluations of, and physiological responses to, autonomous motorway overtakes. The results showed: i) overtake manoeuvres were viewed more positively as pull-in distance increased before reaching a plateau at approximately 28 meters, ii) physiological-based orienting responses occurred for the smallest pull-in distances, iii) participants being overtaken were more forgiving of a closer pull-in if the overtaking vehicle was followed closely by another vehicle, and iv) for two of three cross-experiment comparisons participants were more forgiving of smaller pull-in distances when experiencing lower levels of immersion. Overall, the results suggest that the acceptability of an overtake manoeuvre increases linearly with pull-in distance up to a set point for both overtaking and being overtaken, with some influence of traffic context and levels of immersion. The findings are discussed in terms of implications for the development of assisted and fully autonomous vehicle systems that perform in a way that will be acceptable to both the vehicle occupants and other road users.

3.1. Introduction

Autonomous vehicles may arrive on the roads by approximately 2020 ("Driverless car market watch", n.d.), with one estimate claiming that 75% of all road vehicles will be autonomous by 2040 ("Look Ma, No Hands!", 05/09/2012). Despite this prediction, fully autonomous vehicles will still have to share the road with human-driven vehicles especially during the transition period. Even when most vehicles are autonomous, some human driven vehicles (e.g., bicycles, motorcycles and 'motor enthusiast' vehicles) may remain on the roads. Hence, safe and acceptable interactions between autonomous and human-driven vehicles will remain vital. Furthermore, even if all vehicles become fully autonomous, their behaviour will still need to be acceptable to their occupants and other road users in general. Accordingly, it is important to understand the psychological factors that influence how autonomous vehicle behaviours are perceived.

Survey-based methods have identified factors such as gender (Hohenberger, Spörrle & Welpe, 2016; Howard & Dai, 2014; Hulse, Xie & Galea, 2018; Schoettle & Sivak, 2014), age (Bansal & Kockelman, 2018; Haboucha, Ishaq & Shiftan, 2017; Hohenberger et al., 2016; Hulse et al., 2018; Kyriakidis, Happee & de Winter, 2015), personality (Kyriakidis, et al., 2015; Choi & Ji, 2015), cultural differences (Haboucha et al., 2017; Kyriakidis et al., 2015; Schoettle & Sivak, 2014) daily driving behaviours (Howard & Dai, 2014) and experience (Bansal & Kockelman, 2018) as influencing acceptance and intentions to use autonomous vehicles. A review of several surveys found that attitudes towards autonomous vehicles were most positive amongst males, younger people, those living in urban environments, drivers with previous experience of assistance systems and individuals who were more familiar with the news on technological developments (Becker & Axhausen, 2017).

A review by Hoff and Bashir (2015) defined 'trust' as comprising three distinct components: dispositional, situational, and learned. The survey-based work above mainly investigated how dispositional trust is influenced by various demographic factors, rather than evaluations after interacting with and gaining experience with a system (situational and learned trust). To examine the latter two aspects of Hoff and Bashir's (2015) taxonomy, experimental designs in which participants interact with autonomous vehicles and/or in-vehicle interfaces are required.

Previous experimental work has focused on judgements relating to the overall trust of a vehicle, based on features of the vehicle itself or information provided about the vehicle/by an in-vehicle interface. For example, simulator-based work found higher levels of trust for

autonomous vehicles and driving agents that possess more anthropomorphised features (Lee, Kim, Lee & Shin, 2015; Waytz, Heafner & Epley, 2014). Providing feedback on an autonomous vehicle's level of uncertainty influenced trust (Beller, Heesen & Vollrath, 2013; Helldin, Falkman, Riveiro & Davidsson, 2013), experience with an autonomous vehicle led to increased positivity of initial attitudes (Hartwich, Witzlack, Beggiato & Krems, 2018), and drivers paid less attention to non-driving tasks in an automated vehicle when given information on the vehicle's limitations (Körber, Baseler & Bengler, 2018). With respect to takeover activity, Payre, Cestac and Delhomme (2016) found higher trust was associated with slower manual control recovery, but this effect disappeared when participants were given more elaborate practise with the vehicle.

However, comparatively little work has examined evaluations of specific vehicle behaviours between autonomous and human-driven vehicles. Drivers seem to be concerned about human drivers taking advantage of autonomous vehicles (Tennant, Howard, Franks, Bauer & Stares, 2016) and autonomous vehicles struggling with the 'informal' rules of driver-driver interactions (Tennant et al., 2015). Hence, we need to determine the parameters of behaviour for autonomous vehicles to be acceptable to their occupants and other road users. In starting to address this issue, one simulator study found that participants preferred a more cautious autonomous driving style than their own (Basu, Yang, Hungerman, Singhal & Dragan, 2017). An on-road trial found that occupants gave higher trust ratings of an autonomous vehicle when driving on an empty road than when overtaking a parked vehicle. Higher ratings were also given when overtaking a parked vehicle with oncoming traffic compared to when there was no oncoming traffic, potentially due to the vehicle slowing down with no obvious justification in the latter case (Venturer Trial 2, 2017). A further simulator study showed that participants preferred an autonomous vehicle to use larger lateral distances and begin steering into the middle lane earlier than the driver's own behaviour when overtaking bicycles and scooters (Abe, Sato & Itoh, 2017). Another simulator trial found that participants were more likely to overtake (move to the fast lane) after two vehicles had passed rather than pull into the gap between them, regardless of whether the second vehicle was autonomous or human-driven, or the distinctiveness of the autonomous vehicle's appearance (TRL PPR807, 2017).

However, neither the Venturer trial (2017) nor Abe et al. (2017) or TRL (2017) studies examined the role of pull-in distance on ratings of autonomous overtaking behaviour, Abe et al. (2017) did not examine trust in the context of an autonomous vehicle overtaking another car, and TRL (2017) focused more on the behaviour of human drivers towards autonomous vehicles rather than the evaluation of their behaviour. In addition, participants in both studies

were asked to rate their general level of trust in the vehicle, rather than evaluate a specific vehicle behaviour or manoeuvre, and Abe et al. (2017) did not manipulate individual parameters independently. However, we need data on what makes specific manoeuvres more or less acceptable if autonomous vehicles / systems are to be programmed optimally. Thus, the present study focused on peoples' ratings of acceptability as a number of relevant parameters of a specific manoeuvre were systematically varied. Autonomous overtaking was selected as the manoeuvre of interest for several reasons. First, being at an unsafe distance (Chapter 3) or constantly being overtaken (Chapter 4) are examples of very stressful and frustrating real-world driving scenarios. Second, despite this, overtaking has not received much attention in terms of trust/acceptability of autonomous vehicles. Third, use cases involving overtaking scenarios were identified as a key area of interest following discussions with the project's industry partner. Participants were asked to evaluate overtaking behaviours as a function of: i) pull-in distance after the overtake, ii) occupant perspective: driving a vehicle that is being overtaken or being in an autonomous vehicle that is overtaking another vehicle, and iii) the effect of traffic context; whether ratings of an overtaking manoeuvre are influenced if the overtaking vehicle is being followed by a third vehicle. Acceptability was determined via a 15-item 'Vehicle Character Questionnaire' (VCQ) (Table 3.1).

Table 3.1. Full Vehicle Character Questionnaire (VCQ).

Question number	Question text
1	With 1 being “unpleasant”, and 7 being “pleasant”, please rate the driving scenario you have just experienced
2	With 1 being “incompetent”, and 7 being “competent”, please rate the behaviour of [the overtaking vehicle/your vehicle] in the driving scenario you have just experienced
3	With 1 being “erratic”, and 7 being “predictable”, please rate the behaviour of [the overtaking vehicle/your vehicle] in the driving scenario you have just experienced
4	With 1 being “rude”, and 7 being “polite”, please rate the behaviour of [the overtaking vehicle/your vehicle] in the driving scenario you have just experienced
5	With 1 being “untrustworthy”, and 7 being “trustworthy”, please rate the behaviour of [the overtaking vehicle/your vehicle] in the driving scenario you have just experienced
6	With 1 being “unhelpful”, and 7 being “helpful”, please rate the behaviour of [the overtaking vehicle/your vehicle] in the driving scenario you have just experienced
7	With 1 being “aggressive”, and 7 being “timid”, please rate the behaviour of [the overtaking vehicle/your vehicle] in the driving scenario you have just experienced
8	With 1 being “selfish”, and 7 being “considerate”, please rate the behaviour of [the overtaking vehicle/your vehicle] in the driving scenario you have just experienced
9	With 1 being “unjustified”, and 7 being “justified”, please rate the extent to which [the overtaking vehicle/your vehicle]’s behaviour was justified in the driving scenario you have just experienced
10	With 1 being “uncooperative”, and 7 being “cooperative”, please rate the behaviour of [the overtaking vehicle/your vehicle] in the driving scenario you have just experienced
11	With 1 being “reckless”, and 7 being “safe”, please rate the behaviour of [the overtaking vehicle/your vehicle] in the driving scenario you have just experienced
12	With 1 being “unacceptable”, and 7 being “unacceptable”, please rate the behaviour of [the overtaking vehicle/your vehicle] in the driving scenario you have just experienced
13	Would you be more or less likely to purchase a vehicle that behaves like [the one/your vehicle] in the driving scenario you have just experienced? 1 = highly unlikely, 7 = highly likely
14	Would you be happy to be driven by a vehicle that behaves like [the one/your vehicle] in the driving scenario you have just experienced? 1 = highly unlikely, 7 = highly likely
15	Would you use a system in your own car that behaved like [the vehicle/your vehicle] in the driving scenario you have just experienced? 1 = highly unlikely, 7 = highly likely

VCQ scores ranged from 1-7, with lower scores being more negative and higher scores being more positive.

This chapter presents six experiments; three simulator-based (1a-c) and three video-based (2a-c) replications which examined the potential influence of participant immersion on the results. Electrodermal activity (EDA, E1a-c) heart rate (E1b-c), and driver input data (E1a-c) were also measured. These objective measures were included to investigate, for example, if attention was captured and potentially if someone was surprised, angered or threatened by a manoeuvre (Boucsein, 2012; Critchley, 2002; Dewe, Watson & Braithwaite, 2016; Frith & Allen, 1983). Table 3.2 provides an overview of the research questions and key findings of all six experiments.

Table 3.2. Overview and summary of the methodologies, key variables, and central findings of all six experiments presented in Chapter 3.

Experiment	Methodology	Key independent variables	Key results
1a	Simulator	Pull-in distance, vehicle perspective	Sharp increase in trust with increasing pull-in distance up to ~28m; no effect of vehicle perspective
1b	Simulator	Pull-in distance, following distance/traffic context	Sharp increase in trust up to ~28m; no effect of presence of a following vehicle
1c	Simulator	Pull-in distance, following distance/traffic context	Sharp increase in trust up to ~28m; drivers more forgiving of sharper pull-ins with a closer following third vehicle
2a	Video	Pull-in distance, vehicle perspective, immersion level	Sharp increase in trust up to ~28m; drivers more forgiving of sharper pull-ins on video compared to when in the simulator
2b	Video	Pull-in distance, following distance/traffic context, immersion level	Sharp increase in trust up to ~28m; drivers more forgiving of sharper pull-ins on video compared to when in the simulator
2c	Video	Pull-in distance, following distance/traffic context, immersion level	Sharp increase in trust up to ~28m; no difference in pattern of ratings when compared to simulator

3.2. Experiments 1a – 1c: Evaluation of autonomous overtaking as a function of pull-in distance, perspective and following distance of a third vehicle

Experiment 1a examined the acceptability of overtaking manoeuvres as a function of pull-in distance from two perspectives (being overtaken while driving vs. overtaking as a passenger in an autonomous vehicle). In the *being overtaken condition*, participants drove in the left lane of a three-lane motorway and a second vehicle overtook in the middle lane and then and pulled back into the driver's lane. The distance between the two vehicles when the pull-in manoeuvre commenced varied between 1m to 88m. In the *overtaking condition*, participants were placed in an autonomous vehicle which overtook another vehicle and then pulled back into the left lane at different pull-in distances. In Experiments 1b and 1c, participants were presented with the same *being overtaken condition* except that a third vehicle was present which followed the overtaking vehicle at either a relatively safe following distance of 2s (Experiment 1b) or at a relatively unsafe distance of 0.5s (Experiment 1c).

3.2.1. *Method*

Participants

There were 20 participants in each Experiment (Table 3.3). The number of participants recruited was comparable to similar work on autonomous driving styles (e.g. Abe et al., 2017; Basu et al., 2017; Gouy, Wiedemann, Stevens, Brunett & Reed, 2014; Nowakowski, O'Connell, Shladover & Cody, 2010), and initial pilot work also showed a strong effect of pull-in distance with five participants. Each held a driving licence and was screened for motion sickness, migraine, photosensitivity, heart conditions, vertigo, postural instability, and current pregnancy and completed the Simulator Sickness Questionnaire (Kennedy, Lane, Berbaum & Lilienthal, 1993). Each was paid £5 for their participation.

Table 3.3. Participant details for all six experiments in Chapter 3 (F = Female, RH = right-handed).

Experiment	Number of participants (number females)	Handedness	Age range, mean and standard deviation	Participants with driving license	Driving license duration and type
1a	20 (12F)	16 RH	21 – 37 years M=25.15 SD=4.34	20	10 months – 19 years 13 UK
1b	20 (10F)	18 RH	18 – 35 years M=21.65 SD=3.4	20	6 months – 12 years 9 UK
1c	20 (11F)	18 RH	18 – 28 years M=20.85 SD=2.20	20	4 months – 7 years 8 UK
2a	30(20F)	25 RH	20 – 61 years M=25.76 SD=7.76	20	1 – 27 years 4 UK
2b	30(18F)	28 RH	18 – 36 years M=22.34 SD=4.30	22	1 – 15 years 2 UK
2c	30(23F)	26 RH	18 – 50 years M=23.03 SD=6.10	20	6 months – 20 years 7 UK

Driving simulator setup

The simulator consisted of a fixed-base setup, with the front half of Jaguar XJ 2009 used for the cabin (*Figure 3.1*) running SCANeR Studio 1.4 software. The simulated vehicle was right-hand drive and all roads used were UK-based, with the traffic flow on the left-hand side. There were three projection screens in front of the vehicle (SXGA+ resolution, ~135° horizontal visual angle) and three rear screens reflected in the left, rear and right mirrors. Sound was provided by a 5.1ch surround system. Driver input was via the vehicle's original pedals and steering wheel (with force feedback). Additional tactile feedback was provided by a shaker located under the driver's seat (this was a feature of the driving simulator intended to increase immersion, rather than manipulated as part of the experimental design). A dashboard LCD panel displayed vehicle speed (mph), engine RPM and current gear. The vehicle used an automatic transmission with no gear changing or turn signal input required from the participant. An additional LCD panel located within the centre console was used to administer the questionnaire measures. The environment consisted of a 9.02km stretch of UK motorway, with three lanes and hard shoulder on the left, the equivalent lanes in the opposite direction and a 2m central reservation in between with road lamps every 40m.

(A)

(B)

Figure 3.1. Participants views of the overtaking scenario in the University of Warwick Psychology/WMG fixed-base driving simulator, showing the inside of the cabin, mirrors, steering wheel, and visual panels. Panel (A) shows the perspective of a driver being overtaken. Panel (B) shows the perspective of an occupant of an autonomous vehicle performing an overtake.

Physiological measurements

Electrodermal activity (EDA) was obtained via disposable EL507 electrodes attached to the distal phalanges of the fingers of the non-dominant hand. Initial pilot work indicated that this arrangement provided reliable measurements and was not intrusive to the driving experience. Experiments 1b and 1c also recorded heart rate via Lead II ECG, with disposable EL501 electrodes attached to the participant's right wrist and left ankle. Measurements were recorded using a Biopac MP36R data acquisition unit and analysed with Acqknowledge v4.1. ECG and EDA signals were sampled at 1000Hz. EDA measurements used a gain of x2000, with low pass filters at 66.5 and 38.5Hz and band stop line frequency filter at 50Hz. ECG measurements used the same filters but with a high-pass filter at 0.5Hz and a gain of x1000. Relevant simulator events were automatically coded in the physiological data file.

Simulator driver input

The stimulator experiments examined gas pedal position (0-1), brake pedal force (Newtons), steering wheel angle (radians), lane gap (distance from the centre of the current lane in metres) and speed (mph) over a four-second window from: i) one second before the start of the pull-in to three seconds after the pull-in, and ii) one second before to three seconds after the overtaking vehicle passed the participants' vehicle. All data were from the 'being overtaken' conditions.

Design and procedure

Experiment 1a used a fully within-subjects design. Participants first completed two acclimatization trials (one driving the vehicle, one in an autonomous mode vehicle). They then completed 18 test trials which were combinations of pull-in distance[1] (1m, 3m, 8m, 13m, 18m, 28m, 48m, 68m and 88m – selected using initial piloting) and perspective (being in the autonomous overtaking vehicle vs. being in a non-autonomous vehicle that was being overtaken). *Figure 3.2* provides a top-down schematic of a single trial. Trial order was randomized, and participants were given a break outside of the simulator lab after the first 8 test trials. All 18 trials were randomized (such that participants switched perspective more than once) in order to prevent participants losing interest in the task (e.g., after prolonged periods of automated driving) and to reduce order effects especially in respect to potential physiological habitation. After each trial, participants were presented with the VCQ with each question presented sequentially on a touch-screen LCD panel located in the centre dashboard

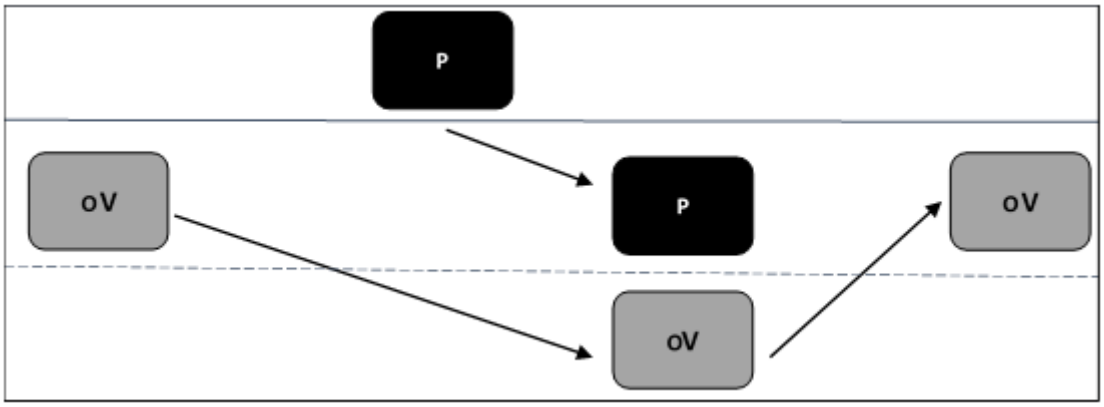
[1] Measurements correspond to the distance between the front bumper of the lead vehicle and the rear bumper of the overtaking vehicle at the commencement of the pull-in to the left lane.

console and participants responded by touching an option on a 7-point Likert scale[2]. For each trial, the average of the scores across all questions was calculated to give a VCQ rating that ranged between 1 and 7. The VCQ scale showed good internal consistency (Cronbach's α = .978 for Experiment 1a; .986 for both 1b & 1c), suggesting that the individual items were measuring similar things, and providing support for using the mean VCQ rating as a measure of vehicle character evaluations. In the *being overtaken condition*, participants drove at a specified speed (60mph) in the left lane after which they were overtaken by an autonomous vehicle (starting at 60mph before accelerating to 70mph) which pulled in at varying distances. In the *overtaking condition*, participants were placed within an autonomous vehicle (starting at 60mph before accelerating to 70mph) which pulled out to overtake a vehicle in the left-hand lane (travelling at 60mph) and pulled back into the left lane at varying distances. After completing the final test trial, participants were screened for simulator sickness, debriefed, and thanked. The experiment lasted 60 to 90 min.

In Experiment 1b and 1c, participants experienced the same *being overtaken condition* from Experiment 1a, except that a third vehicle followed behind the overtaking vehicle. In Experiment 1b, the third vehicle followed 2 seconds (64.69m) behind and in Experiment 1c it followed 0.5 seconds (16.17m) behind. These were selected based on guidelines recommending 2 seconds as a safe following distance ("The Highway Code", n.d.). On each trial, the third vehicle remained at a fixed distance behind the overtaking vehicle, pulled out to the middle lane at the same time, was yoked to the overtaking vehicle's speed, and remained in the middle lane after the pull-in. Each trial stopped ten seconds after the following vehicle passed the participant, or ten seconds after the overtaking vehicle was completely in the left lane after the pull-in, whichever came last.

[2] Each point of the 7-point scale was represented by a grey dot with a white number inside. When participants pressed a dot to respond, the dot turned red to indicate that their response had been received. When the next question was presented, the dot participants had pressed in the previous question remained red until they responded to the next question. This meant that their response to the previous question was visible in the same way as a previous response on a paper-based questionnaire would be. It is unlikely that this had any influence on the results, but this methodological detail has been included for completeness.

(A)



(B)

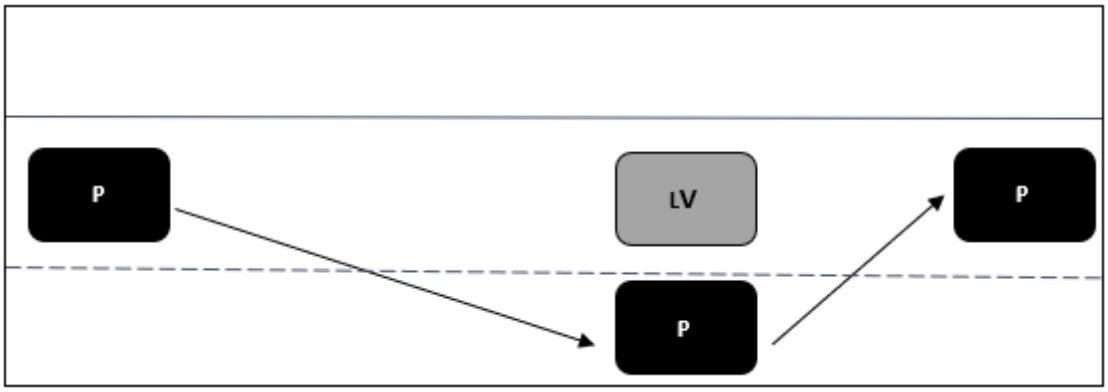


Figure 3.2. Overview of a single trial from Experiment 1a in Chapter 3. (A) In each being overtaken condition trial, the participant (P) moved from the hard shoulder to the left lane and accelerated to 96 km/h while being followed at a distance of 50m by the overtaking vehicle (OV). Three seconds after the participant exceeded 90 km/h, the overtaking vehicle accelerated to 16 km/h above the participant's speed and moved to the middle lane, pulling back into the left lane at the set pull-in distance. (B) In each *overtaking condition* trial, the participant's vehicle (P) accelerated from zero to 96 km/h over ten seconds, followed the lead vehicle (LV) at a distance of 50m for ten seconds, moved to the middle lane and accelerated to 112 km/h, then pulled back into the left lane at the set pull-in distance. For both perspectives, the trial was stopped ten seconds after the overtaking vehicle was completely in the left lane.

3.2.2. Results

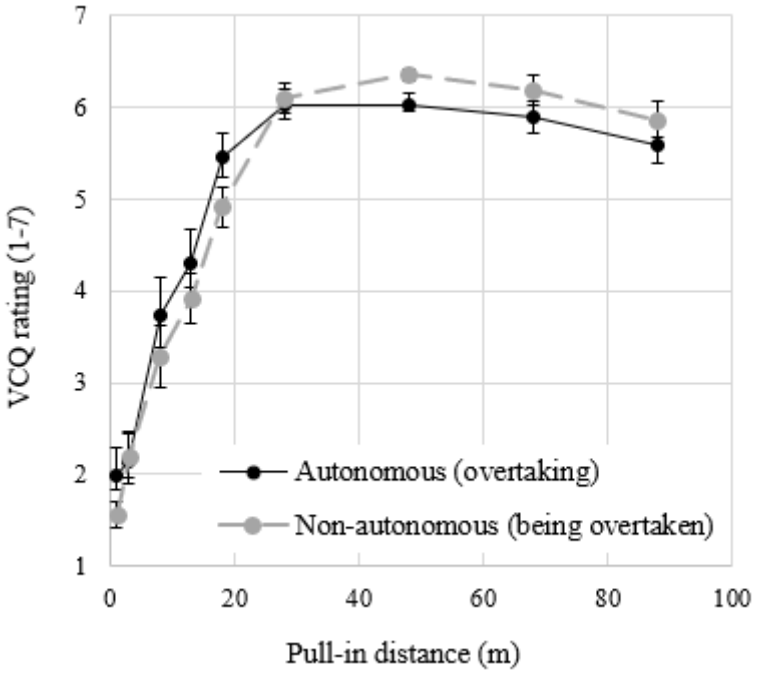
Vehicle character ratings

Experiment 1a: All of the individual question items (each ranging from 1 – 7, all forward-coded) were averaged to compile a final score of between 1 and 7. Mean VCQ ratings (*Figure 3.3A*) increased as a function of pull-in distance up to approx. 28m, plateaued and decreased slightly after approx. 48m. The effects of Vehicle perspective (*being overtaken* vs. *overtaking*) and Pull-in distance (1m to 88m) on VCQ ratings were analysed with a 2 × 9 within-subjects ANOVA (Greenhouse-Geisser corrections applied where necessary). There was a significant main effect of pull-in distance, F (4.367, 82.968) = 102.989, $p < .001$, $\eta^2 = .844$. However, neither the main effect of perspective, F (1, 19) = 3.12, $p = .583$, $\eta^2 = .016$, nor the perspective × pull-in distance interaction reached significance, F (4.279, 81.306) = 1.985, $p = .100$, $\eta^2 = .095$.

Experiment 1b: Mean VCQ ratings for Experiment 1b (*Figure 3.3B*) were analysed with a one-way repeated-measures ANOVA which revealed a significant main effect of pull-in distance, F (4.187, 79.550) = 58.47, $p < .001$, $\eta^2 = .755$. To assess the influence of having a third vehicle follow the overtaking vehicle, the VCQ ratings from the *being overtaken condition* of Experiment 1a were compared with those of Experiment 1b with a mixed 2 (Experiment: 1a, 1b) × 9 (Pull-in-distance) ANOVA with Experiment as the between-subjects factor. This revealed a significant main effect of pull-in distance F (5.201, 197.642) = 145.556, $p < .001$, $\eta^2 = .789$. However, neither the main effect of experiment, F (1, 38) = .006, $p = .940$, $\eta^2 = .000$, nor the experiment × pull-in distance approached significance, F (5.201, 197.642) = 0.849, $p = 0.521$, $\eta^2 = .005$.

Experiment 1c: Mean VCQ ratings for Experiment 1c are shown in *Figure 3.3B.* A one-way repeated-measures ANOVA revealed a significant main effect of Pull-in distance, F (4.025, 76.468) = 34.33, $p < .001$, $\eta^2 = .644$. To determine if evaluations were influenced by the closeness (2s vs. 0.5s) of a vehicle that followed the vehicle doing the overtaking, VCQ ratings for Experiments 1b and 1c were compared using a 2 (Experiment: 1b, 1c) × 9 (Pull-in distance) mixed ANOVA with Experiment as the between-subjects factor. There was a main effect of pull-in distance F (5.086, 193.273) = 87.530, $p < .001$, $\eta^2 = .683$ and a significant pull-in distance × experiment interaction, F (5.086, 193.273) = 2.602, $p = .026$, $\eta^2 = .020$. This mostly likely indicates that the shorter pull-in distances were rated more positively in the 0.5s following condition than in the 2s following condition. The main effect of experiment did not approach significance, F (1, 38) = .389, $p = .537$, $\eta^2 = .010$.

A) Experiment 1a: Effect of perspective and pull-in distance



B) Experiment 1b and 1c: Effect of following and pull-in distance

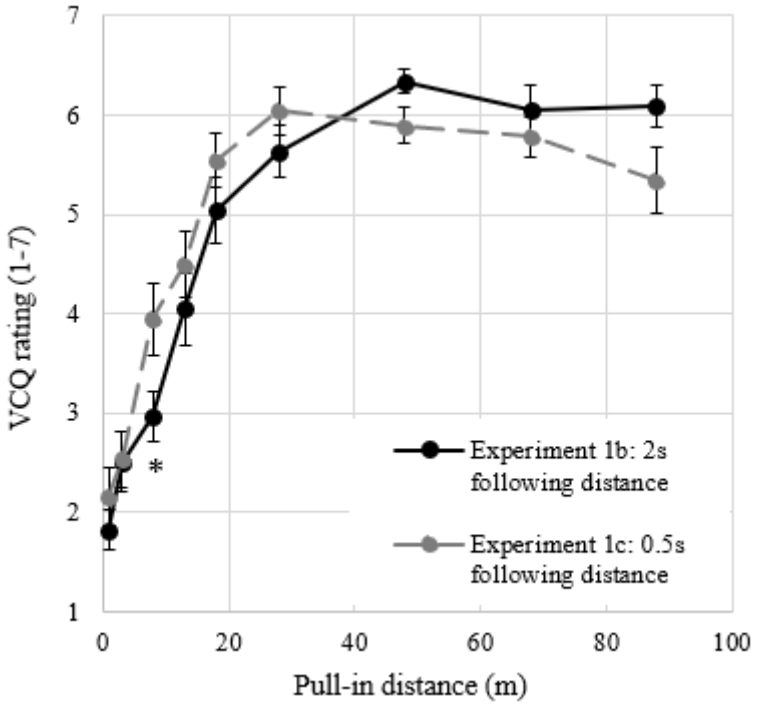


Figure 3.3. Mean VCQ ratings as a function of pull-in distance and perspective in Experiment 1a (Panel A) and ratings as a function of pull-in distance and following distance for Experiment 1b and 1c (Panel B). Error bars correspond to ±1SE. . Asterisks below data points (Panel B) indicate a significant difference between the same pull-in distance across Experiment 1b and 1c ($p < .05$, two-tailed).

Electrodermal activity

The probability of a skin conductance response (SCR) when the overtaking vehicle pulled in was determined with Acqknowledge 4.1 using an automated analysis routine. The minimum SCR size was set to 0.02μs, only SCRs that began 0.5 to 7 seconds after the overtaking vehicle crossed the centre line to enter the left lane were counted, SCRs had to start and end within the driving scenario, and motion artefacts were excluded via a visual inspection of the EDA signal. Mean probabilities are shown in *Figure 3.4.*

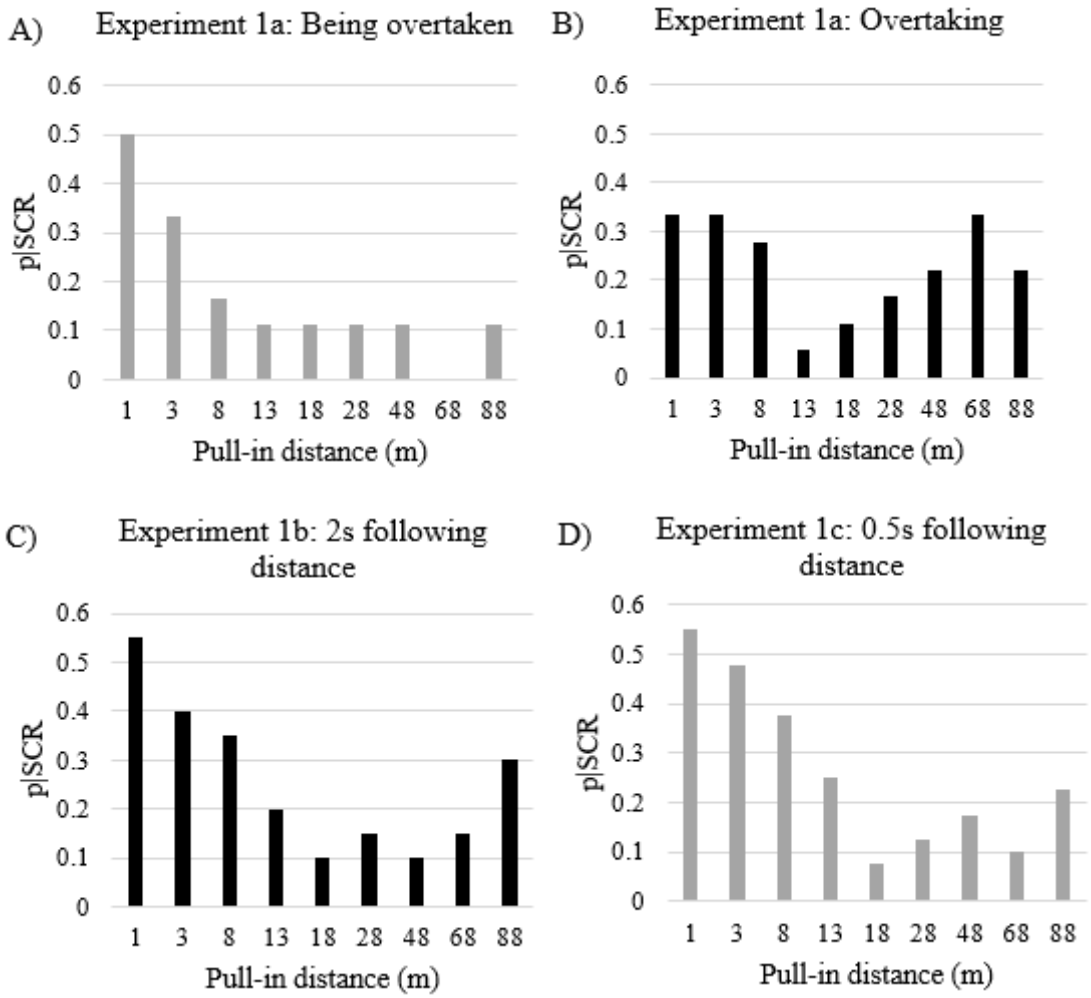


Figure 3.4. Probability of an SCR in response to overtakes as a function of pull-in distance in Experiment 1a (Panel A: Overtaking; Panel B: Being overtaken), Experiment 1b (Panel C) and Experiment 1c (Panel D).

Experiment 1a: Data from 18 of participants from Experiment 1a were analysed (two participants were excluded due to very little deflection in their EDA signal). The analyses for Experiments 1b and 1c included all 20 participants. Cochran's Q test revealed a significant effect of Pull-in distance on SCR probability for the *being overtaken condition*, χ^2 (8) = 23.820, p = .002, however, there was no reliable effect of pull-in distance in the *overtaking condition*, χ^2 (8) = 9.132, p = .33. Exact McNemar tests for the *being overtaken condition* revealed that SCR probability was higher in the 1m condition than all other conditions between 8 and 28m, as well as the 88m (p < .05) and 68m conditions (p < .01). SCR probability was significantly higher in the 3m condition than the 68m condition (p < .05).

Experiment 1b: A Cochran's Q test revealed a significant effect of pull-in distance on SCR probability, χ^2 (8) = 25.630, p < .001. Exact McNemar tests showed that the probability of an SCR was significantly higher in the 1m condition compared to the 13m condition (p < .05) and all other conditions between 18 and 68m (p < .01). The probability of an SCR for a 3m pull-in was significantly higher than for the 18m pull-in (p < .05).

Experiment 1c: A Cochran's Q test revealed a significant main effect of pull-in distance on SCR probability, χ^2 (8) = 33.6, p < .001. Exact McNemar tests showed that SCR probability was significantly higher in the 1m condition than the 18 and 68m conditions (p < .01) and the 28 and 88m conditions (p < .05). SCR probability was significantly higher in the 3m condition than in the 18 and 68m conditions (p < .01) and the 28 and 88m conditions (p < .05). The probability of an SCR was also significantly higher in the 8m condition than in the 18, 28 and 68m conditions (p < .05).

EDA Frequency and Heart Rate Activity

Experiments 1b and 1c included the rate of SCRs per minute and heart rate across the entire overtaking manoeuvre for each pull-in distance. The overall mean number of SCRs per minute was 2.84 (*SD* = 2.81). A 2 (Experiment: 1b, 1c) × 9 (Pull-in distance) revealed no significant main effect of Experiment, F (1, 37) = 1.018, p = .319, η^2 = .027, no main effect of Pull-in distance, F (5.647, 208.934) = 1.552, p = .167, η^2 = .040, or their interaction, F (5.647, 208.934) = .474, p = .817, η^2 = .012. For heart rate (mean = 77.45 bpm, *SD* = 12.16), there was no main effect of Experiment, F (1, 37) = .030, p = .864, η^2 = .001, no main effect of Pull-in distance, F (4.651, 172.105) = .915, p = .467, η^2 = .024, and no interaction, F (4.651, 172.105) = .493, p = .769, η^2 = .013.

Simulator driver input

For the three simulator experiments I examined changes in five dependent variables (gas [accelerator input], brake input, steering input, lane [lateral position of vehicle within

lane and speed]) over a 4s window (binned into nine 0.5s intervals) for each of the nine pull-in distances using a series of one-way ANOVAs. Two windows were examined ranging from, -1 to 3s around: i) the overtaking vehicle passing, and ii) the overtaking vehicle pulling-in. Benjamini-Hochberg corrections (Benjamini & Hochberg, 1995) were applied for each resulting set of nine ANOVAs.

Experiment 1a: There was a significant change in lane gap over time when the overtaking vehicle passed the participant, such that participants drifted towards the left of the lane in the 28m condition, F (1.84, 34.88) = 7.09, $p < .05$. There were no other significant effects.

Experiment 1b: There was a significant change in lane gap over time when the overtaking vehicle passed the participant, such that participants drifted to the right of the lane in the 48m condition, F (1.37, 26.03) = 7.92, $p < .05$. There was a significant decrease in gas pedal input F (2.11, 40.17) = 6.55, $p < .05$, and in speed, F (1.03, 19.60) = 8.07, $p < .05$, in the 1m condition after the pull-in. There was a significant increase in speed after the pull-in in the 68m condition, F (1.06, 20.15) = 8.33, $p < .05$. No other effects were significant.

Experiment 1c: No significant effects were found for any of the behavioural variables.

Experiments 1a-c investigated the influence of pull-in distance, vehicle perspective and traffic context on perceptions of autonomous overtaking manoeuvres. Overall, the findings showed a strong effect of pull-in distance on evaluations of the overtakes, with the sharpest pull-ins being evaluated very negatively and producing a physiological orienting response. However, there was some sensitivity to traffic context: A third vehicle following the overtaking vehicle very closely was associated with more positive evaluations of being overtaken at sharper distances, and more negative evaluations of being overtaken at longer distances. As there is significant variation in the immersion level of simulator setups, and psychological experiments are often run with fairly abstract methodologies presented on computer screens, the following three experiments tested whether the results of Experiments 1a-c held up under a lower level of immersion.

3.3. Experiments 2a-2c: Evaluation of overtaking using a video-based methodology

Experiments 2a-c replicated Experiments 1a-c using a video-based methodology. This allowed the video-based experiments to: i) test the robustness of the findings and provide a replication of the basic effects, and ii) examine the potential influence of the level of immersion (presumed to be higher in the driving simulator) in terms of perceiving and

reporting vehicle character. Investigating potential effects of immersion level was important because many simulator experiments (as well as psychology experiments in general) use lower-immersion setups, where participants are presented with stimuli via a computer screen. If the same preferences/effects do not generalise from the simulator to less immersive methodologies, this could have implications for the appropriateness of these methodologies for investigating evaluations of autonomous driving styles and manoeuvres.

3.3.1. *Method*

Participants

Thirty participants completed each experiment (Table 3.3) and were paid £3. The number of participants was increased slightly from Experiments 1a-c, due to both the expectation that a lower-immersion setup may produce smaller effects of pull-in distance etc., and also the ability to test larger groups of participants in a single session, compared to the simulator. Participants were tested in groups of up to 24 in a large computer laboratory in sessions lasting 30-45m.

Measures and apparatus

Video stimuli: Driver's eye videos of the driving scenarios from Experiments 1a-c were recorded at 1080p resolution and 60 fps and presented on 57 × 35cm LCD screens. A custom HTML script presented the videos, questionnaire measures and recorded responses. All videos were between 24-44s in length, encompassing 4s before the overtaking vehicle pulled into the middle lane, and 4s after the overtaking vehicle pulled into the left lane.

Design and procedure

In Experiment 2a participants saw 18 videos (2 levels of perspective × 9 levels of pull-in distance), in a random order and provided VCQ ratings after each video. Similar to Experiments 1a-c, the VCQ showed good internal consistency, providing support for using a single score (Cronbach's $\alpha = .988$ for Experiment 2a, .981 for 2b). Cronbach's α was not calculated for 2c, due to issues accessing the data caused by the Covid-19 lockdown situation. Experiment 2b and 2c were similar except that participants viewed and rated 9 videos rather than 18.

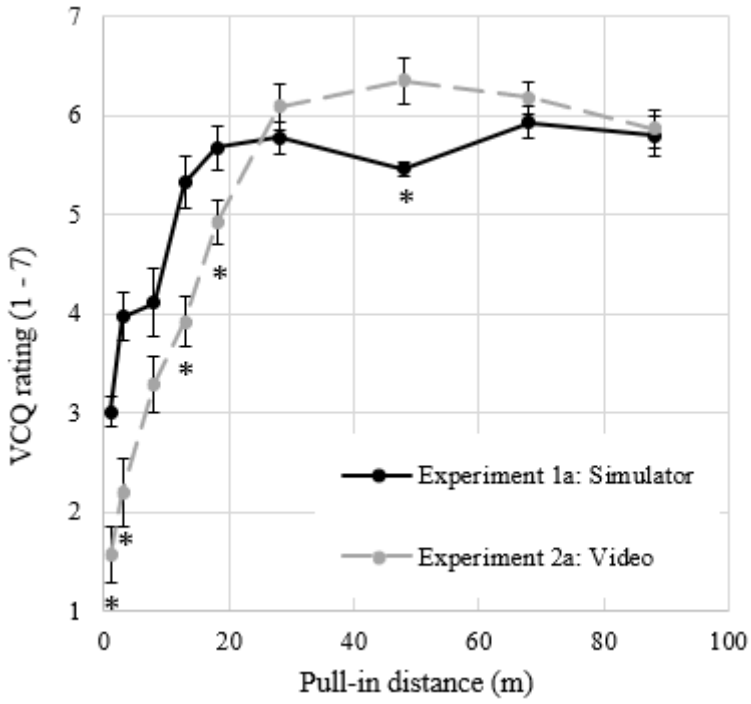
3.3.2. Results

Experiment 2a: Pull-in distance and Immersion

The VCQ data from Experiment 1a were combined with those from Experiment 2a and the resulting set analysed with a 2 (Immersion: video vs. simulator) × 2 (Vehicle perspective: overtaking vs. being overtaken) × 9 (Pull-in distance) mixed ANOVA.

Immersion was the between-subject factor. This revealed a main effect of pull-in distance, F (3.485, 167.284) = 114.111, $p < .001$, $\eta^2 = .677$. There was also a pull-in distance × immersion interaction, F (3.485, 167.284) = 6.526, $p < .001$, $\eta^2 = .039$, which was qualified by an immersion × pull-in distance × vehicle perspective 3-way interaction, F (5.123, 245.896) = 2.851, $p < .05$, $\eta^2 = .055$. Ratings in the video condition were marginally higher (more positive) than in the simulator conditions, F (1, 48) = 3.632, $p = .063$, $\eta^2 = .070$. The main effect of vehicle perspective, F (1, 48) = .632, $p = .430$, $\eta^2 = .012$, and the remaining, vehicle perspective × immersion, F (1, 48) = 2.898, $p = .095$, $\eta^2 = .056$, and pull-in distance × vehicle perspective interactions, F (5.123, 245.896) = 1.103, $p = .360$, $\eta^2 = .021$, were non-significant. As shown in *Figure 3.5*, in the *overtaking condition*, VCQ ratings increased with distance in a similar way for the simulator and video-based conditions. However, for the *being overtaken* condition, VCQ ratings were more positive at the shorter pull-in distances and plateaued at a lower overall level with the video-based methodology than with the simulator. Confirming this, the data were split by perspective and two separate mixed ANOVAs were conducted with the factors of Immersion and Pull-in distance. For the *overtaking* condition there was a main effect of pull-in distance, F (3.714, 178.280) = 72.929, $p < .001$, $\eta^2 = .593$, however, neither the main effect of immersion, F (1, 48) = .905, $p = .346$, $\eta^2 = .019$, nor the immersion × pull-in interaction was significant, F (3.714, 178.280) = 2.104, $p = .087$, $\eta^2 = .017$. For the *being overtaken* data, ratings were overall more positive in the video-based condition, F (1, 48) = 6.520, $p < .05$, $\eta^2 = .120$, and increased as pull-in distance increased, F (4.604, 220.978) = 83.814, $p < .001$, $\eta^2 = .594$. There was also a significant immersion × pull-in distance interaction, F (4.604, 220.978) = 9.258, $p < .001$, $\eta^2 = .066$. The overall pattern was that shorter pull-in distances were rated more positively in the video-based methodology compared to the simulator, with some longer distances being rated more negatively in the video-based version. Ratings in the video-based methodology were significantly more positive at 1m, 3m, 13m (ps < .001), 8m ($p < .01$) and 18m ($p < .05$) compared to the simulator-based methodology. Ratings in the video-based methodology were significantly more negative than those in the simulator-based methodology at 48m, $p < .01$.

A) Experiment 1a and 2a: Effect of immersion and pull-in distance for being overtaken



B) Experiment 1a and 2a: Effects of immersion and pull-in distance for overtaking

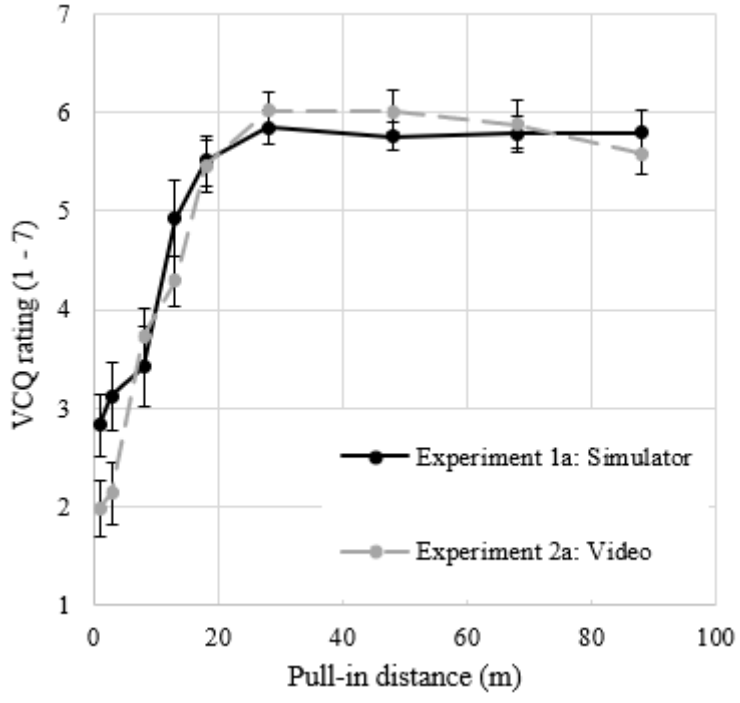
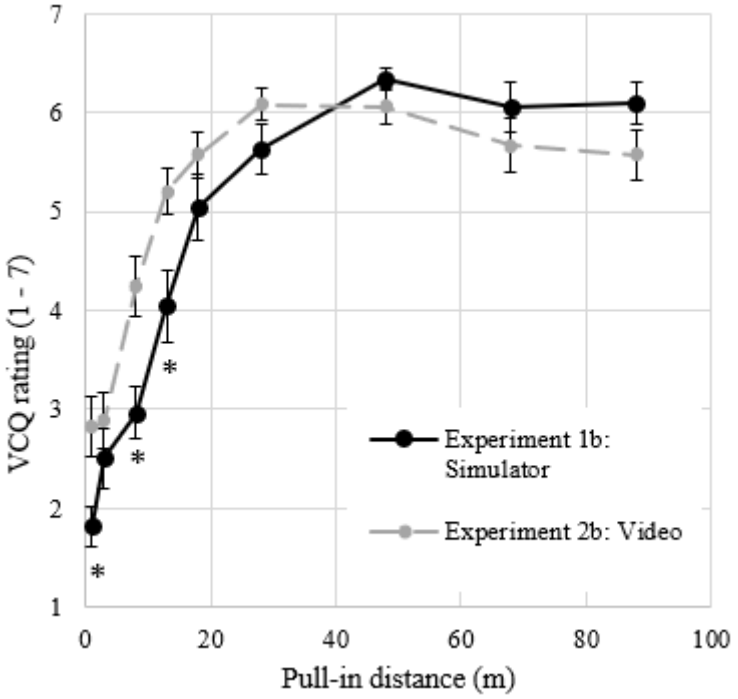


Figure 3.5. Influence of pull-in distance and immersion for Experiments 1a and 2a, for the being overtaken (Panel A) and overtaking (Panel B) conditions. Error bars correspond to ±1SE. Asterisks below data points (Panel B) indicate significant differences between the same pull-in distances across as a function of immersion ($p < .05$, two-tailed).

Experiment 2b: Being overtaken with a third vehicle following the overtaking vehicle with a gap of 2 seconds.

As in Experiment 2a, the VCQ scores from Experiment 1b and 2b were combined. A 2 (Immersion: video vs. simulator) × 9 (Pull-in distance) mixed ANOVA revealed a main effect of pull-in distance, F (4.513, 126.610) = 89.473, $p < .001$, $\eta^2 = .629$, and an immersion × pull-in distance interaction, F (4.513, 216.610) = 4.697, $p < .001$, $\eta^2 = .033$. The main effect of immersion was not significant, F (1, 48) = 2.957, $p = .092$, $\eta^2 = .058$. As shown in *Figure 3.6,* shorter pull-in distances were rated more positively and larger pull-in distances more negatively using the less immersive video-based approach compared to data collected in the driving simulator.

A) Experiment 1b and 2b: Effect of immersion and pull-in distance



B) Experiment 1c and 2c: Effect of immersion and pull-in distance

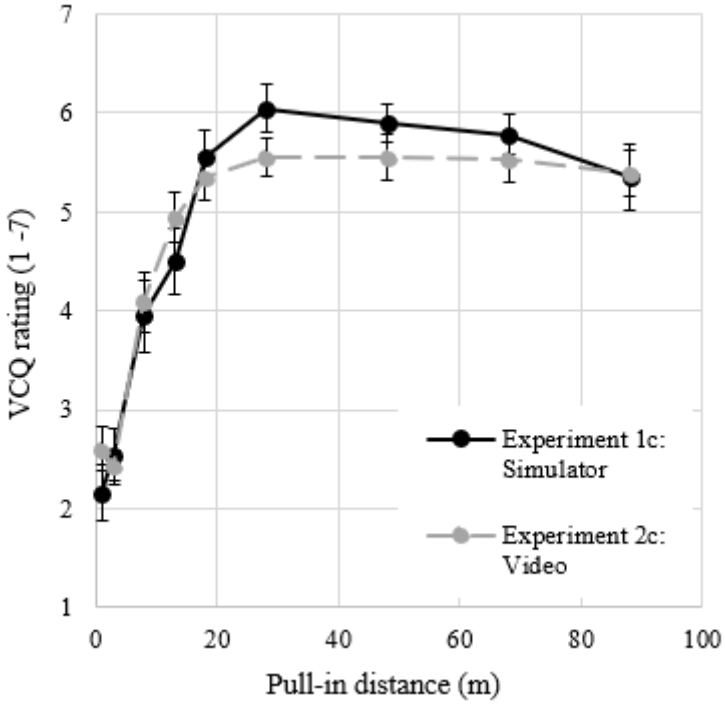


Figure 3.6. Influence of pull-in distance and immersion for Experiment 1b & 2b (Panel A) and Experiment 1c and 2c (Panel B). Error bars correspond to ±1SE. Asterisks below data points (Panel A) indicate significant differences between the same pull-in distances as a function of immersion. ($p < .05$, two-tailed).

Experiment 2c: Being overtaken with a third vehicle following the overtaking vehicle with a gap of 0.5 seconds.

The VCQ scores from Experiment 1c and 2c were combined and analysed using a 2 (Immersion: video vs. simulator) × 9 (Pull-in distance) mixed ANOVA with immersion as the between-subjects factor. This revealed a main effect of pull-in distance, F (4.523, 217.101) = 71.256, $p < .001$, $\eta^2 = .598$. However, neither the main effect of immersion, F (1, 48) = .031, $p = .862$, $\eta^2 = .001$, nor the immersion × pull-in distance interaction, F (4.523, 217.101) = 1.062, $p = .380$, $\eta^2 = .009$, were significant. The VCQ scores from Experiment 2b and Experiment 2c (*Figure 3.7*) were combined to examine the influence of following distance on the effect of pull-in distance using only the video-based methodology, using a 2 (Experiment: 2b and 2c) × 9 (Pull-in distance) mixed ANOVA with Experiment as the between-subjects factor. This revealed a main effect of pull-in distance, F (4.317, 250.403) = 77.068, $p < .001$, $\eta^2 = .569$, but neither the main effect of experiment, F (1, 58) = 2.11, p = .152, $\eta^2 = .035$, nor the experiment × pull-in distance interaction, F (4.317, 250.403) = .286, p = .899, $\eta^2 = .002$, were significant.

Experiment 2b and 2c: Effect of following and pull-in distance

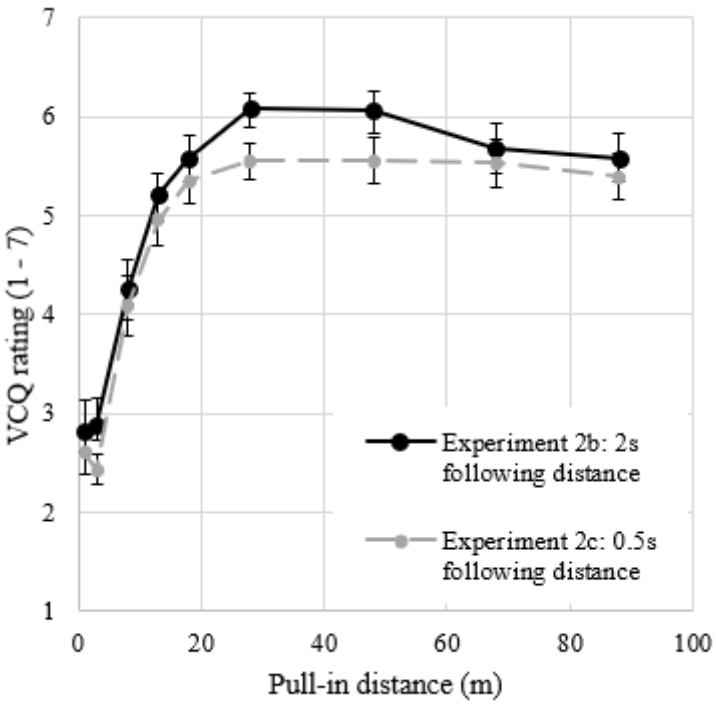


Figure 3.7. Influence of following (2s vs 0.5s) and pull-in distance on VCQ ratings in the low immersion video-based methodology (Experiment 2b and 2c). Error bars correspond to ±1SE.

3.4. General discussion

All six experiments described in this chapter revealed that vehicle character ratings became more positive as pull-in distance increased, before levelling off and beginning a downward trend after ~48m. The pattern of ratings was similar both when the driver was overtaking or being overtaken (suggesting that, in some driving interactions, the range of acceptable behaviours may be similar between human-driven and autonomous vehicles, as in TRL PPR807, 2017). The results suggest that there is a threshold of approx. 28m for what is accepted to be a 'reasonable' pull-in distance. Of note, most of the pull-in distances used were below the recommended 2 second gap rule used as a guide for car following ("The Highway Code", n.d.). The finding that distances as low as 28m (approximately a 1s gap at 96kmh or 60mph) were rated very positively is consistent with findings in the car following literature (Hutchinson, 2008; Nowakowski et al., 2010; Siebert, Oehl, Bersch & Pfister, 2017; Siebert, Oehl & Pfister, 2014; Taib-Maimon & Shinar, 2001) that distances below the officially recommended following time of 2s are often used and rated as comfortable by drivers. The implication is that the parameters of acceptable autonomous driving styles may diverge from formal guidelines in some contexts.

The finding that drivers that were being overtaken (but not the driver/occupant of an overtaking vehicle) were more likely to show an SCR to a pull-in manoeuvre for the shortest distances provides additional, objective support to the questionnaire results. This physiological orienting response provides an objective measure of the effect of different manoeuvres and likely reflects a combination of surprise, anger, threat and attention capture (Boucsien, 2012; Critchley, 2002; Dewe et al., 2016; Frith & Allen, 1983) by manoeuvres which were subjectively rated as being the most unacceptable. The overall pattern of SCR probabilities as a function of pull-in distance replicated across all three simulator experiments, suggesting that an orienting response to a surprising overtake is a robust finding. Most of the significant differences in SCR probability were between the two smallest pull-in distances and the larger distances. Thus, although SCR probability appears to distinguish very unacceptable, potentially threatening, and surprising pull-ins from more acceptable pull-ins, it does not provide fine-grained distinctions for larger pull-in distances. Furthermore, the maximum SCR probability was 50% in Experiment 1a (being overtaken), and 55% in Experiment 1b and 1c. Therefore, although the physiological data support the questionnaire findings and suggest that the sharpest pull-ins were objectively surprising/threatening, it would not be feasible to use the mere presence of an event-related SCR probability as a primary indicator of whether an autonomous overtake is perceived as

acceptable. Of note, there was little influence of pull-in distance in the *overtaking conditions* suggesting that from a physiological response point of view, overtaking with a short pull-in distance might be much less disruptive and attention capturing than being overtaken in the same way. This difference likely reflects the difference in visual distance information and sense of danger experienced by participants from the two differing perspectives. There was also evidence of a comparatively much smaller potential reduction in ratings at the longer pull-in distances suggesting that excessively long pull-in distances may result in negative vehicle character perceptions. This likely occurred because a driver remaining in an outside lane unnecessarily might be considered to be 'lane hogging'.

The driver input recorded from the driving simulator showed that participants took pressure off of the gas pedal and reduced speed after a sharp pull-in and increased speed after a longer pull-in in one simulator experiment and drifted to the left of the lane after a moderate pull-in in another experiment. However, there were no substantial changes in brake pedal force or steering wheel angle in any of the simulator experiments. There was a general trend for participants to reduce gas pedal force and/or speed over time in response to very sharp pull-ins. Overall, these data provide some, albeit weak, evidence, that participants being overtaken at very small pull-in distances did try to slow down, supporting the finding that those distances were perceived as unsafe.

A second main finding was that traffic context also seems to be important in how autonomous vehicle manoeuvres are evaluated. In Experiment 1b, the overtaking vehicle was followed by a third vehicle, at a relatively safe distance of 2s. In this condition, vehicle character ratings did not differ from when there was no following vehicle. However, when a third vehicle was following at a relatively unsafe distance[3] (0.5s), participants were more forgiving of a shorter pull-in distance than when the following vehicle was at a relatively safer following distance. The more positive evaluation of the 8m pull-in when the following vehicle was 0.5s behind, at least in the simulator experiments, suggests that previously unacceptable vehicle behaviours may be viewed as appropriate in some driving contexts. Participants may have perceived the overtaking vehicle as 'getting out the way' of the following vehicle, and therefore rated the shorter pull-in as more acceptable. However, it is important to note that the very shortest pull-in distances (1 and 3m) were not rated more positively with a closer-following vehicle – so there are some limits to this. Of note, these

[3] Please note however, that distances of approximately 0.5s are commonly used and rated as acceptable (Hutchinson, 2008; Nowakowski, et al., 2010; Siebert et al., 2017; Siebert et al., 2014; Taib-Maimon & Shinar, 2001).

were the distances that were most likely to produce a physiological orienting/threat response. Thus, the presence of a physiological SCR response might be able to be used to indicate behaviours that are not so unacceptable that they are unable to be modified by the context of the behaviour. Furthermore, the effect of following distance on evaluations of pull-in distance was much weaker than the overall effect of pull-in distance on character evaluations. This further suggests that there is a limit to the context-dependency of evaluations of autonomous overtaking, with some behaviours remaining unacceptable irrespective of traffic context, and the overall pattern of ratings over distance remaining consistent. Nonetheless, the overall implication for autonomous driving styles is that behaviour that is unacceptable in one driving scenario may be rated more favourably in another, and autonomous vehicles could therefore benefit from using information on traffic context to perform more acceptable overtakes. Manufacturers may benefit from taking account of possible contextual influences on acceptable behaviour and interactions with other road users when developing autonomous vehicles.

This specific context effect was not, however, replicated in the video-based version of the task – perhaps due to a lower level of immersion. The video-based approach also produced some other differences. When viewed as a video on a relatively small computer screen, participants rated manoeuvres at the small pull-in distance less negatively than participants who experienced the same manoeuvre in the driving simulator. This difference likely reflects the greater perception of direct threat perceived in the more immersive simulator environment compared with that perceived by watching a video of the same situation. Conversely, there was also some evidence that the video-based condition led to more negative evaluations at one of the longer pull-in distances for the *being overtaken* perspective. One possible account of this is that participants might have overestimated the distance between themselves and the overtaking vehicle at the longer pull-in distances. This might occur due to the relatively small size and low resolution of the monitor, perhaps making a distant vehicle appear further away, compared to the much greater area and size of the visuals presented in the driving simulator. It is not clear why there was no reliable difference between the video and simulator-based ratings for Experiment 2c. However, there was a trend for some of the longer distances to produce more negative evaluations in the video condition consistent with the results of Experiment 2a and 2b. Although not the main focus of this study, this suggests that lower levels of immersion can result in non-constant (i.e. not just simply a positive bias) changes to valuations of road manoeuvres. This in turn suggests that some of the driving simulator results, although indicative, may not fully reflect

what would happen in real-world driving conditions given that simulated environments may not be as immersive as their real-world equivalents. Verifying the current findings in real-world, road-based conditions would therefore be a valuable goal for future research.

Importantly, this set of experiments also introduced a novel measure of road users' evaluations of autonomous vehicle behaviour (the Vehicle Character Questionnaire, or VCQ). Across both the simulator and (to a lesser extent) the video-based experiments in this chapter, the VCQ was sensitive to autonomous overtaking behaviours, with substantial effects of pull-in distance on ratings. Additionally, the measure showed a good degree of internal consistency, supporting its use as a single measure of road user evaluations. However, the VCQ consisted of questions on several different aspects of road user perceptions (e.g. perceived safety, purchasing intentions). It would be interesting for future research to explore any potential dissociations between different aspects of vehicle character (e.g. road users might prefer other cars to follow the rules, but may be reluctant to use an overly cautious autonomous vehicle).

3.5. Summary and conclusions

Chapter 3 presented six experiments on the influence of pull-in distance, vehicle perspective, traffic context and immersion level on evaluations of autonomous overtaking manoeuvres. In summary, these experiments found a sharp increase in the acceptability of overtakes with increasing pull-in distance (up to approx. 28m), a pattern which was similar from the perspective of overtaking and being overtaken, with a shallower curve when using lower immersion video-based methodologies. Additionally, some influence of traffic context was found, such that closer pull-ins were rated more positively overall when a third vehicle followed the overtaking vehicle more closely, with longer pull-ins being rated more negatively with a closely following third vehicle.

These findings extend previous work on trust and driving styles by focusing more closely on the evaluation of specific autonomous vehicle manoeuvres (in this case overtaking) and their implications for interactions between human-driven and autonomous vehicles in shared traffic spaces. Chapter 4 extends this work by examining the influence of traffic position (overtaking vs being overtaken), vehicle speed, information presentation and traffic context on occupants' experience of autonomous highway journeys. The work presented in Chapter 4 extends existing work by investigating the interaction between autonomous driving style, contextual factors and information presentation, and in doing so

provides a link between the Vehicle Character (Chapters 3 & 4) and System Character (Chapters 5 & 6) workstreams presented in this thesis.

4. Influence of traffic context and information presentation on occupant perceptions of autonomous highway journeys

Abstract

Previous research into perceptions of autonomous vehicles has largely focused on *a priori* attitudes, with less work on the perception of specific traffic situations, context and driving styles. The present study involved three simulator experiments (total N = 150) to examine the combined effects of vehicle speed, lane position, information presentation and traffic context on occupants' levels of satisfaction with autonomous highway journeys. Overall, occupants preferred being in a vehicle that was mostly overtaking compared to being overtaken, regardless of whether the other vehicles were exceeding the speed limit. This effect remained even when occupants were given additional reminders that they themselves were travelling at an appropriate speed (Experiments 1 & 2). Experiment 3 found that occupants preferred overtaking to being overtaken when following another car, but this preference disappeared when they were following a lorry, suggesting that occupants' sensitivity to position amongst the traffic was partially context dependent. Overall, the findings suggest that the inappropriate behaviour of other drivers (e.g., speeding) can negatively impact journey satisfaction for occupants of autonomous vehicles that follow the rules of the road, depending on the specific traffic situation. Potential implications for the integration of autonomous vehicles with other traffic and the need for in-vehicle presentation of information are discussed.

4.1. Introduction

Autonomous vehicles have been predicted to start arriving on roads as early as 2020 ("Driverless car market watch", n.d.), with the possibility of 75% of all road vehicles being autonomous by 2040, according to one estimate ("Look Ma, No Hands!", 05/09/2012). However, it has been acknowledged that there will likely be a transition period in which autonomous vehicles must drive alongside conventional human-driven vehicles. This transition period will likely require safe and acceptable interactions between autonomous vehicles and human drivers in order for the transition to full autonomy to be successful (Hancock, 2019). Therefore, it is important to understand the psychological factors which influence how human drivers and autonomous vehicles interact, and whether autonomous vehicle behaviour is seen as acceptable from the perspective of the occupant(s) and other drivers. This can be from the perspective of drivers evaluating the behaviour of other vehicles but also from the perspective of occupants evaluating the behaviour of the vehicle in which they are travelling.

Much of the work that has been done on trust of autonomous vehicles has focused on survey-based evaluations, investigating the effects of factors such as demographic variables on *a priori* attitudes towards autonomy. Factors such as gender (Hohenberger, Spörrle & Welpe, 2016; Howard & Dai, 2014; Hulse, Xie & Galea, 2018; Schoettle & Sivak, 2014), age (Bansal & Kockelman, 2018; Haboucha, Ishaq & Shiftan, 2017; Hohenberger et al., 2016; Hulse et al., 2018; Kyriakidis, Happee & de Winter, 2015), personality (Kyriakidis, et al., 2015; Choi & Ji, 2015), cultural differences (Haboucha et al., 2017; Kyriakidis et al., 2015; Schoettle & Sivak, 2014) daily driving behaviours (Howard & Dai, 2014) and experience (Bansal & Kockelman, 2018) have been found to be influential for *a priori* acceptance.

While large-scale surveys are invaluable for investigating a priori attitudes, a complete understanding of trust requires methodologies in which users have experience using autonomous vehicles and systems (Hoff & Bashir, 2015). Recently, driving simulator-based work has examined trust in autonomous vehicles based on features of the vehicle/information provided by in-vehicle interfaces. For instance, trust has been shown to be influenced by factors such as feedback over the vehicle's level of uncertainty (Beller, Heesen & Vollrath, 2013; Helldin, Falkman, Riveiro & Davidsson, 2013), the extent to which feedback is worded as instructive or informative (Cramer, Evers, Kemper & Wielinga, 2008), as well as the level of detailed provided by the automated system's descriptions, and the extent to which the system shares the user's driving goals (Verberne, Ham & Midden, 2012). Other influential factors include the autonomous vehicle's level of anthropomorphism (Lee, Kim, Lee & Shin, 2015; Waytz, Heafner & Epley, 2014), information on the limitations of the system (Körber, Baseler

& Bengler, 2018), amount of experience using an autonomous vehicle (Hartwich, Witzlack, Beggiato & Krems, 2018), and the quality of users' practise with the automation (Payre, Cestac & Delhomme, 2016).

Some previous research has also examined the effect of specific vehicle behaviours and driving styles on trust and journey satisfaction. For instance, one simulator study found that occupants of an autonomous vehicle were happier when the vehicle adopted a driving style that was more cautious than their own (Basu, Yang, Hungerman, Singhal & Dragan, 2017). Similarly, a simulator study by Abe, Sato and Itoh (2017) found that occupants preferred an autonomous vehicle to begin passing manoeuvres sooner and use larger lateral distances, compared to how they would overtake when driving the vehicle themselves. A track study found that occupants' trust in an autonomous vehicle was higher when driving on an empty road compared to when overtaking a parked vehicle (Venturer Trial 2, 2017). However, trust was higher when overtaking a parked vehicle when there was oncoming traffic, compared to when no oncoming traffic was present. This suggests a role of traffic context on whether certain manoeuvres are perceived as trustworthy. A simulator-based study by TRL found that participants were more likely to move to the fast lane after two other vehicles had passed, rather than pull into the gap between them (TRL PPR807, 2017). This was irrespective of whether the second vehicle was human driven vs autonomous and was not influenced by the distinctiveness of the autonomous vehicles' appearance.

One key issue concerning the integration of autonomous vehicles into human-driven traffic is the possibility that other road users may act in an aggressive and unpredictable manner towards rule-based autonomous vehicles (Tennant, Howard, Franks, Bauer & Stares, 2016) which may have difficulties dealing with the implicit conventions of interactions between human drivers (Tennant, 2015). Real-world observational work on autonomous vehicle trials by Madigan et al. (2019) found that, while other road users tended to behave cautiously towards autonomous vehicles overall, road users were more likely to adopt riskier behaviours when the infrastructure did not separate autonomous vehicles from other traffic. These types of behaviours could potentially result in a knock-on effect of reduced satisfaction for the occupants of autonomous vehicles, even when the vehicle itself is behaving appropriately. In terms of highway journeys, recent data shows that almost half of all drivers on UK highways exceed the National Speed Limit (National Office of Statistics, Table SPE0111: Free flow vehicle speeds by road type and vehicle type in Great Britain, 2018). Accordingly, if we assume that autonomous vehicles are likely to adopt safe driving styles and follow the rules of the road, this could potentially lead to autonomous vehicles being overtaken frequently, especially in

highway environments. This could lead to frustrating highway journeys for autonomous vehicle occupants, during the transition period to full autonomy. Therefore, it is important to understand what factors lead to reduced occupant satisfaction in such journeys, and to explore what might be done to reduce any frustration caused by the unpredictable behaviour of other road users.

Overall, previous research has shown that features of the vehicle, they type of information displayed to the occupant, traffic context and various parameters of driving style appear to be important in determining the trust of autonomous vehicles and how they are perceived. Despite concerns that unsafe and/or illegal driving behaviour from human drivers could put autonomous vehicles at a disadvantage (and potentially reduce occupant satisfaction), there has been little work directly testing the impact of human drivers' behaviour on the experience of autonomous vehicle occupants. In addition, factors relating to traffic context, information presentation and driving style have largely been examined in isolation, with less work looking at how these factors may interact to influence occupant satisfaction.

With these gaps in the literature in mind, the present study consisted of three simulator-based experiments that investigated journey satisfaction with autonomous highway journeys, from an occupant perspective. The main aim was to examine how journey satisfaction might be influenced by: i) travelling in an autonomous vehicle that was mostly *overtaking* compared with being mostly *overtaken*, and ii) being in a vehicle that was travelling at versus below the maximum legal speed limit. We also examined the influence of: i) traffic context, ii) having information presented or not regarding the vehicle's current speed and the speed limit of the road, and iii) the influence of behavioural nudges (techniques used to influence decision-making unconsciously by making small changes to the environment where choices are being made: See Thaler & Sunstein, 2008) on the above factors.

To preview the results, Experiment 1 found that occupants of an autonomous vehicle were happier overtaking compared to being overtaken, even when their vehicle was travelling at the maximum legal limit. Experiment 2 aimed to improve journey satisfaction when being overtaken while travelling at the legal limit by presenting visual and auditory nudges, but found the same pattern of results as Experiment 1. Experiment 3 found that occupants showed a preference for overtaking over being overtaken when following another car, but this difference disappeared when following a lorry, suggesting an impact of traffic context on journey satisfaction.

4.2. Experiment 1: The influence of lane position, speed, and information presentation on journey satisfaction

Experiment 1 examined the influence of three factors on journey satisfaction, perceived journey duration, vehicle speed, and EDA responses: Lane position (left or right), Vehicle speed (60mph or 70mph) and the presentation of speed Information (present or absent). For lane position, participants travelled in either the left or the right lane of a UK 2-lane (dual) carriageway (left-hand traffic rules). When travelling in the left lane they were constantly overtaken by vehicles in the right lane and when travelling in the right lane they were constantly overtaking vehicles in the left lane. In addition to journey satisfaction ratings we also recorded Electrodermal activity (EDA: Daviaux et al., 2020; Dewe, Braithwaite & Watson, 2016) as a potential measure of occupant frustration.

4.2.1. *Method*

Participants

Experiment 1 recruited sixty participants, fifty-eight of whom provided demographic information (Table 4.1). The number of participants was increased slightly compared to the simulator experiments in Chapter 3. This is because, while Chapter 3 focused on immediate reactions to specific manoeuvres (e.g. a sharp pull-in) which resulted in large effects, the negative influence of a longer journey (where one is repeatedly overtaken) was presumed to have a subtler effect on participants' evaluations of autonomous driving scenarios. All participants were paid £5 for taking part.

Table 4.1. Participant details (F = Female, RH = right-handed) for all three experiments presented in Chapter 4.

Experiment	Number of participants (number female)	Handedness	Age range, mean and standard deviation	Participants with driving license	Driving license duration and type
1	60 (26 F, 1 "Other")	50 RH, 7 LH, 1 A	18 – 53 years M= 22.41 SD= 7.4	60	0 – 36 years 18 UK
2	60 (29 F)	51 RH, 2 LH, 2 A	19 – 38 years M= 23.96 SD= 3.66	60	1 – 18 years 10 UK
3	30 (12 F)	25 RH, 4 LH	18 – 27 years M= 20.18 SD= 2.14	30	0.5 – 8 years 10 UK

Driving simulator setup

The simulator consisted of a fixed-base setup, with the front half of Jaguar XJ 2009 used for the cabin (*Figure 4.1*) running SCANeR Studio 1.4. There were three projection screens in front of the vehicle (SXGA+ resolution, ~135° horizontal visual angle) and three rear screens reflected in the left, rear and right mirrors. Sound was provided by a 5.1ch surround system. Tactile feedback was provided by a shaker located under the driver's seat. A custom dashboard display on an LCD panel displayed vehicle speed (mph) and the speed limit of the road (*Figure 4.2*). An additional LCD panel located within the centre console was used to administer the questionnaire measures. The driving scenario consisted of a 16km dual carriageway style road (along which participants travelled for 6km), with two lanes (all 3.65m wide) on each side of a central reservation (2m wide), and barriers on the outer edge of the slow lane on each side and around the central reservation. There were thirty-five vehicles in the scenario including the participants' own vehicle. There were two vehicles in the participant's own lane, one 80m in front of the participant and one 80m behind. There were twenty vehicles in the lane adjacent to the participant, spaced between 40 and 80m apart

(average distance of 60m). There were twelve vehicles on the opposing carriageway, with nine in the left lane (average spacing of 175m) and three in the right lane (average spacing of 375m). The traffic in the adjacent lane of the participant's carriageway was arranged in a loop based on the distance from the participant's own vehicle, with vehicles respawning 600m behind the participant once they reached 600m in front of the participant (in the being overtaken conditions), and respawning 600m in front of the participant once they reached 600m behind the participant (overtaking conditions). This was done to provide a constant stream of traffic either being overtaken by or overtaking the participants' vehicle. The traffic in the opposing carriageway (for both being overtaken and overtaking conditions) respawned 800m in front on the participant's position when the vehicles reached 800m behind the participants' vehicle.

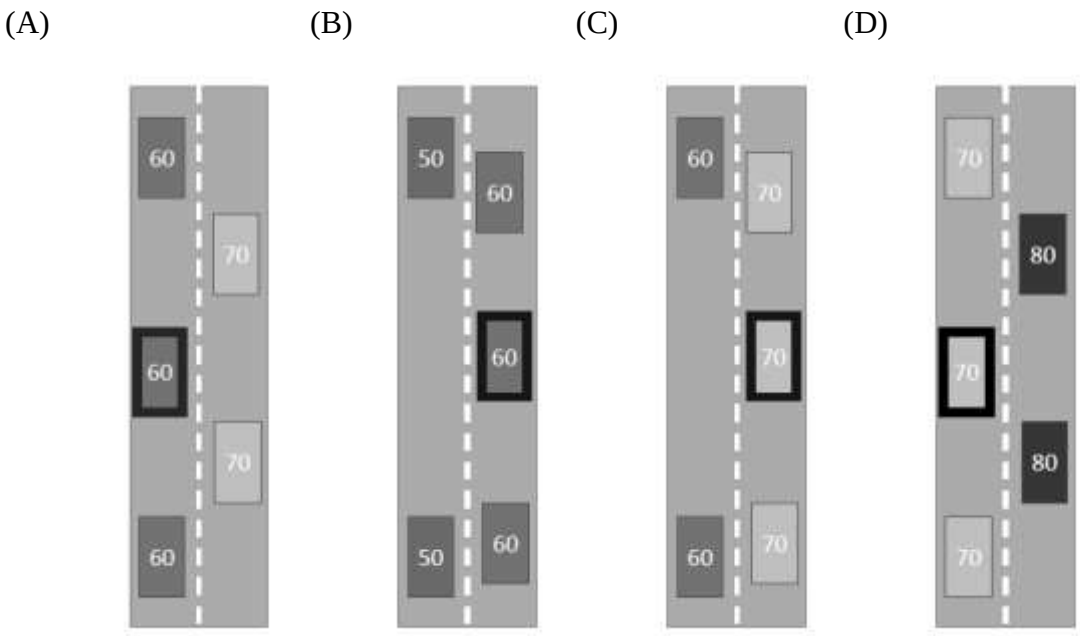


Figure 4.1. Schematic of the four journeys used in Experiment 1 & 2 of Chapter 4 (A: Being Overtaken at 60mph; B: Overtaking at 60mph; C: Overtaking at 70mph; D: Being Overtaken at 70mph). Participants in Experiment 3 were always travelling at 60mph and were either being overtaken/overtaking while following another car/following a lorry. Each journey also contained two lanes of opposing (light) traffic on the other side of a central reservation surrounded by barriers. The rectangle highlighted with bold edges indicates the participants' vehicle in each journey.

Figure 4.2. Example dashboard display from Experiments 1-3 of Chapter 4. Participants who were presented with speed information (half of all participants in Experiments 1 & 2, all participants in Experiment 3) viewed their own vehicle's speed (in mph), the speed limit (always 70mph, top right), and in Experiment 2 viewed a green glow around the dashboard (depicted below) when their vehicle was travelling at the maximum legal limit.

Physiological measurements

Electrodermal activity (EDA) was recorded via disposable EL507 electrodes attached to the distal phalanges of the fingers of the non-dominant hand. Measurements were recorded using a Biopac MP36R data acquisition unit and analysed with Acqknowledge v4.1. EDA signals were sampled at 1000Hz. EDA measurements used a gain of x2000, with low pass filters at 66.5 and 38.5Hz and band stop line frequency filter at 50Hz. Electrodermal activity was used as a potential objective measure of frustration during the simulator journeys, as previous research has shown it to be sensitive to attention being captured by unpleasant or threatening stimuli (Dewe et al., 2016), and EDA has also been linked to stress levels in the context of driving simulation (Daviaux et al., 2020).

Design and procedure

Within any single journey, participants' own vehicle was either in the left lane and being overtaken, or in the right lane and overtaking other traffic. Their vehicle was also either travelling at or below the road's speed limit (60mph vs 70mph, with the speed limit always 70mph). Half the participants were presented with their current speed and the maximum legal speed limit for the road (always 70mph) on the instrument panel. The other half received no speed information. Thus, this was a 2 (Lane position: left/right) x 2 (Vehicle speed: 60mph/70mph) x 2 (Information: yes/no) mixed design with information presentation as the

between-subjects variable. A single participant completed four trials corresponding to the four combinations of land position and speed.

Participants were given consent forms and screened using the Simulator Sickness Questionnaire (SSQ: Kennedy, Lane, Berbaum & Lilienthal, 1993). Consenting participants were first given two short practise journeys with no other traffic present (both at 65mph, one in the left lane and one in the right). Participants then completed four full journeys: one of each combination of lane position and speed. After each journey, participants were asked to verbally estimate the amount of time each journey took, and to estimate their own vehicle's speed (or to report the speed for participants presented with speed information). Participants were then presented with a modified Vehicle Character Questionnaire (VCQ: Table 4.2) with each question presented sequentially on a touch-screen LCD panel located in the centre dashboard console and participants responded by touching an option on a 7-point Likert scale. For each trial, the average of the scores across all questions was calculated to give a VCQ rating that ranged between 1 and 7. As with the original VCQ used in Chapter 3, the modified VCQ showed good internal consistency between items (Cronbach's $\alpha = .802$ for Experiment 1, .873 for 2, and .862 for 3), supporting use of the mean VCQ scores. After completing the final journey participants were screened for simulator sickness, debriefed, and thanked. The experiment lasted 45 to 60 minutes.

Table 4.2. Full Vehicle Character Questionnaire (VCQ) as adapted for Chapter 4.

Question number	Question text
1	With 1 being "unpleasant", and 7 being "pleasant", please rate the journey you have just experienced
2	With 1 being "unsafe", and 7 being "safe", please rate the journey you have just experienced.
3	With 1 being "boring", and 7 being "exciting", please rate the journey you have just experienced.
4	With 1 being "uncomfortable", and 7 being "comfortable", please rate the journey you have just experienced
5	With 1 being "stressful", and 7 being "relaxing", please rate the journey you have just experienced.
6	With 1 being "inefficient", and 7 being "efficient", please rate the journey you have just experienced.
7	With 1 being "slow", and 7 being "fast", please rate the journey you have just experienced.
8	Would you be more or less likely to purchase a vehicle that behaved like your vehicle in the driving scenario you have just experienced? 1 = highly unlikely, 7 = highly likely
9	Would you be happy to be driven by a vehicle that behaved like your vehicle in the driving scenario you have just experienced? 1 = highly unlikely, 7 = highly likely
10	Would you use a system in your own car that behaved like your vehicle in the driving scenario you have just experienced? 1 = highly unlikely, 7 = highly likely

VCQ scores ranged from 1-7, with lower scores being more negative and higher scores being more positive. The VCQ questionnaire was modified from the version used in Chapter 3, in order to better reflect the nature of the task. The questions used in Chapter 3 focused on the acceptability of specific manoeuvres (a sharp pull-in during an overtake), whereas those in Chapter 4 focused on the evaluation of a longer highway journey. Therefore, slightly different wordings and questions were required for the two chapters on vehicle character.

4.2.2. Results

Questionnaire ratings

Questionnaire ratings for Experiment 1 are shown in *Figure 4.3* and were analysed using a 2 (Lane position) × 2 (Vehicle speed) × 2 (Information) mixed ANOVA. There was a significant main effect of speed on journey satisfaction ratings, such that participants preferred travelling at 70 over 60mph, F (1, 58) = 12.423, $p < .001$, $\eta^2 = .020$. There was a significant main effect of lane position, such that participants preferred overtaking to being overtaken, F (1, 58) = 12.030, $p < .001$, $\eta^2 = .034$. There was also a significant interaction between speed and information, F (1, 58) = 4.741, $p < .05$, $\eta^2 = .008$. Simple main effects tests revealed that participants who were given information gave more positive ratings than those not given speed information, but only for journeys where they were travelling at 70mph, F (1, 58) = 5.040, $p < .05$, with no effect of information at 60mph, F (1, 58) = .506, p = .480. Simple main effects also revealed that participants' preference for travelling at 70

over 60mph was only significant for participants given information, F (1, 58) = 13.842, p < .001, with no effect for those not given information, F (1, 58) = 1.099, p = .303.

No other main effects or their interaction were significant: Information, F (1, 58) = 2.542, η^2 = .027, p = .116, speed × lane position, F (1, 58) = 1.097, p = .299, η^2 = .000, information × lane position, F (1, 58) = .538, p = .466, η^2 = .002, speed × lane position × information, F (1, 58) = .060, p = .807, η^2 = .00.

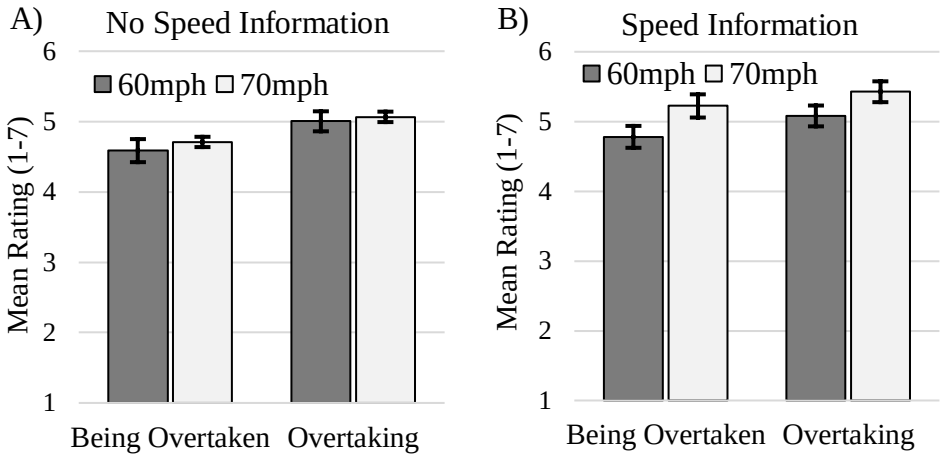


Figure 4.3. Journey satisfaction ratings (higher = more positive) for Experiment 1 by lane position and speed (Panel A: No speed information; Panel B: Speed information).

Estimated speed

Participants' estimates of their own vehicle's speed during each journey are shown in *Figure 4.4* and were analysed in the same way as the questionnaire data. All three main effects were significant. Participants gave higher speed estimates in the 70mph conditions than in the 60mph conditions, F (1, 58) = 69.766, p < .001, η^2 = .515. Higher speed estimates were given for conditions in which participants were overtaking than being overtaken, F (1, 58) = 43.893, p < .001, η^2 = .359. Participants who were presented with information reported higher estimated speeds that those who did not receive any information, F (1, 58) = 35.152, p < .001, η^2 = .377. There was a significant interaction between speed and information, F (1, 58) = 7.708, p < .01, η^2 = .057; the difference between reported speeds in the 60 and 70mph conditions was larger when speed information was present. Simple main effects revealed that participants presented with information reported higher speeds in the 70mph conditions than in the 60mph, F (1, 58) = 245.848, p < .001, and this was also true to participants who were not presented with information, F (1, 58) = 8.894, p < .01. Simple main effects also revealed that 60mph journeys were reported to be significantly faster for participants presented with

information, compared to those not presented with information, F (1, 58) = 22.294, p < .001. 70mph journeys were also reported to be significantly faster for participants presented with information, F (1, 58) = 40.192, p < .001.

There was a significant interaction between lane position and information, F (1, 58) = 20.362, p < .001, η^2 = .167, such that estimated speed differences between lanes were larger for participants not presented with information. Simple main effects revealed that participants presented with information reported higher speeds when they were overtaking than when they were being overtaken, F (1, 58) = 4.949, p < .05. Participants not presented with information also gave higher speed estimates for journeys in which they were overtaking, F (1, 58) = 40.040, p < .001. Simple main effects also revealed that participants presented with information reported higher speeds for journeys in which they were being overtaken, compared to participants not presented with information, F (1, 58) = 49.847, p < .001. Participants presented with information also reported higher estimated speeds for journeys in which they were being overtaken, compared to participants not presented with information, F (1, 58) = 15.645, p < .001.

Finally, there was no significant interaction between speed and lane position, F (1, 58) = 2.727, p = .104, η^2 = .044, nor a significant three-way interaction, F (1, 58) = .774, p = .383, η^2 = .013.

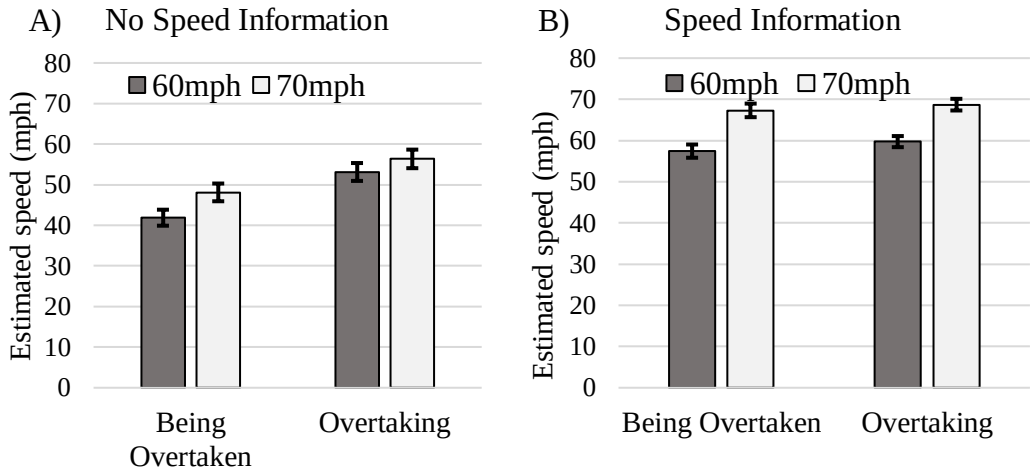


Figure 4.4. Speed estimation data for Experiment 1 by lane position and speed (Panel A: No speed information; Panel B: Speed information).

Estimated journey time

Estimated journey time for Experiment 1 is shown in *Figure 4.5* and was analysed using the same approach as with estimated speed. This revealed only a significant main effect

of speed, with participants reporting that 60mph journeys took longer than 70mph journeys, F (1, 58) = 39.880, $p < .001$, $\eta^2 = .407$. Neither the main effect of lane position, F (1, 58) = .409, $p = .525$, $\eta^2 = .007$, nor the main effect of information was significant, F (1, 58) = .651, $p = .423$, $\eta^2 = .011$. Furthermore, no interactions reached significance: speed × lane position, F (1, 58) = 2.990, $p = .089$, $\eta^2 = .048$, speed × information, F (1, 58) = .067, $p = .797$, $\eta^2 = .001$, lane position × information, F (1, 58) = 1.009, $p = .319$, $\eta^2 = .017$, speed × lane position × information, F (1, 58) = 1.453, $p = .233$, $\eta^2 = .023$.

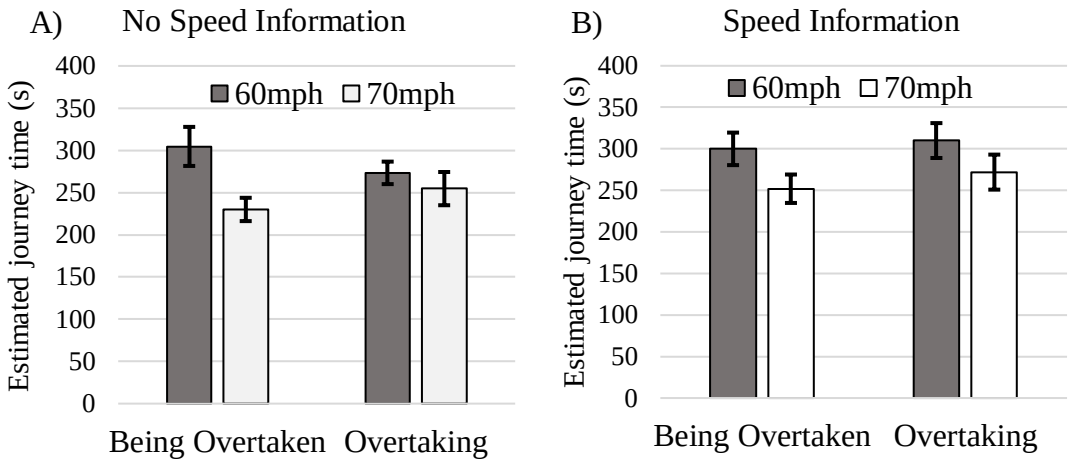


Figure 4.5. Estimated journey time data for Experiment 1 by lane position and speed (Panel A: No speed information; Panel B: Speed information).

Electrodermal activity

Two key measures were analysed for the physiological data: The number of Skin Conductance Responses (SCRs) participants exhibited per minute for each journey, and participants' average Skin Conductance Level (SCL) over each journey. After smoothing out motion artefacts from the physiology trace, an automated analysis routine was used to determine when SCRs occurred (using a minimum SCR size of 0.02μs). Three participants were excluded from the analysis due to showing a very low overall level of reactivity across all journeys. The average SCL per journey was conducted on the entire duration of the journey. While it has been suggested that using the entire trace to analyse SCL can result in the SCL being confounded by SCRs within the trace (Braithwaite, Watson, Jones & Rowe, 2015), this method is appropriate for the present study as both the SCL and frequency of SCRs were used as a secondary measure of frustration levels across the journey. As such,

there was no need to separate these two components of the EDA signal for the purposes of the present study.

Number of SCRs per minute

The number of SCRs per minute is shown in *Figure 4.6* and was analysed using a 2 (Lane position) × 2 (Speed) × 2 (Information) mixed ANOVA. This revealed no significant main effects or two-way interactions: Lane position, F (1, 55) = .449, p = .506, η^2 = .000, speed, F (1, 55) = .009, p = .925, η^2 = .000, information., F (1, 55) = .558, p = .458, η^2 = .008, lane position × speed, F (1, 55) = .277, p = .601, η^2 = .000, lane position × information, F (1, 55) = .008, p = .927, η^2 = .000, speed × information, F (1, 55) = .753, p = .389, η^2 = .000. However, there was a significant three-way interaction between lane position, speed and information, F (1, 55) = 9.126, p < .01, η^2 = .010.

Simple main effects tested were used to follow up the three-way interaction. First, the effect of lane position was examined, with speed as the first moderator factor and information as the second. There was a significant effect of lane position at 60mph for participants given information, such that more SCRs per minute occurred for journeys in which participants were overtaking, F (1, 55) = 4.272, p < .05. The effect of lane position did not reach significance at any other level of speed or information: 60mph with no information, F (1, 55) = .846, p = .366, 70mph with no information, F (1, 55) = 1.703, p = .202, 70mph with information, F (1, 55) = 1.576, p = .220. Second, the effect of speed was examined with lane position as the first moderator factor and information as the second. There was a significant effect of speed for participants who were given information, for journeys in which participants were overtaking, such that there were fewer SCRs per minute when participants were travelling at 70mph compared to 60mph, F (1, 55) = 7.107, p < .05. The effect of speed did not reach significance at any other level of lane position or information: Being overtaken with no information, F (1, 55) = 1.868, p = .183, being overtaken with information, F (1, 55) = 1.273, p = .269, overtaking with no information, F (1, 55) = 1.481, p = .234. Finally, the effect of information was examined with lane position as the first moderator factor and speed as the second. There were no significant effects of information at any combination of lane position and speed: Being overtaken at 60mph, F (1, 55) = 1.513, p = .224, being overtaken at 70mph, F (1, 55) = .023, p = .879, overtaking at 60mph, F (1, 55) = .111, p = .740, overtaking at 70mph, F (1, 55) = 3.207, p = .079.

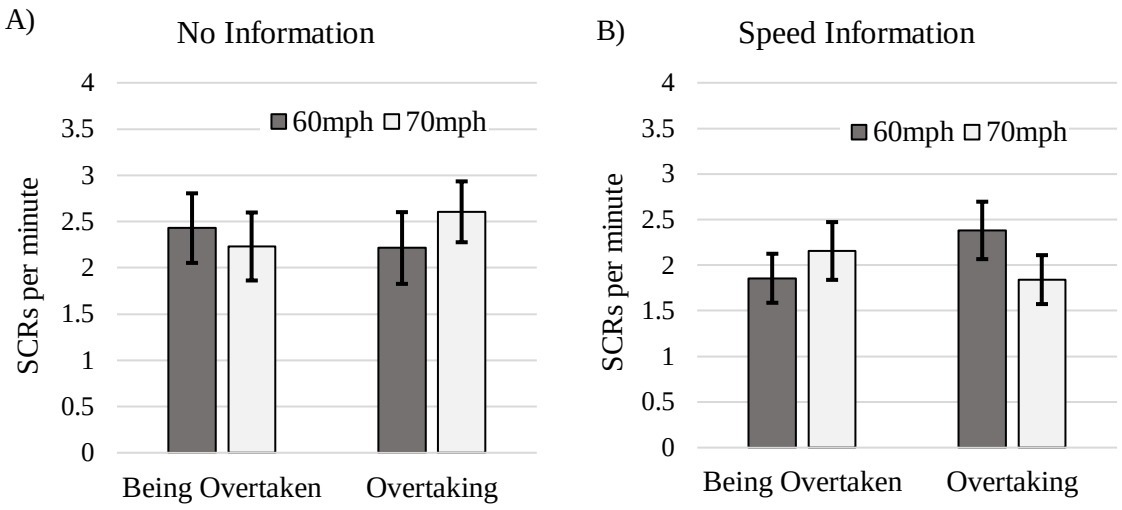


Figure 4.6. Frequency of Skin Conductance Responses (SCRs) per minute as a function of lane position and speed for Experiment 1, for participants provided with no information (N = 29, Panel A) and for participants provided with speed information (N = 27, Panel B).

Overall SCL

Participants' overall SCLs are shown in *Figure 4.7* and were analysed using a 2 × 2 × 2 mixed ANOVA (Lane position × Speed × Information). No main effects or their interaction were significant: Lane position, F (1, 55) = 2.314, p = .134, η^2 = .000, speed, F (1, 55) = .609, p = .438, η^2 = .000, information, F (1, 55) = .338, p = .564, η^2 = .006, lane position × speed, F (1, 55) = .422, p = .518, η^2 = .000, lane position × information, F (1, 55) = .664, p = .419, η^2 = .000, speed × information, F (1, 55) = .187, p = .667, η^2 = .000, lane position × speed × information, F (1, 55) = .007, p = .933, η^2 = .000.

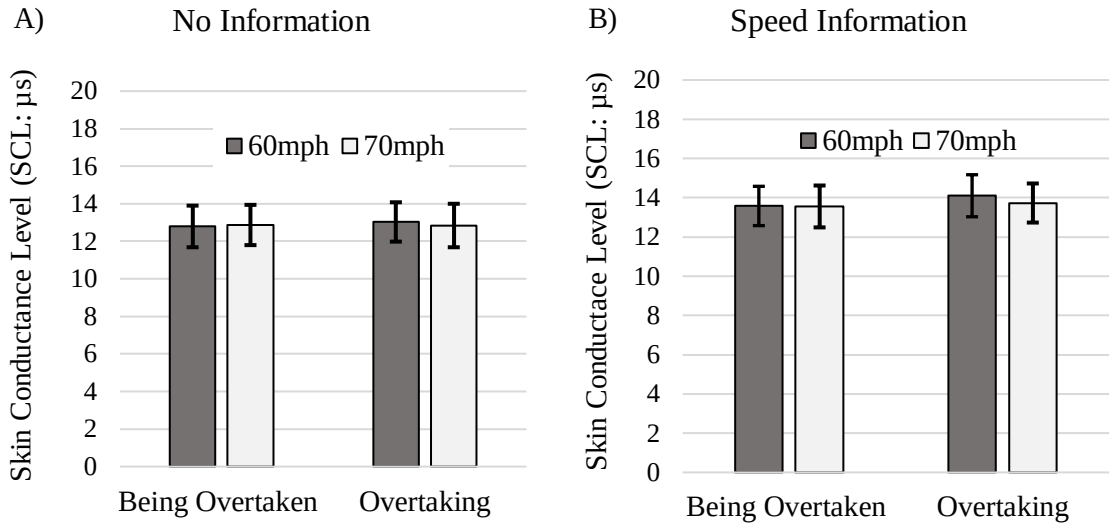


Figure 4.7. Skin Conductance Level (SCL) as a function of lane position and speed for Experiment 1, for participants provided with no information (N = 29, Panel A) and for participants provided with speed information (N = 28, Panel B).

4.2.3. Discussion

Experiment 1 investigated the influence of lane position, vehicle speed and information presentation on evaluations of highway journeys in an autonomous vehicle. Participants preferred overtaking to being overtaken, irrespective of whether they were travelling below or even at the legal limit. Importantly, being overtaken led to decreased journey satisfaction even when the vehicles doing the overtaking were exceeding the maximum legal limit. This suggests that rule-violating behaviour from other vehicles can reduce satisfaction for occupants of autonomous vehicles, even if the autonomous vehicle is acting within the rules of the road. Occupants preferred travelling at the legal limit compared to below it, but only when speed information was presented visually on the dashboard, suggesting that this effect was due to the perception of whether or not the vehicle could travel more quickly, rather than vehicle's speed itself. This has important implications for the extent to which information should be presented to occupants of autonomous vehicles even if that information has no direct role in vehicle control.

The finding that satisfaction was reduced by other vehicles overtaking and travelling above the legal limit is potentially concerning for mixed traffic environments consisting of both human-driven and autonomous vehicles, relating to concerns that human drivers could take advantage of ruled-based autonomous vehicles (Madigan, et al., 2019; Tennant, 2015; Tennant et al., 2016). Experiment 2 tested if occupant frustration could be reduced (particularly when being overtaken) by providing additional auditory and visual nudges

(Thaler & Sunstein, 2008) when occupants were travelling at the maximum limit, to increase the salience of the fact that the occupants' vehicle is travelling as quickly as possible, without breaking the law, despite being overtaken.

4.3. Experiment 2: Influence of additional nudges on journey evaluations

Experiment 2 was designed to address a surprising finding from Experiment 1 (that being overtaken led to frustration even when occupants were travelling at the maximum legal limit). Experiment 2 investigated whether the frustration of travelling at 70mph and being overtaken could be offset by introducing a 'nudge' (Thaler & Sunstein, 2008) to reassure occupants that they were travelling at the maximum limit, under the assumption that participants discounted this information in the previous experiment, possibly because the dashboard display was entirely visual and easy to miss while paying attention to the road environment. Specifically, rather than simply displaying the current speed limit and maximum speed limit in the dashboard display, two much more salient additional cues were added. First was a relatively bright glow that appeared around the outer edge of the dash LCD panel. The colour of the glow depended on the speed; no glow = below the speed limit, green = at the speed limit (or 2mph under/above, i.e. between 68 and 72mph), red = over the speed limit. The second nudge was an auditory prompt which ensured the occupants that they were travelling at the maximum speed and would reach their destination on time.

4.3.1. *Method*

Participants

Experiment 2 recruited sixty participants, fifty-five of whom provided demographic information (Table 4.1). All participants were paid £5 for taking part.

Driving simulator setup

Experiment 2 used the same driving simulator setup, the same road layout and traffic conditions as Experiment 1. The only difference was that participants in the 'nudge' condition viewed a dashboard with a green glow around the edges when their vehicle was travelling at the legal speed limit (70mph), and also received auditory prompts delivered via the simulator's sound system.

Design and procedure

The overall design was very similar to that of Experiment 1. Participants who were not provided with any speed information completed the same four journeys as in Experiment 1 (2 × 2 combinations of Speed and Lane position). Participants who were given speed information also completed these four combinations, but were presented with a green glow

around the edges of the dashboard display (*Figure 4.*2) in the two journeys where their vehicle was travelling at the maximum legal limit (70mph). Participants given speed information were also provided with an auditory prompt from the vehicle every 1km, stating that "You are travelling at 70 miles per hour, which is the maximum legal limit for this road. Your estimated arrival time has not changed"), in journeys where their vehicle was travelling at 70mph.

The overall procedure was similar to that of Experiment 1, participants completed the four combinations of lane position and speed in a randomised order. Participants completed three practise trials. For the first two practise trials, participants were in the left or right lane for one of each. The participants' vehicle accelerated from 0 to 60mph, then from 60 to 70mph, and then from 70 to 80mph (after which the journey was stopped). For participants who were presented with speed information, the background of the dashboard remained black (with no glow) below 68mph. Between 68 and 72mph, a green glow appeared around the edges of the dashboard. When the vehicle's speed exceeded 72mph, the edges of dashboard glowed red. Participants who were not presented with speed information experienced the same speed changes but without any feedback from the dashboard. For the third practise trial each participant received, the participants' vehicle travelled at 70mph, either in the left or right lane (counterbalanced between-subjects), with the green glow around the dashboard as in the first two practises. In this final practise, participants were also played two example voice messages ("this is an example voice message from the vehicle") to familiarise them with the voice-message system.

4.3.2. Results

Questionnaire ratings

Questionnaire ratings for Experiment 2 are shown in *Figure 4.8.* A 2 × 2 × 2 mixed ANOVA was conducted on the questionnaire data. There was a significant main effect of speed on journey satisfaction ratings, such that participants preferred travelling at 70 over 60mph, F (1, 58) = 24.354, $p < .001$, $\eta^2 = .288$. There was a significant main effect of lane position, such that participants preferred overtaking to being overtaken, F (1, 58), = 31.660, $p < .001$, $\eta^2 = .345$. The following main effects and interactions were non-significant: Information, F (1, 58) = .005, $p = .946$, $\eta^2 = .000$, speed × lane position, F (1, 58) = .009, $p = .923$, $\eta^2 = .000$, information × lane position, F (1, 58) = 2.024, $p = .160$, $\eta^2 = .022$, speed × information, F (1, 58) = 2.351, $p = .131$, $\eta^2 = .028$.

An additional analysis was conducted comparing the results of Experiments 1 to those of Experiment 2, in order to assess the effectiveness of the additional behavioural nudges applied in Experiment 2. The data from both experiments were combined into a larger dataset (N = 120) and analysed using a 2 (Experiment) × 2 (Information) × 2 (Lane position) × 2 (Speed) mixed ANOVA. For the purpose of this analysis, the *no nudge* vs *nudge* manipulation from Experiment 2 was considered the same as the *no information* vs *information* manipulation from Experiment 1, resulting in one between-subjects variable for Information across the combined dataset.

There was a significant main effect of lane position on journey satisfaction, F (1, 116) = 41.356, $p < .001$, $\eta^2 = .048$, such that occupants preferred overtaking over being overtaken. There was a significant main effect of speed, F (1, 116) = 36.725, $p < .001$, $\eta^2 = .032$, such that occupants preferred travelling at 70mph over 60mph. There was a significant interaction between speed and information, F (1, 116) = 6.475, $p < .05$, $\eta^2 = .006$. Simple main effects revealed that travelling at 70mph over 60mph was preferred both by participants presented with information, F (1, 116) = 32.859, $p < .001$, and by participants not presented with information, F (1, 116) = 7.075, $p < .05$. Simple main effects also revealed that, for journeys at 60mph, there was no significant influence of information, F (1, 116) = .003, $p = .956$, whereas there was a borderline significant effect of information at 70mph, F (1, 116) = 3.674, $p = .058$, with a trend of participants presented with information rating 70mph journeys as more positive than those not presented with information.

The following main effects were non-significant: Experiment: F (1, 116) = 1.105, $p = .295$, $\eta^2 = .006$, information, F (1, 116) = .981, $p = .324$, $\eta^2 = .005$. The following two-way interactions were non-significant: Experiment × information, F (1, 116) = 1.196, $p = .276$, $\eta^2 = .006$, lane position × experiment, F (1, 116) = 2.324, $p = .130$, $\eta^2 = .003$, lane position × information, F (1, 116) = 2.324, $p = .130$, $\eta^2 = .003$, speed × experiment, F (1, 116) = 3.064, $p = .083$, $\eta^2 = .003$, lane position × speed, F (1, 116) = .297, $p = .587$, $\eta^2 = .000$. The following three-way interactions were non-significant: Lane position × experiment × information, F (1, 116) = .237, $p = .627$, $\eta^2 = .000$, speed × experiment × information, F (1, 116) = .014, $p = .907$, $\eta^2 = .000$, lane position × speed × experiment, F (1, 116) = .137, $p = .712$, $\eta^2 = .000$, lane position × speed × information, F (1, 116) = 2.249, $p = .136$, $\eta^2 = .001$. The four-way lane position × speed × experiment × information interaction was also non-significant, F (1, 116) = 1.652, $p = .201$, $\eta^2 = .000$.

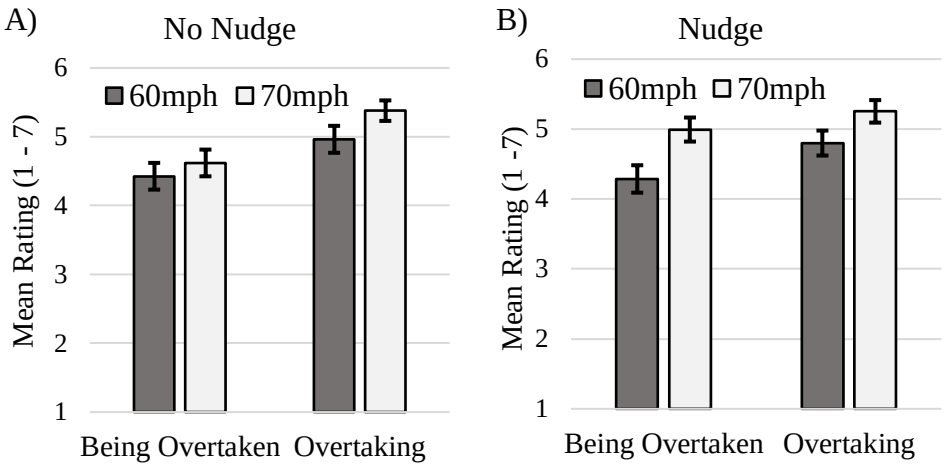


Figure 4.8. Journey satisfaction ratings for Experiment 2 by lane position and speed (Panel A: No speed information; Panel B: Speed information).

Estimated speed

Speed estimations for Experiment 2 are shown in *Figure 4.9.* A 2 × 2 × 2 mixed ANOVA was conducted on participants' estimations of their own vehicle's speed during each journey. There was a significant main effect of speed on estimates of speed, with participants reporting that 70mph journeys were faster than 60mph journeys, F (1, 58) = 100.064, p < .001, η^2 = .599. There was a significant main effect of lane position on estimates of speed, with participants reporting that journeys in which they were overtaking were faster than those in which they were being overtaken, F (1, 58) = 28.093, p < .001, η^2 = .287. There was a significant main effect of information on estimates of speed, with participants who were presented with speed information (and additional nudges compared to Experiment 1) reporting overall that journeys were faster than those who were not presented with information, F (1, 58) = 172.100, p < .001, η^2 = .748.

There was a significant interaction between speed and information, F (1, 58) = 8.940, p < .01, η^2 = .054. Simple main effects revealed that 70mph journeys were reported to be significantly faster than 60mph journeys for participants presented with information, F (1, 58) = 177.832, p < .001. 70mph journeys were also reported to be significantly faster than 60mph journeys for participants not presented with information, F (1, 58) = 16.123, p < .001. Simple main effects also revealed that participants presented with information reported that 60mph journeys were significantly faster, compared to those not presented with information, F (1, 58) = 137.109, p < .001. 70mph journeys were also reported to be significantly faster by

participants presented with information, compared to those not presented with information, F (1, 58) = 161.838, $p < .001$.

There was a significant interaction between lane position and information, F (1, 58) = 11.725, $p < .01$, $\eta^2 = .120$. Simple main effects revealed that there was no significant difference in estimates of speed between overtaking and being overtaken for participants presented with information, F (1, 58) = 2.151, $p = .153$. Participants who were not presented with information reported that journeys in which they were overtaking were significantly faster than those in which they were being overtaken, F (1, 58) = 32.200, $p < .001$. Simple main effects also revealed that participants presented with information reported that journeys in which they were being overtaken were significantly faster, compared to those not presented with information, F (1, 58) = 190.990, $p < .001$. Journeys in which participants were overtaking were also reported to be significantly faster by participants presented with information, compared to those not presented with information, F (1, 58) = 113.319, $p < .001$.

The following interactions were non-significant: Speed × lane position, F (1, 58) = 1.096, $p = .299$, $\eta^2 = .019$, speed × lane position × information, F (1, 58) = .089, $p = .767$, $\eta^2 = .001$.

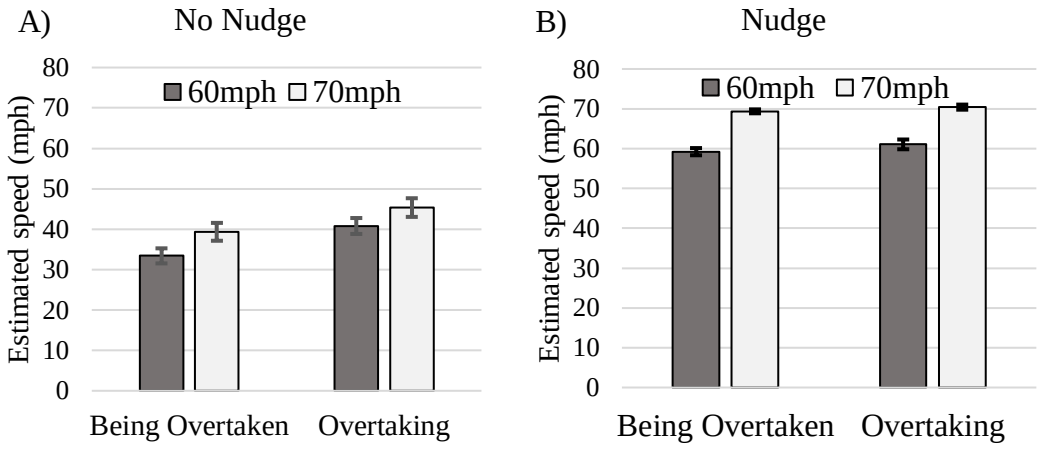


Figure 4.9. Speed estimation data for Experiment 2 by lane position and speed (Panel A: No speed information; Panel B: Speed information).

Estimated journey time

Estimated journey time for Experiment 2 is shown in *Figure 4.10.* A 2 × 2 × 2 mixed ANOVA was conducted on participants' estimations of the duration of each journey. There was a significant main effect of speed on estimates of journey time, with participants reporting that 60mph journeys took longer than 70mph journeys, F (1, 58) = 24.988, $p < .001$, $\eta^2 = .300$. The following main effects and interactions were non-significant: Lane position, F

(1, 58) = 0.000, p = .989, η^2 = .000, information, F (1, 58) = .256, p = .615, η^2 = .004, speed × lane position, F (1, 58) = 2.857, p = .096, η^2 = .047, speed × information, F (1, 58) = .190, p = .664, η^2 = .002, lane position × information, F (1, 58) = .023, p = .879, η^2 = .000, speed × lane position × information, F (1, 58) = .206, p = .651, η^2 = .003.

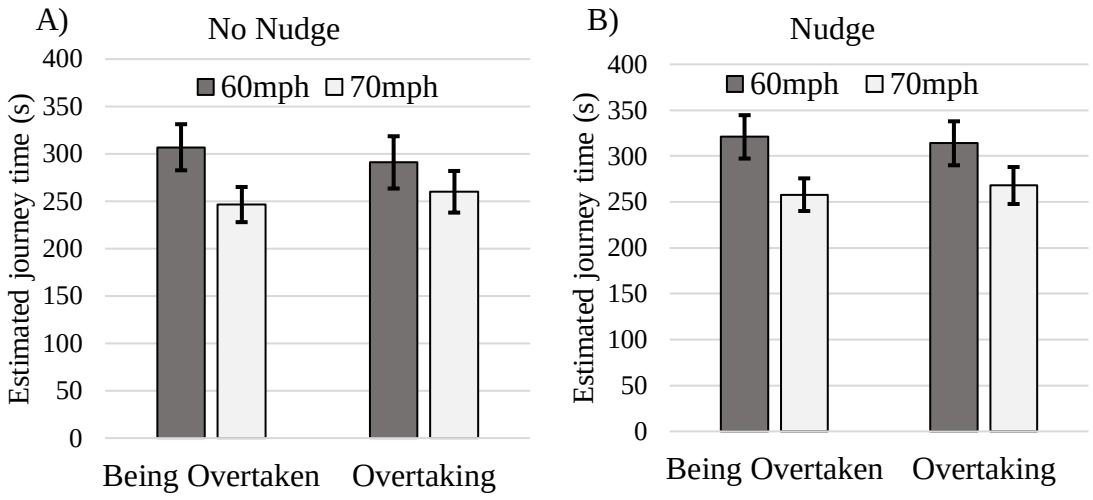


Figure 4.10. Estimated journey time data for Experiment 2 by lane position and speed (Panel A: No speed information; Panel B: Speed information).

4.3.3. Discussion

Experiment 2 was motivated by a surprising finding in Experiment 1, that occupants of autonomous vehicles were less happy being overtaken compared to overtaking, even when their vehicle could not legally travel at a higher speed. Experiment 2 was similar to Experiment 1 in design, but participants who were presented with speed information were also given auditory and visual nudges (Thaler & Sunstein, 2008) to remind them that their vehicle was travelling at the maximum legal limit, with the aim that this would reduce frustration when being overtaken in these sorts of situations.

However, the results of Experiment 2 were very similar to those of Experiment 1, with no indication that occupants being overtaken while travelling at 70mph and being provided with nudges were any happier than those provided with no information. The results also suggest that the high frequency of the auditory nudge (reminding occupants were travelling at the maximum legal limit) may have reduced satisfaction for the 70mph journeys, as the preference for being provided with information (compared to no information) from Experiment 1 was not found in Experiment 2. However, a direct comparison between Experiments 1 and 2 found no differences in the effects of lane position, speed, or information on journey satisfaction across both experiments. This suggests that, while the

addition of behavioural nudges did not appear to improve journey satisfaction, the nudges did not actively make the journeys any less satisfying for occupants. These findings suggest that, in some cases, nudges such as the ones used in this experiment may not be effective at improving occupant satisfaction in frustrating driving scenarios such as being overtaken while travelling at the maximum legal limit.

While Experiments 1 and 2 examined potential interactions between information presentation and traffic flow, Experiment 3 investigated the possible role of traffic context on evaluations of autonomous highway journeys. Experiment 3 tested whether occupants' perceptions of lane position were different depending on the type of vehicle that was in front of them (i.e. a vehicle that was preventing the participants' vehicle from increasing speed). Specifically, participants were either overtaking or being overtaken, and they were either following a car or a lorry.

4.4. Experiment 3: Influence of lead vehicle type on journey evaluation

Experiment 3 aimed to extend the findings of Experiments 1 and 2 by further examining the influence of traffic context on perceptions of autonomous highway journeys. Participants in Experiment 3 were either being overtaken or overtaking, and either following another car or a lorry. If participants are sensitive to the type of lead vehicle, then participants being overtaken and following a lorry might become more frustrated than if following another car, due to the expectation that cars should overtake large, slow-moving vehicles in highway environments (HGVs typically travel at speeds around 10mph below that of cars, with articulated lorries travelling up to 15mph below the speed of cars: National Office of Statistics, Table SPE0111: Free flow vehicle speeds by road type and vehicle type in Great Britain, 2018).

Related to this possibility, the acceptability of overtaking manoeuvres has been shown to be sensitive to traffic context in previous simulator-based work: Drivers in Ritchie et al. (2019) were somewhat more forgiving of being overtaken at a sharp pull-in distance if the overtaking vehicle was changing lanes to allow a closely-following third vehicle to pass. Additionally, it has been shown that drivers are more likely to overtake slow-moving lead vehicles (Asaithambi & Shravani, 2017; Bar-Gera & Shinar, 2004). A video-based study also found that participants demonstrated higher levels of frustration when taking the perspective of a vehicle following long platoons of vehicles and travelling at lower speeds, and that shorter platoons and lower speeds were associated with higher intentions to overtake (Kinnear, Helman, Wallbank & Grayson, 2015). This background suggests that frustration in

autonomous highway journeys may be sensitive to whether one's vehicle is following a vehicle that would be expected to be in a slower lane.

4.4.1. *Method*

Participants

Experiment 3 recruited thirty participants, twenty-nine of whom provided demographic information (Table 4.1). Half the number of participants were recruited compared to Experiments 1 and 2 because Experiment 3 used a fully within-subjects design (compared to Experiments 1 & 2, in which Information was manipulated between-subjects). All participants were paid £5 for taking part.

Driving simulator setup

Experiment 3 used the same driving simulator setup and the same road layout as Experiment 1. However, participants in half of the journeys followed a lorry instead of another car (participants were always following another car in Experiments 1 and 2).

Design and procedure

Participants in Experiment 3 completed four journeys, either in the left or right lane, and either following another car or following a lorry. This resulted in a fully within-subjects 2 (Lane position) × 2 (Lead vehicle) design. All journeys were at 60mph (which was the indicated speed limit, chosen to reflect the speed of lorries in real-world highway environments), and all participants were presented with speed information (visually via the dashboard display, as in Experiment 1). Within a given journey, participants were either being overtaken or overtaking (as in Experiments 1 & 2) and were either following another car or following a lorry. Participants completed all four combinations of lane position and lead vehicle type in a randomised order, after completing two practise journeys (one in the left lane, one in the right) in which the participant's vehicle travelled at 60mph with no other traffic present. The overall procedure was the same as in Experiments 1 and 2.

4.4.2. Results

Questionnaire ratings

Questionnaire ratings for Experiment 3 are shown in *Figure 4.11.* A 2 × 2 within-subjects ANOVA was conducted on the questionnaire data. There was a significant main effect of lane position on journey satisfaction ratings, such that participants preferred overtaking to being overtaken, F (1, 29) = 4.215, $p < .05$, $\eta^2 = .127$. There was a significant interaction between traffic position and lead vehicle type, F (1, 29) = 7.645, $p < .05$, $\eta^2 = .209$. Simple main effects tests revealed that participants preferred overtaking over being

overtaken, but only when following another car, F (1, 29) = 8.648, p < .01, with no effect of traffic position when following a lorry, F (1, 29) = .192, p = .664. Simple main effects also revealed that, when overtaking, participants preferred following another car over following a lorry, F (1, 29) = 5.091, p < .05, with no effect of lead vehicle type when being overtaken, F (1, 29) = 1.627, p = .212. There was no significant main effect of lead vehicle type, F (1, 29) = .743, p = .396, η^2 = .025.

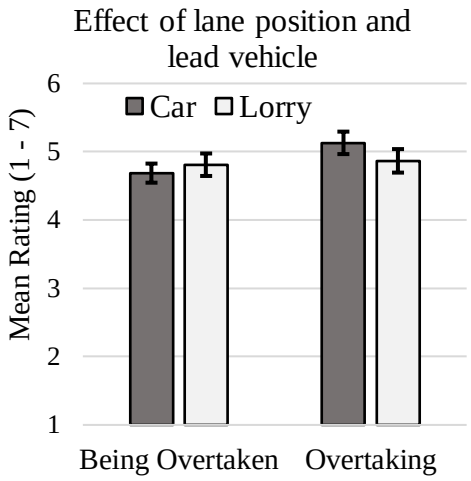


Figure 4.11. Journey satisfaction ratings for Experiment 3 by lane position and lead vehicle.

Estimated speed

Speed estimations for Experiment 3 are shown in *Figure 4.12.* A 2 × 2 within-subjects ANOVA was conducted on participants' estimations of their own vehicle's speed during each journey. No main effects or interactions were significant: Lane position, F (1, 29) = 1.960, p = .172, η^2 = .063, lead vehicle, F (1, 29) = .024, p = .879, η^2 = .001, lane position × lead vehicle, F (1, 29) = .024, p = .879, η^2 = .001.

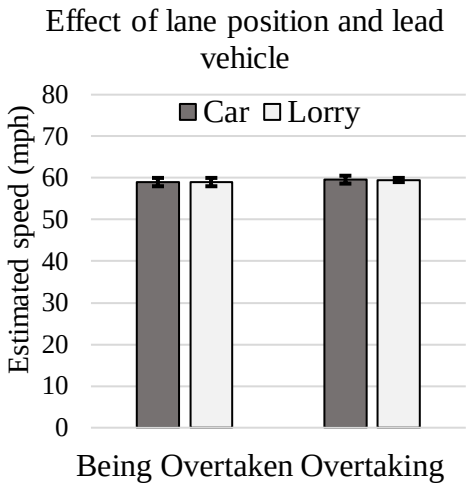


Figure 4.12. Speed estimation data for Experiment 3 by lane position and lead vehicle.

Estimated journey time

Estimated journey time for Experiment 3 is shown in *Figure 4.13.* A 2 × 2 within-subjects ANOVA was conducted on participants' estimations of the duration of each journey. No main effects or interactions were significant: Lane position, F (1, 29) = .008, p = .929, η^2 = .000, lead vehicle, F (1, 29) = .698, p = .410, η^2 = .024, lane position × lead vehicle, F (1, 29) = 2.920, p = .098, η^2 = .091.

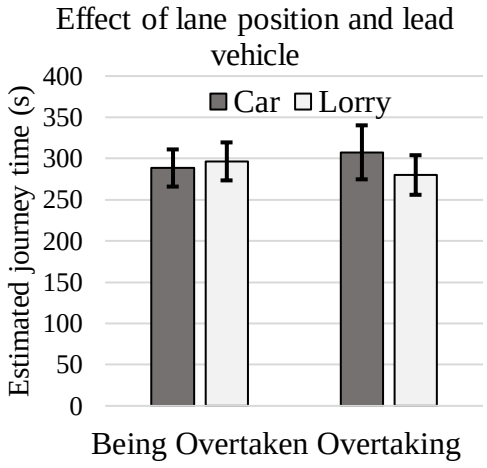


Figure 4.13. Estimated journey time data for Experiment 3 by lane position and lead vehicle.

4.4.3. Discussion

Experiment 3 investigated possible mediating effects of traffic context on occupants' ratings of satisfaction in autonomous highway journeys. The participants' vehicle was either overtaking or being overtaken, and either following a lorry or another car as the lead vehicle. As in Experiments 1 and 2, participants preferred overtaking to being overtaken, but showed

no overall preference for following a lorry or another car. However, the interaction uncovered between lane position and lead vehicle revealed that occupants only showed a preference for overtaking over being overtaken when following another car, with no effect of lane position when following a lorry. Following a larger, slow-moving vehicle may have counteracted the positive influence of overtaking other vehicles. Additionally, occupants preferred following another car, but only when overtaking (with no effect of vehicle type when being overtaken). In this case, the frustration caused by being overtaken may have counteracted any potential sensitivity to the type of vehicle being followed.

Overall, these results suggest a limited degree of sensitivity to traffic context in autonomous highway journeys, in which the type of vehicle ahead of the participant's vehicle only impacted satisfaction when overtaking. This limited influence of context is arguably similar to the small influence of a third vehicle on evaluations of autonomous overtaking manoeuvres in Chapter 3 (see also Ritchie et al., 2019), with pull-in distance of the overtaking manoeuvre exerting a much larger overall effect on character ratings. The lack of an overall effect of vehicle type on satisfaction ratings is surprising given previous findings that following slower-moving vehicles can lead to increased frustration and intentions to overtake (Kinnear et al., 2015). This may be because Experiment 3 did not manipulate speed (the occupants' vehicle speed was kept at 60mph throughout to more realistically reflect real-world speeds of lorries). If the lorry was travelling at a much-reduced speed, then the results may differ. Accordingly, future research may benefit from examining whether frustration when following a lorry differs dependent on vehicle speed and speed limit of the road.

4.5. General discussion

Three simulator-based experiments investigated the effect of lane position, vehicle speed, information presentation and traffic context on satisfaction with highway journeys in an autonomous vehicle. Experiment 1 found that occupants preferred overtaking to being overtaken (even when travelling at the maximum legal limit) and that participants preferred travelling at 70mph over 60mph, but only when this information was presented visually. Experiment 2 gave occupants additional visual and auditory 'nudges' when travelling at the maximum legal limit, but this did not improve journey satisfaction when being overtaken at 70mph. Experiment 3 found that occupants preferred to follow another car compared to following a lorry, but only when they were overtaking. Participants also showed a preference for overtaking over being overtaken (as in Experiments 1 & 2), but only preferred overtaking when following another car, with no preferences for lane position while following a lorry.

One of the key findings from Experiments 1 and 2 – that being overtaken led to decreased satisfaction, even when travelling at the maximum legal limit – is potentially concerning for integrating autonomous vehicles into human-driven traffic. This is because the occupants' satisfaction with the journey was reduced by the behaviour of other vehicles (over which the autonomous vehicle has no control), even when the behaviour of the occupants' own vehicle was legal and safe. This finding is in line with survey-based work that has suggested motorists are concerned that human-driven vehicles could take advantage of law-abiding, rule-based autonomous vehicles (Madigan, et al., 2019; Tennant, 2015; Tennant, et al., 2016). While the individual drivers in this case are not directly taking advantage of the autonomous vehicle, the traffic context (other vehicles travelling at 70mph in the left lane with the right lane traffic travelling above the limit) provides human drivers with frequent opportunities to overtake, resulting in reduced occupant satisfaction. The influence of other vehicles' unsafe behaviour on evaluations of autonomous vehicles was also observed in Chapter 3 (Ritchie et al., 2019), in which sharper overtakes were rated more positively (and longer pull-ins more negatively) when the autonomous vehicle was being tailgated by another driver.

Furthermore, human drivers frequently overtaking rule-abiding autonomous vehicles could become an issue for journey satisfaction even in cases where human drivers are not in violation of the speed limit. This is because speedometers tend to have a degree of error, often over-reading the vehicle's actual speed by a small amount. In the United Kingdom, speedometers are not permitted to under-read the vehicle's true speed and can legally over-read by a maximum of 10% of the vehicle's true speed + 6.25mph (*The Motor Vehicles (Approval) Regulations*, 2001). A recent survey of the speedometer accuracy of several different vehicles showed that over-reading the vehicle's true speed by 3mph is common ("UK speed camera tolerances revealed: is your car's speedo accurate?", 24/04/2019). For example, if a vehicle's true speed is 70mph, then the maximum legal indicated speed would be 83.25mph (70mph + 10% (7mph) + 6.25mph), with a likely indicated speed of around 73mph. If we assume autonomous vehicle speedometers would have a similar degree of error; this could cause some issues for overtaking. For example, if a human driver were travelling at 70mph in the overtaking lane, their speedometer might indicate 73mph, but the driver might estimate their actual speed as 70mph because of their knowledge of the built-in error. In the case of autonomous vehicles, if a manufacturer wished to avoid the speedometer suggesting the vehicle was driving over the limit (to reduce occupant anxiety), then a speedometer with the same degree of error might display the indicated speed as 70mph when the autonomous

vehicle is actually travelling at 67mph. This could result in autonomous vehicles frequently being overtaken by human drivers travelling at the legal limit, with occupants possibly perceiving that they are being overtaken while driving at the legal limit (if they are not aware of the speedometer error, for instance).

Another key finding from Experiments 1 and 2 is the inconsistent influence of presenting more information to occupants. Participants presented with speed information (their vehicle's own speed relative to the speed limit) in Experiment 1 were happier overall with their journeys compared to those with no information, with follow-up analyses revealing that this positive effect of information was limited to journeys where participants were travelling at the legal limit. This suggests that, in some cases, presenting occupants with more vehicle information is beneficial, but that it may be preferable for autonomous vehicles not to display speed information in some cases (such as when the vehicle is travelling below the speed limit and the occupant is unable to intervene).

In relation to the 'transfer effects' of other drivers' illegal behaviour observed in Experiment 1, Experiment 2 aimed to improve satisfaction in journeys where participants were travelling at the maximum limit but still being overtaken. This was done by presenting participants with a voice message reminding them when they were travelling at the maximum legal limit, and a green glow around the dashboard, with the aim of making the fact the vehicle was travelling as quickly as legally possible more salient. This manipulation was added as a behavioural 'nudge', which are small changes to the environment that have been shown to influence decision making (Thaler & Sunstein, 2008). In terms of application to autonomous vehicles, the auditory and visual nudges used were added as a potentially simple and cost-effective method of reducing occupant frustration while being overtaken (but unable to legally increase speed). Presentation of simple visual cues has been suggested as a method of improving driver behaviour, for example by presenting information on other drivers' fuel consumption (Rakotonirainy, Schroeter & Soro, 2014).

However, this manipulation did not appear to be successful, with no interactions between lane position and speed observed, and no overall effect of information observed in Experiment 2 (unlike in Experiment 1, where participants presented with information were happier overall when travelling at 70mph). One possibility is that the frequency of auditory prompt in the 70mph journeys (once every 1km, for 6km journeys) may have been too frequent (some participants reported becoming frustrated with the reminders), and therefore led to a decrease in satisfaction in journeys where participants were travelling at the maximum limit. This may have brought journey satisfaction down to the levels observed in

journeys where the participant was travelling below the legal limit, resulting in no effect of information presentation. Future research could benefit from exploring the frequency of information presentation in relation to satisfaction with autonomous highway journeys.

Experiment 3 examined the effect of traffic context on journey satisfaction, with occupants either following another car or a lorry and either overtaking or being overtaken. Occupants were indifferent to the type of lead vehicle when being overtaken, suggesting that any potential nuanced effects of traffic context may have been diluted by the more substantial effect of being overtaken on frustration levels (similar to how participants in Experiments 1 and 2 were less satisfied when being overtaken, even when travelling at 70mph). Conversely, the finding that participants preferred following another car when overtaking suggests that following a large, slow-moving vehicle leads to frustration, even when driving in the 'fast lane' and passing other vehicles frequently. Participants also preferred overtaking over being overtaken when following another car, but this preference disappeared when following a lorry. This suggests that, when following a larger, slow-moving vehicle, the positive effect of overtaking other traffic is neutralised by the frustration generated by the traffic context.

These findings could be explained by occupants' expectations that they should be overtaking, rather than following, larger, often slow-moving vehicles (National Office of Statistics, Table SPE0111: Free flow vehicle speeds by road type and vehicle type in Great Britain, 2018). Following slower vehicles has been linked to higher levels of frustration and self-reported intentions to perform an overtake (Kinnear et al., 2015). The violation of this expectation may have led to increased frustration, cancelling out any positive influence of travelling in the 'fast lane'. Taken together, the findings of Experiment 3 suggest that occupants of autonomous vehicles may place different levels of weight on traffic context (type of vehicle) and rank position amongst other traffic (overtaking vs being overtaken). These findings have potential implications for how to integrate autonomous vehicles into human-driven traffic. For instance, in the case of platooning (Hjälmdahl, Krupenia & Thorslund, 2017; Lu, Li & Huang, 2017), occupants of autonomous vehicles may become more frustrated when following platoons of larger vehicles, compared to when following other cars, suggesting that separate lanes may be preferable in some mixed traffic systems.

Across all three experiments, participants were asked to estimate the amount of time each journey took, as well as either reporting or estimating the speed at which their vehicle was travelling (depending on whether this was presented on the dashboard). In both Experiments 1 and 2, slower journeys were reported to take longer, with no effects of information presentation or lane position, and no effects of the traffic variables in Experiment

3. Overall, the estimated journey time data showed that participants in Experiments 1 and 2 correctly reported 60mph journeys were slower than 70mph journeys, and also reported that journeys in which they were being overtaken were slower than journeys in which they were overtaking. However, follow-up tests on the effects of speed information revealed that the effects of lane position (being overtaken vs overtaking) on estimated journey speed were larger for participants not presented with speed information. This suggests that, in the absence of feedback on their vehicle's speed, occupants of an autonomous vehicle appear to use rank-based cues like lane position to estimate their speed. In line with the results for journey satisfaction ratings, this could be an issue for occupants of autonomous vehicles that are travelling at the legal limit but being overtaken, as journeys could feel slower and potentially more frustrating for occupants not presented with speed information.

The lack of an effect of lane position on estimated journey time is surprising given previous findings that negative emotional states can lead to overestimation of the duration of tasks (Angrilli, Cherubini, Pavese & Manfredini, 1997; Tipples, 2008). Previous research in the domain of public transport has also shown that travellers frequently overestimate the duration of journeys (Meng, Rau & Mahardhika, 2018). Given that journeys in which participants were being overtaken were rated more negatively than when they were overtaking across all three studies, participants would be expected to report the 'being overtaken' journeys as taking more time than the 'overtaking' journeys, although this was not found. The lack of effects in the present investigation may be due to the fact that the objective journey time for all of the journeys was fairly short (either 3:15 or 3:45), particularly for highway journeys. In future research, providing participants with longer overall journey times may give time estimation differences between different journeys more time to appear.

Experiment 1 utilised physiological data – frequency of Skin Conductance Responses (SCRs) and overall Skin Conductance Level (SCL) – as secondary, objective measures of levels of journey satisfaction. Some previous research has linked physiological measurements such as these to threating stimuli (Dewe et al., 2016) and stress levels during driving tasks (Daviaux et al., 2020: Although, Beggiato, Hartwich & Krems (2019) did not find EDA to be sensitive to driving discomfort). It was anticipated that participants may have had higher levels of reactivity for more stressful journeys. Alternatively, participants may have also showed lower levels of reactivity for journeys that were less interesting.

There were no effects of any of the traffic variables on overall SCL, suggesting that tonic levels of activity were not an effective measure of journey satisfaction in Experiment 1. Overall, there was little effect of any of the traffic variables on the frequency of SCRs, with

the exception of a three-way interaction between lane position, speed information, with two key findings. First, when participants presented with speed information were travelling below the legal limit (60mph), they demonstrated more SCRs per minute when overtaking, compared to being overtaken. This could either be due to a higher level of excitement when overtaking. Second, participants who were presented with speed information and overtaking showed few SCRs when travelling at 70mph, compared to 60mph. This could potentially be due to a higher level of surprise when travelling below the legal limit even when overtaking, or perhaps a higher level of frustration. Although the use of physiological measurements revealed some interesting effects in Experiment 1, it was not carried forward for several reasons: i) Experiment 2 used vocal prompts from the vehicle which may have elicited SCRs, which may have been indistinguishable from those relating to satisfaction; ii) overall, the physiological measures were not sufficiently sensitive to the traffic variables used to make them a reliable secondary measure of journey satisfaction.

Similar to the experiments presented in Chapter 3, the modified VCQ scale showed a good level of internal consistency as a single measure of autonomous journey evaluations. However, future research with a focus on different aspects of vehicle character and trust could explore potential dissociations in relation to autonomous journeys. For example, what trade-offs do occupants make between perceived safety and excitement when choosing between different speed profiles? While beyond the scope of the present work, this would be an interesting direction for future research.

4.6. Summary and conclusions

Overall, the present study found that occupants of an autonomous vehicle were less satisfied when being overtaken by other traffic, even in situations when the overtaking vehicles were exceeding the maximum legal limit. Participants in Experiment 1 were happier when provided with speed information, but Experiment 2 demonstrated that presenting more information may not always be beneficial to occupants (and may reduce satisfaction in some cases). Some sensitivity to traffic context was observed in Experiment 3, in which occupants only preferred overtaking when following another car, with no preference for lane position when following a lorry (potentially due to increased overall frustration when following a large, slow-moving vehicle). Overall, these findings suggest that occupant satisfaction with autonomous vehicle journeys is dependent on a variety of interacting factors both inside (information presentation, vehicle speed, lane position) and outside of the vehicle's control (other vehicles exceeding the speed limit, the type of lead vehicle present). These findings

provide implications for both the development and deployment of autonomous vehicles. In order to maximise occupant satisfaction, autonomous vehicles/systems may benefit from tailoring the presentation of information to the occupant based on the specific traffic context, although occupant frustration may be difficult to avoid in mixed traffic environments.

The present investigation adds to previous research by exploring interactions between driving styles, the traffic environment and information presentation on occupant satisfaction with autonomous highway journeys. However, the experiments reported in this chapter are somewhat limited by being conducted in a driving simulator. While this provides a reasonably immersive environment (more so than 'driver's eye' videos: Ritchie et al., 2019) for testing scenarios that would be difficult to test in real life, future research would benefit from exploring the effects of rank position amongst traffic and information presentation in real-world highway scenarios.

5. Influence of accuracy, rank position and task difficulty on the evaluation of context-neutral instrument display systems

Abstract

Previous research has identified factors such as system accuracy and the type of information presented to users as important for trust in computerised systems. However, that work has tended to focus on systems that users have had previous experience with/expectations of, potentially confounding the influence of prior experience and the effect of the system's behaviour on trust. This chapter presents nine experiments which used a relatively context-neutral methodology, focusing on the effects of accuracy, rank position, task difficulty, and distance from correct solution on system trust. Most experiments found a steep drop in trust quickly after accuracy fell below 100%. This pattern was not influenced by rank position or task difficulty, but users were slightly more forgiving of less accurate systems that got closer to the correct solution. Similar to those of Chapters 3 and 4, these findings suggest a large influence of some central factors (here accuracy on system trust) on users' trust of systems, with a limited degree of sensitivity to other contextual factors.

5.1. Introduction

In the domain of computerised systems, 'trust' has been operationalised as *"a positive expectation regarding the behavior of somebody or something in a situation that entails risk to the trusting party"* (Marsh & Dibben, 2003, pp. 470). System trust has been further divided into *dispositional* (based on characteristics of the user), *situational* (based on the systems' behaviour and the context) and *learned* trust (built up over experience with the system: Hoff & Bashir, 2015; Marsh & Dibben, 2003). Previous research has also distinguished between factors relating to the human user, the automation itself, and the environment, with automation-related factors displaying the largest impact on trust (Hancock et al., 2011; Schaefer, Chen, Szalma & Hancock, 2016). Factors influencing user trust in computerised systems are important to understand because levels of trust predict usage intentions (Choi & Ji, 2015), with poor calibration of trust linked to both over- and under-use of automated systems (Parasuraman & Riley, 1997).

Prior work into system trust has shown some influence of the appearance of systems/computerised agents on user trust. For instance, robotic driving agents have been shown to be trusted more if they display a more anthropomorphised (human-like) appearance, compared to a more robotic appearance (Lee, Kim, Lee & Shin, 2015). A simulator-based study found that occupants reported an autonomous vehicle with a more anthropomorphised appearance to be more trustworthy than one with a less human-like appearance (Waytz, Heafner & Epley, 2014). Participants in the Waytz et al. (2014) study also placed less blame on the more anthropomorphised vehicle when an accident occurred, compared to a less human-like autonomous vehicle (although both types of autonomous vehicle received more blame than if the participants were controlling a human-driven vehicle). In the medical domain, one study gave younger and older adults a series of health-related questions to answer, either with no decision aid, an anthropomorphised decision aid or a non-anthropomorphised decision aid (Pak, Fink, Price, Bass & Sturre, 2012). Pak et al. (2012) found that younger adults were more trusting of the anthropomorphised system, with there being no effect of anthropomorphism on older adults' trust levels, suggesting a degree of interaction between user (here age) and system characteristics on system trust. While not relating to anthropomorphism, one driving simulator study found higher levels of participant trust for navigation systems that had a more realistic and detailed display, compared to those that displayed the road network in a simplified manner (Barthou, Kemeny, Reymond, Mérienne & Berthoz, 2010).

Previous research has also shown that the level/nature of information provided by computerised systems can influence trust. For example, simulator-based studies in the vehicle

domain have found occupants' level of trust in an autonomous vehicle to be influenced by feedback on the systems' level of uncertainty (Beller, Heesen & Vollrath, 2013; Helldin, Falkman, Riveiro & Davidsson, 2013). Verberne, Ham and Midden (2012) provided participants with picture and text-based descriptions of the goals and automation levels of Adaptive Cruise Control (ACC) systems. Verberne et al. (2012) found that ACC systems that provided more detailed information were rated more positively than those that provided less information, and that participants preferred systems that were described as sharing their own driving goals (described by presenting participants with pictures showing how the system ranked goals such as comfort, safety, etc.). Another study concerned with information presentation found that in-vehicle agents that gave more instructive information were rated more positively than those that simply provided information about traffic conditions, but only in light traffic, with no effect of information in heavy traffic (Cramer, Evers, Kemper & Wielinga, 2008). This suggests some degree of context-dependency in how the information provided by in-vehicle systems is evaluated by users.

McGuirl and Sarter (2006) provided aeroplane pilots with either static or continually updating feedback on the confidence of an automated decision aid assisting them to detect in-flight icing scenarios in a simulated flight task. McGuirl and Sarter (2006) found that pilots provided with continually updating feedback rated the system's accuracy as higher, and experienced fewer stalls than those given static feedback from the system. Verame, Constanza and Ramchurn (2016) investigated the impact of system confidence information on acceptance of a simulated handwriting recognition system. Participants could either blindly accept the systems' decision, delay responding to it, or review the systems' decision, and either complete the task manually or allow the system to do so. Additionally, participants were either given system confidence information or no information. Verame et al. (2016) found that participants given system confidence information accepted more system decisions and were more likely to allow the system to complete the task, particularly in situations where the systems' confidence was high. These results suggest that, in addition to self-reported trust, intentions to use computerised systems can be influenced by the information the system presents.

In a laboratory-based study on system-generated explanations, Lim, Dey and Avrahami (2009) had participants interact with decision tree-style systems which aimed to predict a binary state (whether or not someone was exercising, based on factors such as temperature and heart rate). Participants were presented with boxes containing numerical values for input variables (temperature, heart rate, motion), which fed into a box containing the output (exercising vs. not exercising). Participants were presented with trials in which one of the

input/output boxes was empty and were required to fill in any information that was missing, as well as giving written answers on their understanding of how the system worked. The explanations given by the system were also manipulated: The key contrast was whether the system's account explained why it reached a particular decision (*why* explanations), or instead explained why it did not reach a different decision (*why not* explanations). Participants' answers were more accurate when given explanations on the systems' decision (compared to no explanation), although participants given feedback on why the system made a particular decision gave more complete answers on the workings of the system and trusted it more, compared to systems that gave feedback on why they did not make a different decision. Lim et al.'s (2009) findings were similar when repeating the study with a more abstract methodology, in which the system's goal was to predict a state of either "a" or "b", based on the values of three input variables ("A", "B", and "C"). Participants who were given 'why' explanations gave more accurate answers and rated the system as more trustworthy, compared to those given 'why not' explanations. Overall, these results suggest that the type of feedback given by computerised systems (in addition to the amount given) appears to be important for influencing user trust.

Another interesting finding from Lim et al. (2009) is that, when participants were asked to give subjective reports on their understanding of how the systems worked (their first experiment), participants' pre-existing knowledge about variables such as body temperature interfered with their ability to understand how the system predicted states of exercise vs. no exercise, prompting a second experiment with a more abstract, context-free methodology. This suggests that user evaluations of computerised systems may be biased by their previous knowledge/experience, and that more abstract methodologies may be helpful for measuring the influence of system behaviour on user trust in a way that minimises such confounds.

The behaviour and performance of computerised systems has also been shown to influence user trust. For instance, Madhavan and Phillips (2010) gave participants a task in which they were assisted by a simulated airline baggage-checking system, which was either 70 or 90% accurate at the task. They found that the 90% accurate systems were rated as more trustworthy (compared to 70%), but also that participants with higher levels of computer self-efficacy (trust in their own ability to use computers) displayed both higher levels of trust and performed better at the task when using the 90% accurate system. This suggests that system performance factors such as reliability can interact with characteristics of the user to influence trust. Chancey, Proaps and Bliss (2013) investigated the influence of system accuracy (20 vs 40%) on user trust and frequency of responses to an alarm system. Chancey et al. (2013) found

that, although alarm systems with higher accuracy were rated as more trustworthy than those with lower accuracy, participants' ratings of trust did not mediate the relationship between system accuracy and participants' frequency of responding to the alarms (measured by how often participants agreed with the system's decision), suggesting that subjective trust may have a limited influence of some behavioural aspects of system use in some contexts. One study focusing on use of luggage screening assistance systems found no overall effects of system accuracy on trust, but did find that participants were more accurate at a screening task when using a highly accurate system, and made more errors at identifying targets when the system provided cues that appeared more plausible in the event of a false alarm (Chavaillaz, Schwaninger, Michel & Sauer, 2020).

In addition to studies with static levels of accuracy, researchers have also investigated the effects of changes in system accuracy on user trust. For instance, Chavaillaz, Wastell and Sauer (2016) had participants complete a fault-identification task with the aid of an assistance system that had an accuracy of either 100, 80 or 60%, after an initial training session in which the system was 100% accurate. The level of assistance provided by the system could be manually altered by the participant throughout the task. Chavaillaz et al. (2016) found larger drops in trust with lower levels of accuracy, and that participants were more prone to errors when using less accurate systems. However, there was no influence of system accuracy on the level of automation participants elected to use, suggesting that trust does not always have a direct impact on some other aspects of system interaction (in line with Chancey et al., 2013). Similarly, Chavaillaz and Sauer (2017) gave participants either a highly reliable system (100% accuracy) or a less reliable system (60% accuracy) during an initial training phase, and participants then used a system with either 100% or 60% accuracy and provided their level of trust for that system. There was a marginal interaction between accuracy during training and during the main phase of the experiment, suggesting that the lowest levels of trust were found when participants used the low-accuracy system during both training and testing, with overall trust also being lower for the less accurate system independent of training. This suggests some influence of the system's behaviour when learning to use the system on later levels of trust.

Using a laboratory-based methodology, Yu et al. (2017) found a sharp drop in trust when accuracy dropped below 80% when participants rated their trust in a fault-checking system for glass manufacturing. Yu et al. (2017) also found that, over successive trials in which the system's decision was incorrect, trust dropped more sharply than it increased over successive trials in which the system was correct. This suggests an asymmetry in accuracy's effect on trust, with trust being lost more rapidly in poorly performing systems than it can be

gained when the system performs well. Kaltenbach and Dolgov (2017) investigated the interaction between reliability and transparency in an attempt to separate the influence of these factors on trust. Participants controlled a simulated coffee-manufacturing machine and were required to press a button to release hot water at a certain time. Participants were either given single or multi-line feedback on the systems' operations (low vs high transparency), and the system carried out the users' commands on either 65 or 95% of trials (low vs high reliability). Kaltenbach and Dolgov (2017) found that higher levels of transparency led to lower levels of trust when reliability was low, suggesting that giving users more feedback on system operations may not always be desirable. There was no overall effect of reliability on system trust, which the authors suggest could be due to the fact that participants could easily correct the systems that were less reliable. One question for future research is whether users would be more sensitive to reliability with fully autonomous systems which cannot be manually corrected, compared to tasks in which the system only assists the user in performing a task manually.

The work previously presented in Chapters 3 and 4 focuses on evaluations of autonomous vehicles based on the information presented by in-vehicle systems (Chapter 4, specifically), the autonomous vehicles' behaviour and factors relating to traffic context (Chapters 3 & 4). The work in these chapters is of a more applied nature and largely simulator-based, in order to study the impact of specific driving behaviours such as overtaking on trust. In contrast, the work presented in Chapter 5 takes a more 'context-neutral' approach, investigating the factors that influence participants' trust in novel systems using more abstract methodologies. This methodological approach was taken to determine the factors that influence trust while minimising the influence of participants' prior experience with specific technologies.

The current chapter presents nine laboratory-based experiments on the influence of system accuracy (all nine experiments) rank position (Exp 1a-c), task complexity (Exp 2a-b) and distance from the correct solution (Exp 3a-b; 4a-b).

5.2. Experiments 1a – 1c: Effect of accuracy and rank position on system trust

Experiments 1a-c had two key goals. The first of which was to examine the effect of accuracy on user trust when evaluating context-neutral systems, building on previous work showing that trust in specific technologies and interfaces is sensitive to accuracy (Chancey et al., 2013; Chavaillaz & Sauer, 2017; Chavaillaz, et al., 2016; Madhavan & Phillips, 2010; Yu et al., 2017). In order to investigate trust in novel systems while minimising any potential influence of prior experience with specific technologies, Experiments 1a-c presented

participants with systems that tried to generate a specific target colour (red). This was done because previous research has tended to focus on trust of specific existing technologies/systems, for which users likely have prior expectations and experience. Use of a more abstract, context-free methodology not only reduces these potential confounds, but may also shed light on factors influencing users' trust of completely novel systems and interfaces, which may be found in autonomous vehicles. The system had the ability to generate two possible colours: Red (correct) or a single alternative incorrect colour (blue, green, or yellow). The level of system accuracy (success at generating the target colour) was varied across conditions. Participants were asked to give a rating of their trust in the system after viewing a block of trials in which the system aimed to reach the target colour.

In addition to examining the effect of objective accuracy on trust of context-neutral systems, Experiments 1a-c also investigated the influence of rank position on system trust. The relative influence of objective value vs rank position has received considerable attention in the field of decision-making, with debate surrounding the extent to which objective value is represented in the brain (see Vlaev, Chater, Stewart & Brown, 2011, for a review). One position ('Decision by Sampling': Stewart, 2009; Stewart, Chater & Brown, 2006) argues that the value of any given option is always calculated by comparisons to all the other options available, with no internal representation of 'objective' value. In support of this approach, previous research in behavioural science has found people to be very sensitive to the ranking of options (Brown, Gardner, Oswald & Qian, 2008; Ert & Erev, 2013; Mullett & Tunney, 2013; Stewart, 2009; Stewart et al., 2006; Walasek & Stewart, 2015; Walasek & Stewart, 2019).

For instance, Walasek and Stewart (2015) presented participants with a series of 50-50 gambles and measured how often participants accepted them, varying the size and variability of the possible gains and losses to test the influence of relative ranking on loss aversion (the phenomenon whereby people are more sensitive to losses than gains of equivalent size: Kahneman & Tversky, 1979). Across four experiments, Walasek and Stewart (2015) found that the loss aversion effect was stronger when the range of possible losses was wider than the range of possible gains, but the typical the pattern of loss aversion was reversed when the range of gains was wider than the range of losses. Using a similar paradigm, Walasek and Stewart (2019) examined the effect of rank position more directly, by presenting participants with 50-50 gambles and varying the skew of the distributions from which the payoffs were sampled. Walasek and Stewart (2019) found that participants were most likely to accept gambles where the potential gains were selected from a positively skewed distribution and the potential losses came from a negatively skewed distribution. These findings suggest that people's perception

of risky decisions are sensitive to how the payoffs of a given decision rank against all other possible payoffs.

These findings suggest that relative rank position can exert powerful effects on human judgements and decision making. Experiments 1a-c aimed to test if the evaluation of computerised systems is also sensitive to rank position, and how this might interact with the effect of objective accuracy. For instance, are users of a less accurate system any more forgiving of the system if it is the best-performing option out of the available alternatives? In the context of the present study, 'rank position' refers to the position of a system's level of accuracy, in comparison to all of the other systems presented to participants within a given experiment (which was varied across experiments). Alternatively, evaluation of system trust might go against the typical findings that would be expected from relative rank theory. This would have implications for both relative rank theory and for predicting user trust in multi-system situations. Accordingly, Experiments 1a-c each used a different distribution of accuracies (Exp 1a: 100, 75 & 50%; Exp 1b: 75, 62.5 & 50%; Exp 1c: 100, 87.5 & 75%). In addition to within-subjects comparisons of trust over accuracy, a between-experiment analysis of all three experiments was performed to test the potential interaction between relative rank position and objective accuracy.

5.2.1. *Method*

Participants

There were thirty participants in each experiment, recruited via an online subject panel and given £3 or course credit for taking part. The sample sized used in this experiment and across Chapter 5 are comparable to similar studies on system trust (e.g. Barg-Walkow & Rogers, 2016; Chavaillaz et al., 2016; Madhavan & Phillips, 2010; Yu et al., 2017), and a small pilot with sixteen participants showed a substantial effect of system accuracy on trust. Demographic information for all nine experiments is available in Table 5.1. Participants were tested in groups of up to twenty-four in a large computer laboratory in sessions lasting up to 30 minutes.

Table 5.1. Demographic information for participants in all nine experiments presented in Chapter 5.

Experiment	Number of participants (number female)	Age range, mean and standard deviation	Handedness (A = Ambidextrous)
1a	30 (23)	18 – 25 years M = 18.90 *SD* = 1.40	29 RH 1 LH
1b	30 (24 female, 5 male)	18 – 20 years M = 18.76 *SD* = 0.64	27 RH 1 LH 1 A
1c	30 (20)	18 – 55 years M = 23.62 *SD* =7.14	26 RH 1 LH 3 A
2a	30 (24)	18 – 35 years M = 21.66 *SD* = 3.40	29 RH 1 LH
2b	30 (23)	19 – 33 years M = 21.33 *SD* = 3.40	26 RH 3 LH 1 A
3a	30 (11)	19 – 38 years M = 22.24 *SD* = 4.02	26 RH 4 LH
3b	30 (26)	18 – 20 years M = 18.61 *SD* = 0.63	28 RH 2 LH
4a	30(22)	18 – 31 years M = 19.6 *SD* = 2.49	26 RH 4 LH
4b	30(24, 5 male)	18 – 30 years M = 19.34 *SD* = 2.36	23 RH 6 LH

N.B: In cases where the figures do not add to the total sample of the experiment, this was due to participants either not providing some/all demographic information, writing their age illegibly or selecting the 'prefer not to say' option on the consent form.

Measures, stimuli, and apparatus

All displays were presented at 1920 × 1080px resolution on 57 (width) × 35cm (height) LCD computer monitors. System trust ratings were collected using a modified version (see Table 5.2) of a system trust scale developed by Jian, Bisantz and Drury (2000). The wording of some of the items was modified (as in several previous studies: Helldin, et

al., 2013; Rupp, Michaelis, McConnell & Smither, 2016; Rovira, Pak & McLaughlin, 2017; Verberne et al., 2012) to better reflect the nature of the systems participants experienced. Participants indicated their response (a number between 1 and 7) by using a mouse to click numbered buttons presented on the screen. Participants in Experiments 1a-c viewed two rectangles (one to the left of the centre of the screen, one to the right) on a black background. After each block of trials, participants completed the modified Jian et al. (2000) questionnaire by clicking circular numbered buttons to indicate their response. A schematic of the colour decision task is provided in *Figure 5.1*.

Table 5.2. System trust questionnaire adapted from Jian et al. (2000).

Question number	Question text
1 (R) *	The system did not provide me with all the information it could have
2 (R)	The system behaves in an underhanded manner
3 (R)	I am suspicious of the system's intent, action or outputs
4 (R)	I am wary of the system
5 (R)	The system's actions could have a negative outcome
6	I am confident in the system
7	The system provides security
8	The system has integrity
9	The system is dependable
10	The system is reliable
11	I can trust the system
12 *	I feel that I know how the system works and how it makes its decisions

Scores ranged from 1-7, with lower scores being more negative and higher scores being more positive. Each question began with “With 1 being 'Not at all', and 7 being 'Extremely', please rate the following statement in relation to the system in this block”.

“R” = Reverse-scored

“*” = Wording modified from original scale

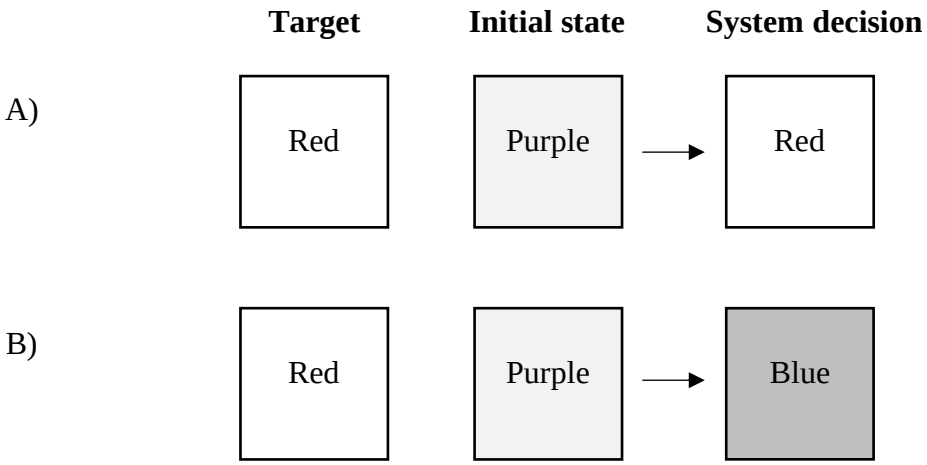


Figure 5.1. Schematic of the colour selection task used in Experiments 1a-c, 2a and 2b of Chapter 5. Panel A shows an example of a "correct" trial, whereas Panel B shows an example of an "incorrect" trial.

Design and procedure

Participants in Experiments 1a-c viewed a system that was aiming to generate a target colour and achieved this at varying levels of accuracy across different conditions. Participants viewed two rectangles on a black screen: The rectangle on the left indicated the target colour (i.e. the colour that the system *should* reach), and the rectangle on the right represented the system's decision. The levels of accuracy were 100, 75 and 50% in Experiment 1a, 75, 62.5 and 50% in Experiment 1b, and 100, 87.5 and 75% in Experiment 1c. Participants viewed one level of accuracy per block and completed three blocks of trials, with the order counterbalanced between-subjects. There were twenty-four trials in each block, with the order of 'correct' and 'incorrect' trials randomised. The target colour (which the system aimed to generate) in all three experiments was always red. The system's alternative (incorrect) colour choice was also fixed for each participant but randomized across participants (being either green, blue, or yellow). This resulted in each participant being assigned a 'colour pair' for the two options the system could choose across the entire experiment: Red-Blue, Red-Green, or Red-Yellow. Thus, on every trial the system had a choice of two options – the correct colour or a single alternative and fixed incorrect colour.

In each trial, participants viewed two rectangles on a black screen (100 × 100 pixels each, one to the left and one to the right (with 100 pixels of blank space between both rectangles). The system could either choose the correct colour or one incorrect colour on each trial, with the single incorrect colour option assigned between-subjects using the colour-pair

system described above. The RGB values used for each colour are included in parentheses. The rectangle on the left was always red (R: 255, G: 0, B: 0) in Experiments 1a-c and represented the target colour. The rectangle on the right started purple (R: 127, G: 0, B: 127) and instantly changed colour to either the correct colour (always red) or the incorrect colour, which was either blue (R: 0, G: 0, B: 255), green (R: 0, G: 255, B: 0), or yellow (R: 255, G: 255, B: 0) after four seconds. One second after the colour changed, text above both rectangles appeared in the middle of screen stating, "I have reached a decision". Two seconds after this, the rectangles and text disappeared, and participants were required to click on buttons marked "Yes" or "No" to indicate if the system had made the correct decision. When the participant had made their response, the screen went blank for two seconds before the next trial began. Participants were prompted to press the space bar after each block of trials to load the questionnaire, which they completed using mouse-click buttons, and then were prompted to press space again to load the next block of trials or to complete the experiment. *Figure 5.2* provides a schematic of the general procedure for Experiments 1a-c.

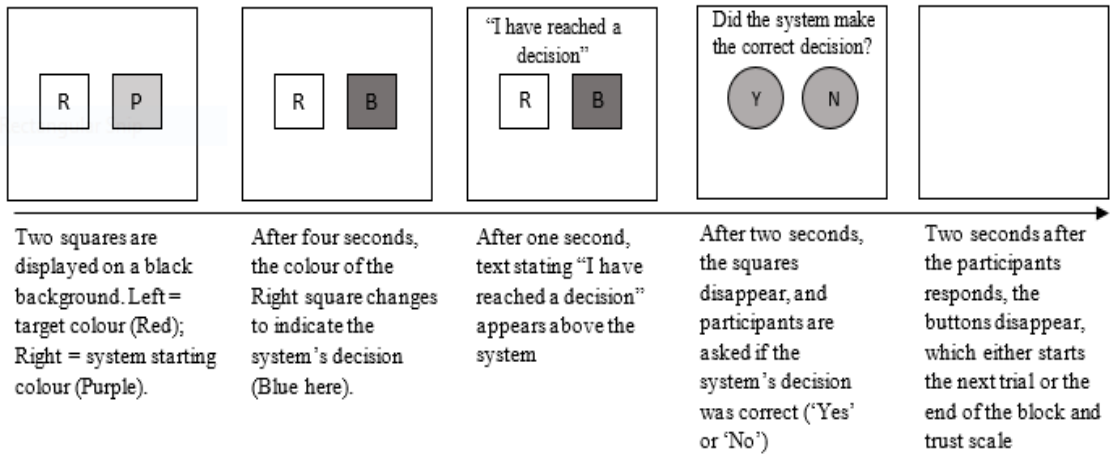


Figure 5.2. Schematic of a single trial in Experiments 1a-c and Experiments 2a-b.

5.2.2. Results

The main dependent measure of interest was participants' ratings of system trust as a function of how accurate the system was (see *Figure 5.3*). This was assessed via within-subjects one-way ANOVAs and associated post-hoc tests for each Experiment 1a-c. Where indicated Greenhouse-Geisser corrections for violations of sphericity were applied. For follow-up tests in which normality assumptions were violated, non-parametric tests were used where indicated.

Experiment 1a: Effect of accuracy on system trust

There was a significant main effect of accuracy on system trust, F (1.385, 40.153) = 23.845, $p < .001$, $\eta^2 = .451$. Bonferroni-corrected post-hoc tests revealed that ratings were significantly more positive in the 100% accuracy condition, compared to both the 75% and 50% conditions, ps < .001. Ratings in the 75% condition were significantly more positive than in the 50% condition, $p < .05$. A follow-up Wilcoxon's signed rank test revealed that the drop in ratings between 100 and 75% accuracy was significantly larger than the drop between 75 and 50%, $W = 464$, $p < .001$.

Experiment 1b: Effect of accuracy on system trust

There was no significant effect of accuracy on system trust, F (1.307, 37.891) =1.663, $p = .207$, $\eta^2 = .054$.

Experiment 1c: Effect of accuracy on system trust

There was a significant main effect of accuracy on system trust, F (1.321, 38.320) = 24.068, $p < .001$, $\eta^2 = .454$. Bonferroni-corrected post-hoc tests revealed that ratings were significantly more positive in the 100% accuracy condition, compared to both the 87.5% and 75% conditions, ps < .001. There was no significant difference in ratings between the 87.5 and 75% conditions, $p = .327$. Further follow-ups revealed that the drop in ratings between 100 and 87.5% accuracy was significantly larger than the drop between 87.5 and 75%, t (29) = 3.798, $p < .001$.

Combined analysis on effect of rank position

The datasets from Experiments 1a-c were combined to examine the influence of rank position (relative accuracy of a system, in relation to the performance of the other two systems presented) and absolute accuracy on system trust ratings. Rank position was based on the relative level of accuracy for each system across the three studies. That is: rank position 1 relates to the 100%, 75% and 100% accuracy levels of Experiment 1a, 1b and 1c, respectively. Similarly, rank position 2 corresponds to accuracy levels of 75% (1a), 62.5% (1b) and 87.5% (1c), and rank position 3 corresponds to 50% (1a and 1b) and 75% (1c). To give a specific example, the 75% accuracy condition represented the second-best performing system in Experiment 1a, whereas 75% was the highest level of accuracy in Experiment 1b, and the lowest in Experiment 1c. Ratings were analysed using a 3 (Rank position: 1st, 2nd, 3rd) × 3 (Experiment: 1a, 1b, 1c) mixed ANOVA with Experiment as the between-subjects factor. There was a significant main effect of rank position on ratings, F (1.402, 121.985) = 46.786, $p < .001$, $\eta^2 = .313$. Bonferroni-corrected post-hoc tests revealed that systems ranked 1st were rated more positively than those ranked 2nd or 3rd, ps < .001, and systems ranked 2nd were rated more positively than those ranked 3rd, $p < .01$. There was a significant main effect of

experiment on system trust ratings, F (2, 87) = 5.567, $p < .01$, $\eta^2 = .113$. Bonferroni-corrected post-hoc tests revealed no significant difference in overall ratings between Experiments 1a and 1b ($p = 1.000$), however, overall ratings were higher in Experiment 1c, compared to both Experiments 1a and 1b, ps < .05. There was also a significant rank position × experiment interaction, F (2.804, 121.985) = 7.858, $p < .001$, $\eta^2 = .105$. Simple main effects analyses revealed that there was a significant main effect of rank position in Experiments 1a and 1c (ps < .001), but not in Experiment 1b ($p = .129$). Furthermore, there was only a significant main effect of experiment when comparing systems ranked 1st ($p < .001$). There was no difference across experiments when comparing systems ranked 2nd ($p = .210$), and a borderline effect of experiment when comparing systems ranked 3rd ($p = .052$).

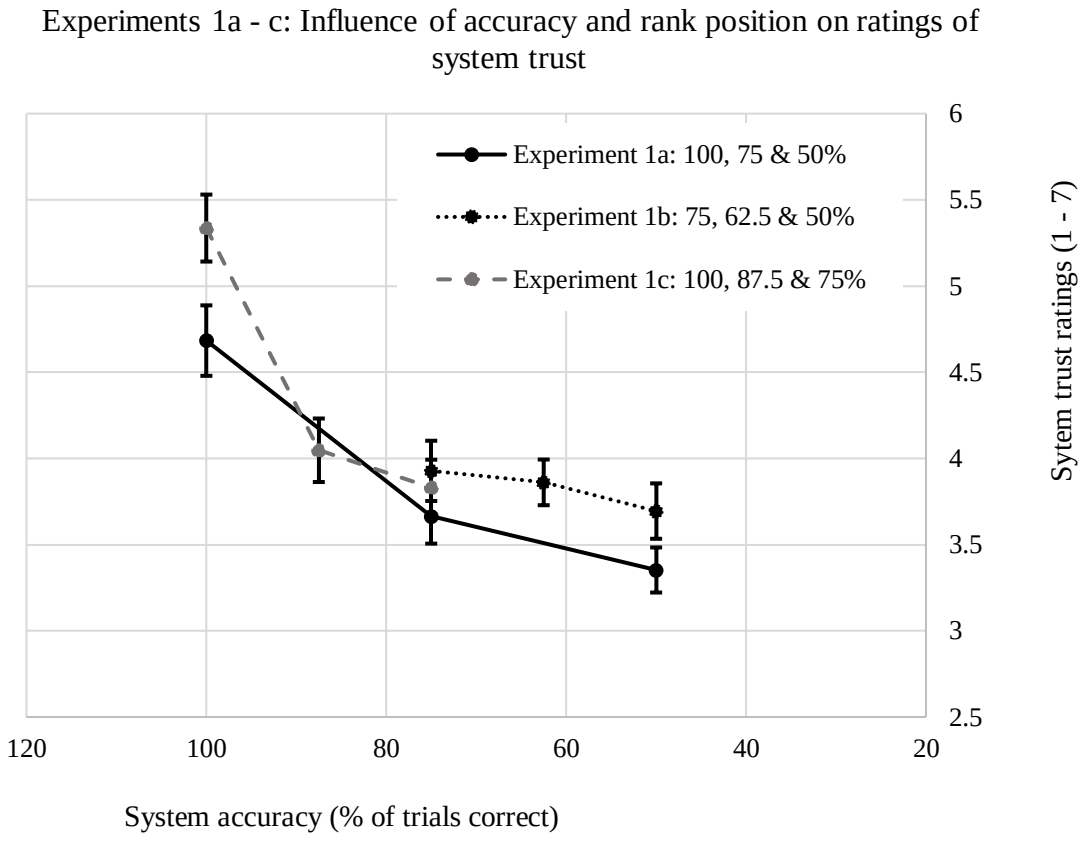


Figure 5.3. Trust questionnaire ratings for Experiments 1a-c: Influence of accuracy and rank position. Experiment 1a used accuracies of 100, 75 and 50%. Experiment 1b used accuracies of 75, 62.5 and 50%. Experiment 1c used accuracies of 100, 87.5 and 75%. N.B: The system trust scale provided an average score between 1 and 7, however the axis below is scaled to make the pattern of results for each individual experiment clearer to read. Error bars represent ±1 standard error of the mean.

5.2.3. Discussion

Experiments 1a – c investigated the influence of accuracy and rank position on trust of a context-neutral system attempting to generate a target colour. In the context of the present study, 'rank position' refers to the position of the system's accuracy in relation to all the other systems the participant views across the whole experiment. Overall, there was a sharp drop in trust when accuracy dropped below 100% (between 100 and 75% in Experiment 1a and between 100 and 87.5% in Experiment 1c), with a much flatter drop between 75 and 50%. There was no influence of accuracy on system trust in Experiment 1b (using accuracies between 75 and 50%), suggesting that participants were sensitive to absolute accuracy and not merely evaluating systems on rank position. This suggestion is further evidenced by the interaction between rank position ($1^{st}/2^{nd}/3^{rd}$) and distribution of accuracies (Experiment 1a/b/c). If participants were basing their trust merely based on ranking the three systems that they saw in the experimental session (based on accuracy) then the interaction should not have occurred. That is, the most accurate system should have been assigned a similar high trust value (e.g., 6 out of 7) irrespective of its absolute level of accuracy, resulting in three parallel lines across the three experiments, with the same pattern occurring regardless of objective accuracy. A similar argument can be made for the middle and lowest ranked systems. The finding of the interaction tells us that trust not only depended on rank position but also on the absolute level of accuracy.

The findings also suggest that, in the case of the present study's system evaluation task, users were very sensitive to small decreases in accuracy below 100% performance (in the 87.5% accuracy condition for instance, the system generated the incorrect colour on only three of twenty-four trials), rather than trust decreasing linearly alongside accuracy. The strong influence of absolute accuracy on ratings of system trust provided evidence that users are sensitive to the performance of systems when very little (if any) real-world context is provided for the task (compared to evaluation of existing systems/technologies). The finding that rank position did not override the effect of absolute accuracy on trust is in strong contrast to previous research showing substantial effects of relative rank position on decision making (Brown et al., 2008; Ert & Erev, 2013; Mullett & Tunney, 2013; Stewart, 2009; Stewart, et al., 2006; Vlaev et al., 2011; Walasek & Stewart, 2015, 2019). The finding that participants' ratings of trust were not sensitive to rank (at least not over and above the effect of objective accuracy) is surprising given the central role of ranking in the judgement and decision-making literature surveyed above. With Experiments 1a-c finding that users were no more forgiving of less accurate systems that performed the best of the available alternatives,

Experiments 2a and 2b focused on the nature of the task and how this may influence perceptions of trust. Specifically, Experiments 2a and 2b investigated the effect of varying the number of targets and possible solutions on how system trust varied over accuracy.

5.3. Experiments 2a – 2b: Effect of size of problem space on system trust ratings

In Experiments 1a-c the main aim was to determine: i) the level of people's trust of a system as a function of how accurate the system was, and ii) whether trust was based on absolute judgements of accuracy or was characterised by rank-based considerations. However, across all three experiments the system always had a single target colour (in this case, red) and when it was inaccurate it chose a single alternative colour (e.g., blue). Thus, the system and its associated problem space were relatively simple in that there was only a single target and only two choices to make (the correct one or a single incorrect alternative). In this sense we might expect people to be relatively less forgiving of errors compared to situations in which the problem space and level of complexity is higher. To take a real-world example, selecting the appropriate cabin temperature of a vehicle might be seen as a relatively simple task. There is only one temperature to be determined and there are probably relatively few factors that influence was the temperature should be set at. In contrast, destination prediction is a considerably more complex process with a greater number of solutions and a greater number of factors that could influence the prediction. Is peoples' trust in a system influenced by the complexity of the problem that the system is trying to solve?

Accordingly, Experiments 2a and b explored trust of systems in which the problem space was larger and more complex than in Experiments 1a-c. In Experiment 2a there was again a single target colour (red, as in Exp 1a-c) but on every trial there were four possible choices that system could make; red which was correct, or green, blue, and yellow, which were incorrect solutions. Thus, as in Experiments 1a-c there was only a single target but now the system had to choose from 4 rather than 2 possible options – hence its task was more complex with more room to make a potential error. In Experiment 2b the complexity of the problem was further increased. Rather than there being a fixed target colour (red) the target colour now varied on each trial from 4 possible alternatives (red, green, blue or yellow) and again there were four possible choices the system could make (one being correct and three being incorrect). If people are sensitive to the difficulty of a task that a system is trying to solve then we would expect higher values of trust to emerge than in Experiments 1a-c.

5.3.1. *Method*

Stimuli, measures, and apparatus

These were the same as in Experiments 1a – c.

Design and procedure

Participants in Experiments 2a-b viewed three blocks of twenty-four trials in which a system aimed to generate a target colour, with the same accuracy levels, counterbalancing of block order and overall procedure as those of Experiment 1a (100, 75 and 50% accuracy). The target colour was always red in Experiment 2a, and when the system was incorrect it chose either blue, green, or yellow. Thus, in the 50% accurate condition in a block of 24 trials the system chose 12 red, 4 blue, 4 green and 4 yellow. In Experiment 2b, rather than being fixed, the target colour was red, blue, green or yellow, each on one quarter of all trials (such that every block contained 6 trials where the target was red, 6 blue, 6 green, and 6 yellow). In each display the target colour was indicated by the colour of the left box and now changed over trials rather remaining fixed (red). In the 50% accuracy condition in Experiment 2b, three trials of each target colour were incorrect, with the system choosing all three non-target colours across these three trials (e.g. for the 6 trials where the target was red, the system would choose red for 3, blue for 1, green for 1 and yellow for 1). For the 75% condition, each participant was assigned two target colours for which the system would choose correctly on five out of six trials, and two for which the system chose correctly on four out of six trials. The six pairs of target colours where the system was correct for five out of six trials were randomly paired with the six block orders. The overall experiment lasted up to 30 minutes.

5.3.2. Results

Experiment 2a: Effect of accuracy and problem space on system trust

A one-way within-subjects ANOVA revealed a significant main effect of accuracy on system trust ratings (*Figure 5.4*), F (1.551, 44.986) = 14.036, $p < .001$, $\eta^2 = .326$. Bonferroni-corrected post-hoc tests revealed that ratings were significantly more positive in the 100% accuracy condition, compared to both the 75% ($p < .01$) and 50% conditions ($p < .001$). There was no difference in ratings between the 75 and 50% conditions, $p = .204$. Further follow-ups revealed that the drop in ratings between 100 and 75% accuracy was not significantly larger than the drop between 75 and 50%, t (29) = 1.857, $p = .073$.

Experiment 2b: Effect of accuracy and problem space on system trust

There was a significant main effect of accuracy on system trust ratings (*Figure 5.4*), F (1.439, 41.745) = 38.509, $p < .001$, $\eta^2 = .570$. Bonferroni-corrected post-hoc tests revealed that ratings were significantly more positive in the 100% accuracy condition, compared to both the 75 and 50% conditions, ps < .001. Ratings in the 75% condition were significantly

more positive than those in the 50% condition, $p < .05$. A follow-up Wilcoxon's signed-rank test revealed that the drop in ratings between 100 and 75% accuracy was significantly larger than the drop between 75 and 50%, $W = 340.500$, $p < .01$.

Combined analysis on effect of size of problem space

We combined the datasets from Experiment 1a, Experiment 2a and Experiment 2b to examine the influence of size of problem space (number of targets and solutions) and accuracy (100 vs 75 vs 50%) on system trust ratings. Trust ratings (see *Figure 5.4*) were analysed using a 3 (Accuracy: 100, 75, 50%) × 3 (Experiment: 1a, 2a, 2b) mixed ANOVA with Experiment as the between-subjects factor. There was a significant main effect of accuracy on ratings, F (1.490, 129.666) = 69.673, $p < .001$, $\eta^2 = .442$. Bonferroni-corrected post-hoc tests revealed that ratings were significantly more positive in the 100% accuracy condition, compared to both the 75 and 50% conditions (ps < .001). Ratings in the 75% condition were significantly more positive than those in the 50% condition, $p < .001$. Neither the main effect of experiment, F (2, 87) = 1.770, $p = .176$, $\eta^2 = .039$, nor the accuracy × experiment interaction, F (2.981, 129.666) = .428, $p = .732$, $\eta^2 = .005$, were significant.

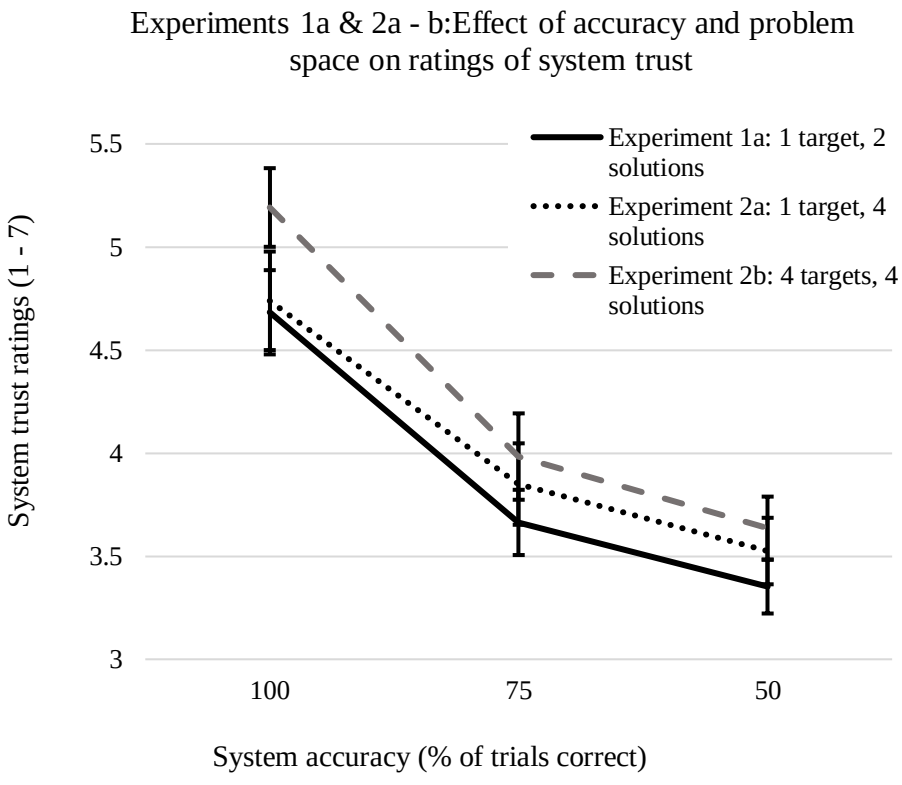


Figure 5.4. Comparison of trust questionnaire ratings for Experiments 1a (1 target, 2 choices), 2a (1 target, 4 choices) and 2b (4 targets, 4 choices): Influence of accuracy and size of problem space/task difficulty. Error bars represent ±1 standard error of the mean.

5.3.3. Discussion

Experiments 2a and 2b were designed to test the influence of task complexity on the pattern of system trust ratings over accuracy. Experiment 2a expanded the number of possible solutions (colours the system could choose) from two to four, while keeping the number of target colours at one, whereas systems in Experiment 2b could choose four potential colours and had to solve for four different targets. If users of context-neutral systems are sensitive to task complexity, it could be expected that participants would be more forgiving of less accurate systems attempting a more difficult task (i.e. one in which there are four possible targets). This would have led to an interaction between system accuracy and task complexity which increased over Experiments 1a and 2a-b. However, overall, the results of Experiments 2a and 2b were very similar to those of Experiment 1a, with system trust falling rapidly between 100 and 75% accuracy, with a shallower drop between 75 and 50%. There was no evidence to suggest that users were sensitive to the number of solutions or target colours, with no interaction between accuracy and experiment.

One account for the lack of an effect of complexity in Experiment 2a (although of course one has to be cautious interpreting null results) is that the greater variance in answers (four different responses rather than 2) might have led participants to consider the system as less intelligent. For example, in contrast to Experiment 1, the system in Experiment 2a 'got it wrong' in multiple different ways (e.g., choosing 3 different kinds of wrong answer). In contrast, in Experiment 1a the system only ever presented one wrong answer and so may have appeared more reliable. Thus, rather than participants perceiving the task as more difficult (because the system had to choose from a wider range of options) they may have perceived the system as less reliable because of the wider range of errors it made. However, this account cannot explain the lack of difference across the three experiments, because participants in Experiment 2b did not see the system generating colours which were never present as a target (unlike in Experiment 1a and 2a). The findings provide no evidence that participants consider the complexity of a task when considering how trustworthy a system is – at least within the current experimental context and tasks.

The experiments thus far used a colour generation task where the system attempted to generate the target colour. In order to test the generalizability of the findings, Experiments 3a-b and 4a-b examined the same levels of accuracy as Experiments 2a and 2b (100, 75 and 50%), but used a dial-based task in which the system attempted to reach a target angle. While participants were not provided with any context for what the dials were measuring (meaning the systems were still context-neutral), the dial-based task did enable testing of the effects of

accuracy in a slightly more 'real-world' domain. That is, people are generally much more familiar with systems that present (analog) information in a needle and dial format than systems that use colour to indicate values. In addition, the dial experiments were used to investigate the impact of distance from the correct solution (i.e. is the system 'almost there' when it makes the incorrect decision?) on ratings of system trust. This contrasts with Experiments 1 and 2 in which the solution was either correct or not (given the categorical nature of colour perception: Ozturk, Shayan, Liszkowski & Majid, 2013; Winawer et al., 2007) and so was more 'digital' in nature.

5.4. Experiments 3a & 3b: Effect of closeness to solution on system trust ratings

Experiments 3a and 3b replaced the coloured patches system with a dial-based system. Although Experiments 3a-b and 4a-b focused on user trust of systems which attempted to reach a target position on a dial (which is arguably less contextually neutral, compared to colour generation), the dials values did not represent any specific measurement (e.g. vehicle speed). Therefore, the dial task used in Experiments 3a-b and 4a-b was a contextually neutral spatial task. On each trial, participants were presented with a single dial that consisted of a pointer and a red dot on the edge of the dial window. The red dot indicated the target position that the system was supposed to indicate. In Experiment 3a, on each trial the system either reached the correct answer (aligning with the red dot) or was incorrect by fixed 90-degree angle. In Experiment 3b, when incorrect the system missed the target by 10 degrees. These manipulations were included to examine if the results of Experiments 1 and 2 generalised to a different, more familiar, presentation method. However, the approach also facilitated testing of the effects of 'closeness to solution' on levels of trust. That is, here accuracy could quite clearly differ on two dimensions. The first was how often the system was successful versus giving an incorrect answer (as in Experiments 1 and 2). However, in Experiment 3 there was also a clear indication of just how wrong the system was when it made an error (being 10deg off the solution or 90deg away from the solution). It is possible that trust depends both on the frequency of errors and on the size of the errors made.

5.4.1. *Method*

Stimuli, measures, and apparatus

The questionnaire measures and overall apparatus were the same as those used in Experiments 1a-2c. However, participants were presented with a dial centred in the middle of a black screen (with a diameter of 200 pixels), which consisted of a hollow circle with a white circumference. In the centre of the dial, there was a white line (100 pixels in length)

which acted as a ‘pointer’. A red dot on the circumference of the dial represented the target dial position for each trial. That is, the point that the system was supposed to move the needle to. A schematic is provided in *Figure 5.5.*

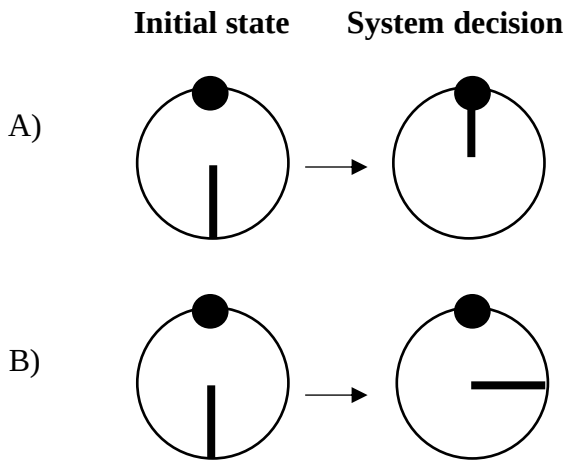


Figure 5.5. Schematic of the dial position task used in Experiments 3a-b and 4a-b. Panel A shows an example of a “correct” trial, whereas Panel B shows an example of an “incorrect” trial. The “target” dial position in Experiments 3a-b and 4a-b was always indicated by a red dot at the 12 o’clock position of the dial, with the dial hand always starting at the 6 o’clock position.

Design and procedure

The levels of accuracy, counterbalancing of block order, questionnaire measures and overall procedure in Experiments 3a-b were the same as those in Experiment 1a. Participants viewed a system which aimed to reach a target position on a dial display. The target position was always at 12 o’clock and was indicated by a red dot on the outside of the dial. The needle always started at the 6 o’clock position. When the system made the incorrect decision (in all three experiments), the needle finished either to the left or right of the correct solution (between-subjects, randomly paired with block order: For each participant, all three systems landed on either the left or the right side of the correct solution when incorrect). This was analogous to the assignment of ‘incorrect’ colours in Experiments 1a-c, such that each participant viewed a system that either chose the correct position or one other (incorrect) position (always 90 degrees in Exp 3a, always 10 in Exp 3b). Participants completed three blocks of twenty-four trials, with the trust scale administered after each block. At the start of

each trial, the dial appeared on the screen with the pointer facing downwards, and the red dot indicating the goal angle. After four seconds, the pointer moved instantaneously to indicate the system's decision. One second later, a message appeared above the dial stating, "I have reached a decision" (as in Exp 1 & 2). After two seconds, the dial disappeared, and participants were prompted to click 'Yes' or 'No' on buttons to indicate whether the system had made the correct decision in that trial. Two seconds after their response, the buttons disappeared, which either began the next trial or a prompt to load the trust scale for that block. A schematic of a single trial in Experiments 3 and 4 is provided in *Figure 5.6.*

In Experiment 3a, the needle always finished 90◦ away from the target position on 'incorrect' trials. In Experiment 3b, the needle finished 10◦ away from the target position. Each experiment lasted up to 30 minutes.

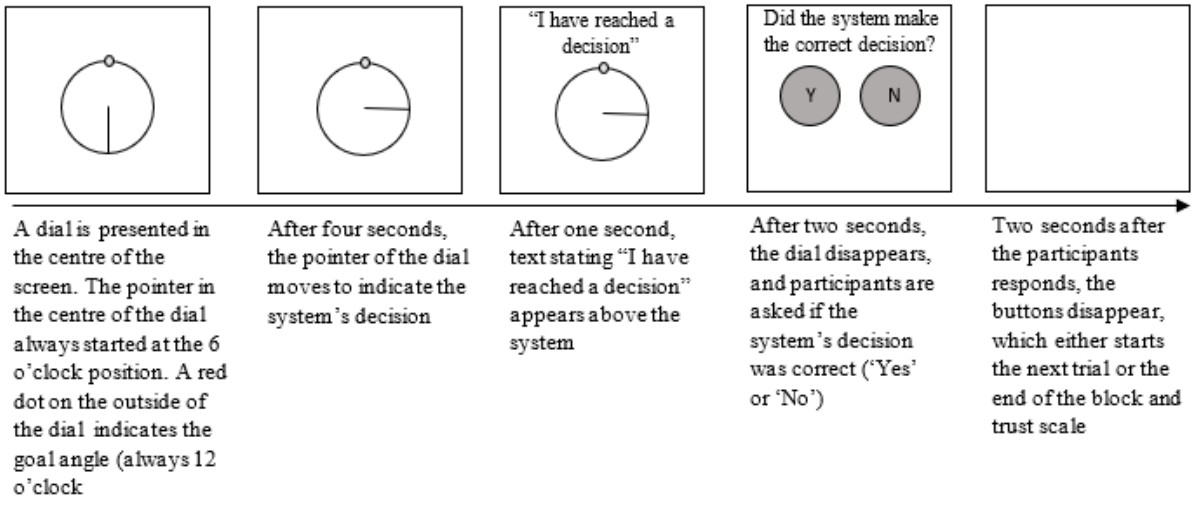


Figure 5.6. Schematic of a single trial in Experiments 3a-b & 4a-b.

5.4.2. Results

Experiment 3a: Influence of accuracy and closeness to solution on system trust

A one-way within-subjects ANOVA revealed a significant main effect of accuracy on system trust ratings, F (2, 58) = 44.674, $p < .001$, $\eta^2 = .606$ (*Figure 5.7*). Bonferroni-corrected post-hoc tests revealed that ratings were significantly more positive in the 100% accuracy condition, compared to both the 75 and 50% accuracy conditions, ps < .001. There was no difference in ratings between the 75 and 50% conditions, $p = .112$. Further follow-ups revealed that the drop in ratings between the 100 and 75% conditions was significantly larger than the drop between 75 and 50%, t (29) = 3.073, $p < .01$.

Experiment 3b: Influence of accuracy and closeness to solution on system trust

There was a significant main effect of accuracy on system trust ratings, F (2, 58) = 33.972, $p < .001$, $\eta^2 = .539$ (*Figure 5.7*). Bonferroni-correct post-hoc tests revealed that ratings were significantly more positive in the 100% condition, compared to both the 75 and 50% conditions, ps < .001. Ratings were significantly more positive in the 75% condition than in the 50% condition, $p < .01$. Further follow-ups revealed that the size of the drop in ratings between 100 and 75% accuracy did not differ from the drop between 75 and 50%, p = .202.

The results from Experiment 3a and 3b were compared with using a 2 (Closeness to solution: 90◦ vs 10◦) x 3 (Accuracy: 50, 75 or 100% correct) mixed ANOVA with accuracy as the within-subjects factor. This revealed that trust decreased as accuracy decreased, F (1, 116) = 77.309, $p < .001$, $\eta^2 = .321$. Post-hoc tests on the main effect of accuracy found that trust in the 100% accuracy condition was significantly higher than in the 75 and 50% conditions (ps < .001), and that trust was significantly higher in the 75% accuracy condition, compared to the 50% condition, $p < .01$. There was a significant accuracy x closeness to solution interaction, F (2, 116) = 4.763, $p < .05$, $\eta^2 = .020$. As shown in *Figure 5.7*, the interaction appears to be driven by a greater drop in trust between 100 and 75% accuracy for the 90-degree condition. Simple main effects revealed that the main effect of accuracy was significant across both levels of closeness to solution, ps < .001. There was no main effect of closeness to solution at 100 (p = .132), 75 (p = .200) or 50% accuracy (p = .214). A follow-up Mann-Whitney U test revealed that the drop in ratings between 100 and 75% accuracy was significantly larger for systems that were 90◦ away from the target (compared to 10◦), t (58) = 2.931, $p < .01$. There was no difference in the drop in ratings between 75 and 50% as a function of closeness to solution, U = 375.500, p = .273. The main effect of closeness to solution was not significant, F (1, 58) = .194, p = .661, $\eta^2 = .001$.

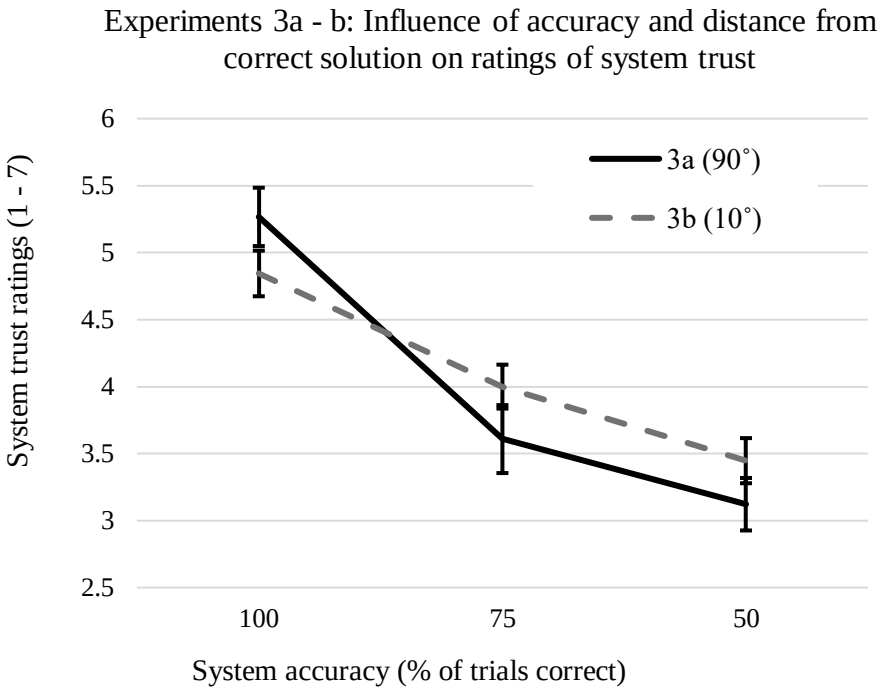


Figure 5.7. Trust questionnaire ratings for Experiments 3a (90˚ distance from correct solution) and 3b (10˚ distance from correct solution): Influence of accuracy and closeness to solution (between-study comparison). Error bars represent ±1 standard error of the mean.

5.5. Experiments 4a & 4b: Effect of closeness to solution on system trust ratings

Experiments 3a and 3b revealed that, when a system attempting to reach a specified point on a dial was closer to the correct solution when it made the wrong decision, participants' trust of the system dropped less severely with decreasing accuracy, compared to if the system missed by a larger amount. This suggests some degree of sensitivity to closeness to solution when making judgments of system trust, in conjunction with the effect of objective accuracy. Experiments 4a and 4b were conducted as a direct replication of Experiments 3a and 3b (respectively), and were therefore identical in terms of stimuli, design, and procedure. The aim of Experiments 4a and 4b was to replicate the interaction between accuracy and distance from solution revealed by Experiments 3a and 3b, in order to test whether the observed relationship between accuracy and closeness to solution was reliable. The reason this experiment in particular was replicated is because it was the first and only experiment in this chapter to show an interaction between accuracy and any other aspects of system performance. In contrast, the previous experiments across this chapter showed no influence of variables such as rank position and task complexity on the relationship between

accuracy and system trust. It was therefore important to strengthen this interesting new finding from Experiments 3a-b.

5.5.1. *Method*

Stimuli, measures, and apparatus

In Experiment 4a, these were identical to Experiment 3a. In Experiment 4b, these were identical to Experiment 3b.

Design and procedure

The design and procedure of Experiment 4a was identical to Experiment 3a. The design and procedure of Experiment 4b was identical to Experiment 3b.

5.5.2. Results

Experiment 4a: Influence of accuracy and closeness to solution on system trust

A one-way within-subjects ANOVA revealed a significant main effect of accuracy on system trust ratings (*Figure 5.8*), F (1.447, 41.954) = 28.017, $p < .001$, $\eta^2 = .491$. Bonferroni-corrected post-hoc tests revealed that ratings were significantly more positive in the 100% accuracy condition, compared to both the 75 and 50% accuracy conditions, ps < .001. There was no difference in ratings between the 75 and 50% conditions, $p = .264$. Further follow-ups revealed that the drop in ratings between the 100 and 75% conditions was significantly larger than the drop between 75 and 50%, t (29) = 2.969, $p < .01$.

Experiment 4b: Influence of accuracy and closeness to solution on system trust

A one-way within-subjects ANOVA revealed a significant main effect of accuracy on system trust ratings (*Figure 5.8*), F (2, 58) = 20.798, $p < .001$, $\eta^2 = .418$. Bonferroni-corrected post-hoc tests revealed that ratings were significantly more positive in the 100% accuracy condition, compared to both the 75 and 50% accuracy conditions, ps < .001. There was no difference in ratings between the 75 and 50% conditions, $p = .263$. Further follow-ups revealed that the drop in ratings between the 100 and 75% conditions was not significantly different than the drop between 75 and 50%, $W = 300.000$, $p = .076$.

The results from Experiment 4a and 4b were compared with using a 2 (Closeness to solution: 90◦ vs 10◦) x 3 (Accuracy: 50, 75 or 100% correct) mixed ANOVA with accuracy as the within-subjects factor. This revealed a significant drop in trust with decreasing accuracy, F (1.708, 99.091) = 48.784, $\eta^2 = .286$. Bonferroni-corrected post-hoc tests revealed that ratings were significantly more positive in the 100% accuracy condition, compared to both the 75 and 50% conditions, ps < .001. Ratings in the 75% condition were significantly more positive than in the 50% condition, $p < .05$. There was a marginal main effect of

closeness to solution, such that ratings (across all levels of accuracy) were more positive for systems that were 10◦ away from the correct solution when incorrect, compared to 90◦, F (1, 58) = 4.019, p = .05, η^2 = .024. There was no significant accuracy × closeness to solution interaction, F (1.708, 99.091) = 1.336, p = .266, η^2 = .008.

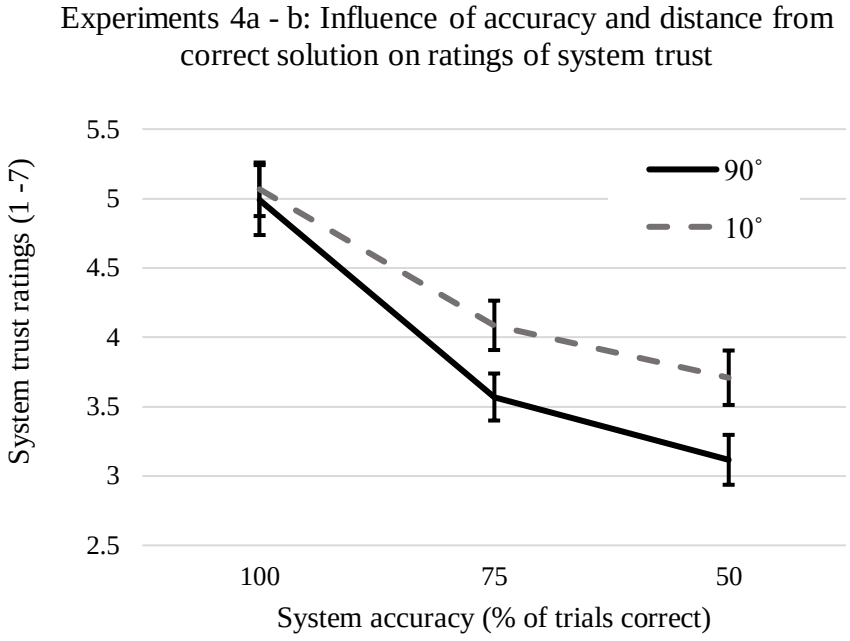


Figure 5.8. Trust questionnaire ratings for Experiments 4a – b: Influence of accuracy and closeness to solution (90˚ vs 10˚ distance from correct solution). Error bars represent ±1 standard error of the mean.

Combined analysis on effect of closeness to solution (Experiments 3a-b & 4a-b)

The dataset from Experiments 3a-b and 4a-b were combined into a larger dataset to examine the influence of accuracy and closeness to solution on system trust ratings. System trust ratings for the combined dataset are shown in *Figure 5.9* and were first analysed using a 3 (Accuracy: 100, 75 and 50%) × 2 (Closeness to correct solution: 90˚ or 10˚) mixed ANOVA. There was a significant main effect of accuracy on system trust ratings, F (1.802, 212.584) = 123.094, p < .001, η^2 = .303. Bonferroni-corrected post-hoc tests revealed that ratings in the 100% accuracy condition were significantly more positive compared to both the 75 and 50% conditions, ps <.001. Ratings in the 75% condition were significantly more positive than in the 50% condition, p < .001. There was a significant accuracy × closeness to solution interaction, F (1.802, 212.584) = 5.300, p < .01, η^2 = .013. Simple main effects revealed that there was a significant main effect of accuracy, both for systems that were 90◦ and 10◦ away from the target on incorrect trials, ps < .001. It was also revealed that systems

that were 10◦ away from the correct solution were rated more positively than those that were 90◦ away, when examining the 75 and 50% accuracy conditions, *p*s < .05. There was no effect of closeness to solution when examining the 100% accuracy condition, $p = .414$. Further follow-ups revealed that the drop in trust between 100 and 75% accuracy was significantly larger than the drop between 75 and 50%, $W = 4943.500$, $p < .001$. There was also a significantly larger drop between 100 and 75% accuracy when the system was 90˚ away from the correct solution, compared to 10˚, $U = 1302.000$, $p < .01$, with no difference in the drop between 75 and 50% as a function of closeness to solution, $U = 1982.000$, $p = .340$. The main effect of closeness to solution did not reach significance, $F (1, 118) = 2.775$, $p = .098$, $\eta^2 = .009$.

A separate mixed ANOVA was conducted to examine a possible interaction between Accuracy (100, 75 and 50%) and Experiment (3a, 3b, 4a and 4b). This analysis also revealed a significant drop in trust with decreasing accuracy, $F (1.802, 209.021) = 121.640$, $\eta^2 = .303$. As above, Bonferroni-corrected post-hoc tests revealed that ratings in the 100% accuracy condition were significantly more positive compared to both the 75 and 50% conditions, *p*s <.001, and ratings in the 75% condition were significantly more positive than in the 50% condition, $p < .001$. Neither the main effect of experiment, $F (3, 116) = 1.276$, $p = .286$, $\eta^2 = .013$, nor the accuracy × experiment interaction, $F (5.406, 209.021) = 1.948$, $p = .082$, $\eta^2 = .015$, were significant.

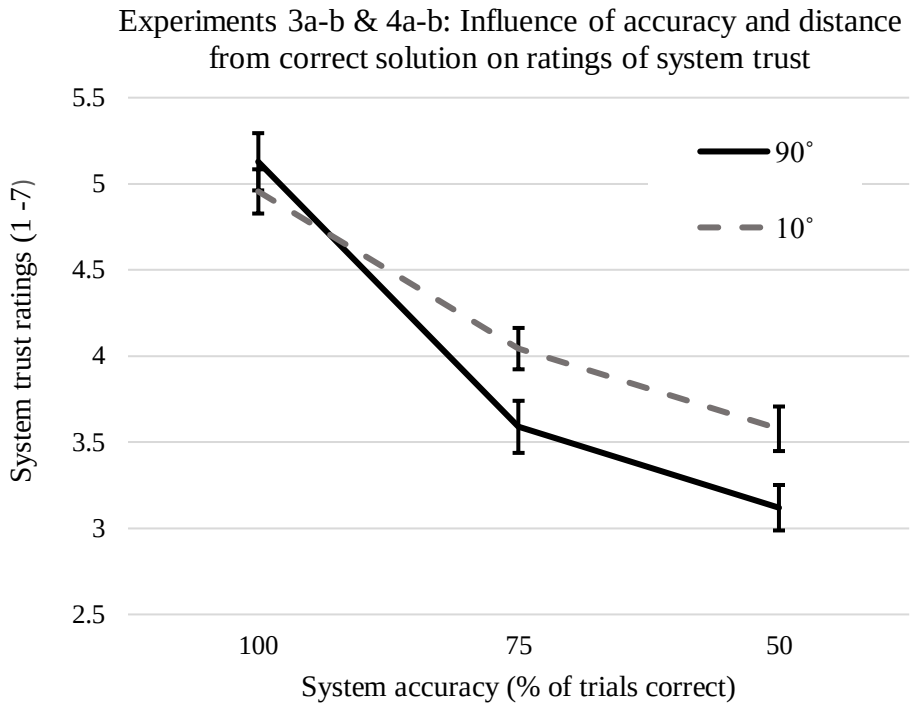


Figure 5.9. Trust questionnaire ratings for Experiments 3a – b and 4a – b: Influence of accuracy and distance from correct solution (combined dataset, N = 120). Error bars represent ±1 standard error of the mean.

5.5.3. Discussion

Experiments 3a-b and 4a-b investigated the influence of accuracy and distance from the correct solution on ratings of system trust, using a relatively context-neutral yet slightly more realistic task in comparison to the previous five experiments presented in this chapter. All four experiments presented participants with systems that were either 100, 75 or 50% accurate in their attempts to reach a target position on a dial. In Experiments 3a and 4a, the needle of the dial was always 90˚ away from the correct solution on trials where the system made the incorrect decision, whereas the distance from the correct solution in Experiments 3b and 4b was 10˚.

When comparing Experiments 3a and 3b, an interaction between accuracy and distance from the correct solution was revealed, such that the drop in trust between 100 and 75% accuracy was shallower when the system was closer to the correct solution (10˚ away). This was not replicated in the comparison between Experiments 4a and 4b, but a combined dataset of Experiments 3a-b and 4a-b revealed the same effect, providing some evidence that participants were more forgiving of less accurate systems when their attempts were closer to the correct solution. It is possible that the lack of an effect in Experiments 4a-b may have

been due to a power issue, which may have been remedied by combining the datasets. Although the drop in trust between 100 and 75% accuracy was shallower for systems that were closer to the correct solution, the pattern resulting from Experiments 3a and 4a was similar to most of the colour decision task experiments presented in this chapter, in that users experienced a more rapid drop in trust between 100 and 75% accuracy, compared to between 75 and 50%.

5.6. General discussion

The present investigation consisted of nine experiments designed to examine the influence of absolute accuracy, rank position, size of problem space and closeness to solution on evaluations of system trust. In all eight experiments which included the 100% accuracy condition, there was a sharp drop in system trust when the level of accuracy dropped below 100%, with a much smaller drop between 75 and 50% accuracy. Experiment 1b (where system accuracy was between 75 and 50%, with no 100% condition) found no effect of system accuracy on trust ratings. Experiments 1a-1c revealed that rank position (whether a system was ranked $1^{st}/2^{nd}/3^{rd}$ in a hierarchy) had no mediating influence on the effect of absolute accuracy. Experiments 2a-2b revealed that the size of the problem space (whether the system chose between two or four targets) also had no effect on the effect of accuracy on trust. Experiments 3a-b and 4a-b replicated the effect of accuracy in a different task domain, and provided some evidence that users were more forgiving of less accurate systems when the incorrect decisions of those systems were close to the correct solution.

The key finding across eight of nine experiments (all of those where a 100% accuracy condition was included) was a steep drop in system trust between the highest and second-highest levels of accuracy, with a much shallower drop in trust between the middle and lowest levels of accuracy. This rapid loss of trust below 100% accuracy is in line with some previous lab-based work in that system users often have a low tolerance for error and that trust in computerised systems can be lost very quickly with decreasing accuracy (Chavaillaz et al., 2016; Yu et al., 2017). In addition to examining the effect of accuracy, the first three experiments presented in this chapter provided participants with different ranges of accuracy to test if users were sensitive to rank position or absolute accuracy when making judgments of system trust (in the context of previous work demonstrating a strong sensitivity to relative rank in judgement and decision making: Ert & Erev, 2013; Mullett & Tunney, 2013; Stewart, 2009; Stewart et al., 2006; Walasek & Stewart, 2015, 2019; Vlaev et al., 2011). Across the three experiments, the pattern of results suggest that users were very sensitive to decreases in

accuracy between 100 and 75% but did not discern between levels of accuracy between 75 and 50%. This suggests that any effects of rank order did not override the effect of absolute accuracy: If participants were basing their trust on the relative levels of accuracy amongst the systems they were presented with, they would be expected to show higher trust in the 75% accurate system in Experiment 1b (which revealed no difference in trust between accuracies of 75 and 50%). However, the findings of Experiments 1a-c suggest that, if a system's absolute performance is poor, comparing it to the performance of worse systems is not enough to increase trust in that system. These findings are in contrast with what would be predicted by the decision by sampling approach (Stewart, 2009; Stewart et al., 2006; Walasek & Stewart, 2015, 2019), which argues that people do not calculate objective value independent of rank-based comparisons, and that judgments are made based on a comparison with all other options available. Future research into rank position might usefully provide participants with more experience with the system (e.g. a larger number of trials and/or blocks): This would give participants a larger number of instances against which to compare the performance of any given system, which could lead to any rank-based effects on trust emerging. However, the results here were clear, participants were highly influenced by the absolute accuracy of a system and not by its relative position amongst alternatives.

Experiments 2a and 2b focused on the effect of task difficulty, particularly the size of the problem space (number of possible solutions). The expectation was that if the task a system is attempting to perform is more difficult (a higher number of potential solutions), then users might be more forgiving of systems with lower levels of accuracy (compared to if the system were attempting a simpler task). However, there was no influence of the number of solutions (colours the system could select) or the number of targets, with the effect of accuracy and the steep drop in trust below 100% remaining very similar irrespective of the size of the problem space (further supporting the findings of Experiments 1a-c). This suggests that, in the case of context-neutral systems performing a discrete task such as generating a target colour, users may not be very sensitive to the number of alternatives the system has to choose from (and therefore the difficulty of the task).

The null results from Experiments 2a-b are at odds with some research suggesting that users are sensitive to the difficulty of the task a system is attempting to perform. For example, one study gave participants a target detection task (indicating the presence of an 'X' on screen) to complete, while being assisted by a decision aid system which was either 60% accurate on the most difficult trials, and 100% accurate on the easiest trials, or the reverse (Madhavan, Wiegmann & Lacson, 2006). Madhavan et al. (2006) found that, although the

overall accuracy of the system was the same across trial difficulty, participants were less trusting of systems that made errors on the easier trials. A similar study gave participants instructions that framed a target-detection task as either easy, moderate difficulty or difficult, finding that participants were more likely to comply with the system's decisions when the task was framed as more difficult, but participants' levels of compliance decreased with increasing difficulty when more difficult trials were associated with larger potential rewards (Schwark, Dolgov, Graves & Hor, 2010). More recently, Faerevaang, Nguyen, Jimenez and Jentsch (2017) presented participants with descriptions of robots providing humans with information at either 80% or 50% accuracy, in either a safe or a dangerous environment. Faerevaang et al. (2017) found that, while participants were sensitive to accuracy overall, they gave more positive evaluations of the robot providing information in the dangerous situation, independent of its accuracy level. The lack of an effect in the present study could arguably be due to the number of potential choices/targets being within a fairly small range (2 – 4), and participants may have therefore perceived the systems' task as fairly simple across all conditions. Future research may benefit from testing the influence of size of problem space/task difficulty in a more continuous domain (e.g. aiming to generate a target number) with a larger possible number of solutions. While the first five experiments presented in this chapter examined user trust in systems aiming to generate a target colour, Experiments 3a-b and 4a-b focused on trust of systems aiming to reach a target position on dial display. This provides more of a link between trust in more realistic in-vehicle systems (as explored in Chapter 3), although the task remained fairly context-neutral (as participants were not told what the dials were intended to measure). However, the key reason for using dial-based systems was to study the impact of distance from the correct solution (i.e. was the system's attempt 'almost there', or very off the mark?) on trust, whereas the systems aiming to generate a target colour were performing a more binary decision task. This provided an important extension to the literature on system trust, as previous studies on the effect of system accuracy have focused on systems that are either fully correct or fully incorrect on a given trial (Chancey et al., 2013; Chavaillaz & Sauer, 2017; Chavaillaz, et al., 2016; Madhavan & Phillips, 2010; Yu et al., 2017). Comparison between Experiments 3a and 3b found that, when the system was closer to the target position when its estimate was not correct, participants were more forgiving of less accurate systems and displayed a shallower drop in trust with decreasing accuracy, suggesting some sensitivity to distance from the correct solution. Although this effect was not replicated by Experiments 4a-b, a combined dataset of all three dial-based experiments showed a similar pattern to the comparison

between 3a and 3b (*Figure 5.9*). Although these findings should be taken with some caution due to the lack of an interaction in Experiments 4a-b, overall, they suggest that users of computerised systems may be more trusting of less accurate systems that are frequently 'almost correct'. This suggests some degree of context-dependency in how such systems are evaluated, although the effect of accuracy remained much stronger overall.

It is also important to note that, although the combined dataset from Experiments 3a-b and 4a-b showed the same pattern of results as Experiments 3a-b, there are issues with combining the original experiments with the replication in this way. Specifically, it is very likely that the significant findings of the combined analysis were driven by the findings of Experiments 3a-b alone, and it is therefore not clear whether combining this with Experiments 4a-b for a more powerful analysis had any effect. It would be more valuable in future research to conduct a single study with a larger sample size, attempting to replicate the original findings of Experiments 3a-b. This would be a more methodologically sound test of whether users of computerised systems really are more forgiving of less accurate systems which get closer to the correct solution.

The finding that participants were sensitive to closeness to solution suggests that, when evaluating the performance computerised systems, users take account of '*how* the system gets it wrong' when it makes the incorrect decision, as well as how often incorrect decisions occur. A parallel can be drawn to research on how users react to different *types* of error. For instance, some previous work has shown that, when users complete a decision task assisted by a computerised system, false alarms (indicating the presence of the target which is not present) have a greater negative impact on performance compared to misses (failing to detect a target that is present: (Chancey, Yamani, Brill & Bliss, 2017; Dixon, Wickens & McCarley, 2007; Wiczorek & Meyer, 2016)). Although this work does not focus on system trust directly, it is arguably in line with the results of Experiments 3a-b (and the combined analysis of 3a-b and 4a-b), in that they suggest users of computerised systems react differently to different types of errors, as well as the frequency of errors, whether the errors differ in terms of direction (Chancey, et al., 2017; Dixon, et al., 2007; Wiczorek & Meyer, 2016) or distance from the correct solution.

5.7. Summary and conclusions

The nine experiments presented in this chapter investigated the effects of accuracy, rank position, task difficulty and distance from correct solution on trust of context-neutral systems. This project was conducted in the context of background literature on system trust

largely focusing on trust of specific, existing technologies and/or systems (Introduction and Chapter 1), which potentially presents the problem of experience with existing systems confounding judgments of a completely novel systems' behaviour (e.g. see Lim et al., 2009). This approach is also in contrast to Chapter 4, which also examined the influence of information presentation on trust, but in the context of in-vehicle systems in simulated autonomous highway journeys.

The experiments presented in Chapter 5 found a sharp drop in system trust below 100% accuracy, with a shallower drop below this initial decrease. This 'two-stage' pattern of trust as a function of system performance can be compared with the results of the six experiments on autonomous overtaking (see Chapter 3), in which evaluations of overtaking manoeuvres became sharply more positive with increasing pull-in distance, before reaching a plateau. While neither rank position nor the difficulty of the task appeared to influence the pattern of trust ratings over accuracy, there was some evidence from the last three experiments in this chapter that users were sensitive to distance from the correct solution, being more forgiving of less accurate systems when they were closer to being correct. In conclusion, the experiments presented in this chapter revealed a limited degree of context sensitivity in users' trust of context-neutral systems, with accuracy exerting the largest influence, and other contextual factors having little or no influence on trust ratings in this particular methodology. These findings are somewhat analogous to those in Chapters 3 and 4, in which one central variable of system performance (pull-in distance in Chapter 3, lane position in Chapter 4) exerted the largest effect on occupants' experience of the manoeuvres/journeys, with traffic context (third vehicle in Chapter 3, lead vehicle type in Chapter 4) having a more limited degree of influence.

The next chapter (Chapter 6) presents a set of experiments that also investigated users' evaluations of computerised systems in a fairly abstract laboratory-based setting. However, while Chapter 5 uses a more context-neutral methodology to answer empirical questions on what factors influence trust in computerised systems, Chapter 6 focuses on applying more theoretical work on how people integrate multiple sources of information to make overall values judgments to trust of in-vehicle systems.

6. Framing-based value integration biases in the evaluation of in-vehicle systems

Abstract

When given the task of estimating the overall value of sequences of numbers, previous research has shown that people's judgments can be biased. For example, their preferences can be reversed by changing how the question is framed (e.g., 'select the high value sequence' or 'reject the low value sequence'). While such framing effects have been observed consistently when focusing on abstract streams of digits, previous work has not yet observed if these same biases apply to evaluation of computerised systems. Five experiments (total N = 300) were conducted, focusing on the influence of task framing, directional bias, level of information provided and differences in variability on participants' judgements of two in-vehicle systems that over time indicated the same mean values but had different standard deviations around the true value. The results were partially consistent with the framing effects found in previous research suggesting that users of in-vehicle systems may be somewhat susceptible to framing biases when comparing the performance of different systems. However, across all five experiments, the results suggest that several specific conditions may be required for system users to be sensitive to variability and framing effects: System performance must be evaluated in a single direction (i.e. underestimating or overestimating); the differences in variability between the two systems needs to be relatively large, and the system decisions must be presented instantly for any framing biases to occur.

6.1. Introduction

Users' trust of computerised systems, defined by Marsh & Dibben (2003) as *"a positive expectation regarding the behavior of somebody or something in a situation that entails risk to the trusting party"*: pp. 470, is important to understand because trust has been shown to influence usage intentions (Choi & Ji, 2015). Previous work into the evaluation of computerised systems has found user trust to be sensitive to characteristics of the user, properties and behaviour of the system and the context of the user-system interaction, although factors relating to properties of the system have been shown to be the most influential (Hancock et al., 2011; Schaefer, Chen, Szalma & Hancock, 2016).

Previous experimental work has shown that the appearance of computerised systems can influence trust, with more anthropomorphised (human-like) systems being rated as more trustworthy compared to those with a more robotic appearance (Lee, Kim, Lee & Shin, 2015; Pak, Fink, Price, Bass & Sturre, 2012; Waytz, Heafner & Epley, 2014). Similarly, the type of information that users are provided with has also been shown to be influential in system trust, with systems that display more information regarding the system's level of confidence being rated as more trustworthy (Beller, Heesen & Vollrath, 2013; McGuirl & Sarter, 2006; Verame, Constanza & Ramchurn, 2016). Providing more detailed information (Verberne, Ham & Midden, 2012) or information that provides clearer instructions on the task (Cramer, Evers, Kemper & Wielinga, 2008) have also been shown to influence trust in those systems. However, providing more information has also been shown to decrease trust, preventing over-trust in partially automated systems (Helldin, Falkman, Riveiro & Davidsson, 2013). Increased system transparency has also been shown to decrease trust if it exposes flaws in a poorly performing system (Kaltenbach & Dolgov, 2017). Trust has also been shown to be sensitive to accuracy, with more highly accurate systems being rated as more trustworthy by users (Chancey, Proaps & Bliss, 2013; Madhavan & Phillips, 2010). Trust also drops sharply with decreases in system accuracy over time (Chavaillaz & Sauer, 2017; Chavaillaz, Wastell & Sauer, 2016; Yu et al., 2017).

Building upon previous work into system trust, other work presented within this thesis has explored various factors that influence evaluations of computerised and contextually-neutral systems (Chapter 5), information presentation systems in in-vehicle contexts (Chapter 4), and the behaviour of autonomous vehicles (Chapters 3 & 4). In relation to performance of computerised systems, the experiments presented in Chapter 5 revealed a sharp drop in trust when a system falls below 100% accuracy, which flattened between 75 and 50%. Furthermore, this overall pattern was not sensitive to the rank distribution of systems experienced within a

single context nor the complexity of the task. Users were, however, more forgiving of less accurate systems if their incorrect attempts at a task were closer to the correct solution. In relation to information presentation, simulator-based experiments in Chapter 4 revealed that presenting occupants of autonomous vehicles with visual information on their vehicle's speed improved perceptions of the journey, but only when the participants' vehicle was travelling at the maximum limit, although one experiment also revealed that providing additional information did not always improve journey satisfaction.

In addition to more intuitive factors such as objective accuracy, (Chancey et al., 2013; Chavaillaz & Sauer, 2017; Chavaillaz, et al., 2016; Madhavan & Phillips, 2010; Yu et al., 2017), the way in which information about computerised systems is framed has been shown to influence trust. For instance, one study asked participants to indicate whether an 'X' was present on screen using a computerised aid system that had an accuracy of 80% (Lacson, Wiedemann & Madhavan, 2005), framed either positively ("80% correct"), negatively ("20% incorrect") or neutrally (both statements). Participants provided with the negative frame were more likely to rely on the system's decision when it did not detect an 'X', resulting in fewer false positives (indicating that the target was present when it was absent), but also more false negatives (missing a target when it was present). Another study provided participants with reports on suspected fraud generated by decision aids, which were framed as either '80% accurate' (positive) or '20% inaccurate' (negative), and asked participants to indicate their intention to follow up on the suspected fraud (Huerta, Glandon & Petrides, 2012). Huerta et al. (2012) found that participants demonstrated lower intentions to investigate fraud when presented with the negatively framed accuracy information. Rice and McCarley (2011) found that participants performed more poorly on a signal detection task and were less likely to use a decision aid when its incorrect decisions were 'false alarms' (detecting a target that is not present), compared to 'misses' (not detecting a target that is present), at the same level of objective accuracy. Cheng and Wu (2010) found that online shoppers were more inclined to purchase a translation device described as having "80% accuracy", compared to one described as having a "20% error rate", although this effect was weakened by providing participants with text-based warnings about the influence of framing effects on product choice.

A specific domain in which framing effects appear, but which has not yet been investigated in the domain of computerised systems, is *value integration*. This refers to the task of combining multiple individual values (e.g. individual values presented within a stream of digits) to make judgments of overall value (e.g. "what is the overall value of this stream?": Glickman, Tsetsos & Usher, 2018; Kunar, Watson, Tsetsos & Chater, 2017; Tsetsos, Chater &

Usher, 2012; Usher, Tsetsos, Glickman & Chater, 2019). Previous research into value integration has demonstrated paradoxical choice biases. For example, in one condition Tsetsos et al. (2012) presented participants with two streams of briefly presented values and asked them to choose the stream with the highest average value. Crucially both streams of digits had the same mean value but different standard deviations. When participants were asked to select the best stream (the one with the highest overall value in a slot machine task), they chose the stream with the higher standard deviation. However, when asked to reject the system with the worst performance (lowest overall value), participants also chose the stream with the higher standard deviation. Thus, the same stream (or system) was both chosen as the best and rejected as the worst depending on how the question was framed. Tsetsos et al. (2012) concluded that the framing of the question participants were asked (choose the best vs reject the worst stream) made different individual values in the stream more salient, biasing participants' judgment of the overall value of the stream, resulting in the paradoxical choices observed. That is, the positive frame biased people to see the high values in a stream and the negative frame biased people to see the low values. Since by definition the stream with the highest variability had some higher values and some lower values than the low variability stream, people chose that stream when presented with either the negative or the positively framed question. *Figure 6.1* provides an example of two distributions that could be used in a typical value psychophysics paradigm.

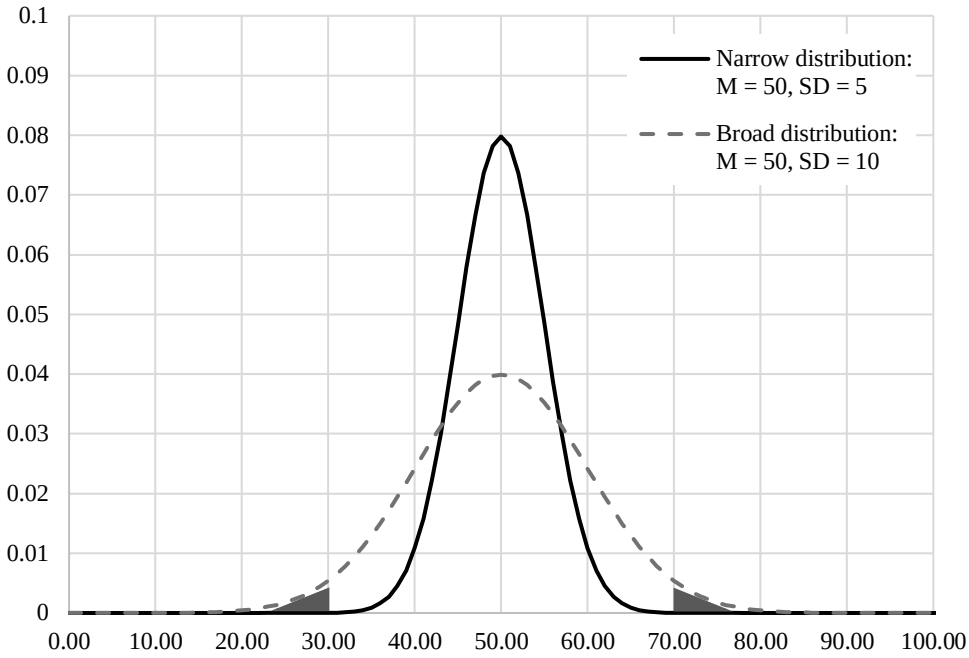


Figure 6.1. Schematic of two example distributions in a typical value psychophysics experiment. Both distributions share the same mean but differ in their standard deviations. The black distribution is narrower and frequently produces values close to the mean, whereas the grey dotted distributed is broader and produces both very high and very low values in comparison. The darker shaded triangular areas on the left and right edges of the broad distribution: When instructed to select the stream with the *highest* value in Tsetsos et al. (2012), participants' attention was biased towards the higher end of the distributions (right), resulting in the broader distribution being chosen more frequently. Paradoxically, the broader system was also selected when participants were asked to choose the *lowest* value stream, as attention was biased towards the lower end of the distribution (left), resulting in the framing bias effect.

Kunar et al. (2017) built on these findings by examining the effect of irrelevant but salient outliers (they were a different colour than the other items in the stream) on value judgments. They found that including a high-value outlier in a stream of numbers led to participants reporting the overall value as higher, with streams including low outliers judged as lower value overall. Participants were also less accurate at identifying high target values when a low-value outlier was presented in the stream and were less accurate at identifying low target values when the outlier was high. In another study focusing on attentional mechanisms, participants were provided with two streams of numbers and asked to select the stream with

the highest mean (Glickman, et al., 2018). On some trials, Glickman et al. (2018) placed a red dot above one of the numbers in the pair (either the higher number or the lower number), and participants were asked to report whether the dot was present. Glickman et al. (2018) found that participants instructed to select the highest number of each pair were more accurate at identifying the dot when it appeared on the higher number of each pair, whereas participants asked to select the lowest number were more accurate at identifying the dot when it appeared on the lower number. In a second experiment, Glickman et al. (2018) also found that participants who preferred to sample payoffs from a broader (compared to a narrow) distribution also showed a higher bias for identifying the dot when it appeared on the higher payoff out of each pair. These findings support the framing biases found by Tsetsos et al. (2012), and also suggest a key role of attention in driving preferences for 'riskier' distributions and framing biases.

Integrating findings on biases influencing value-based decision-making, Usher et al. (2019) presented a theory of 'selective integration' to explain these effects. Usher et al.'s (2019) theory posits that overall judgments of value (for instance, which stream of numbers has a higher overall value) are formed in a bottom-up manner by making a series of smaller comparisons one-by-one, selectively attending to features that are relevant to the goal of the task. This explains the framing-based preference reversals found in previous research (Glickman et al., 2018; Tsetsos et al., 2012), as changing the framing of the question would have changed the features that participants were attending to when making each individual comparison (e.g. higher vs lower values).

This chapter presents five experiments which aimed to examine the potential role of framing biases in value integration (Glickman et al., 2018; Tsetsos et al., 2012) in the domain of evaluation of in-vehicle computerised systems. Clearly, in terms of commercial evaluation, it is important to know if framing biases apply in these situations given that framing effects could lead to the very same instrument or system being favoured or rejected depending on how the question is asked. Similarly, the success of a system might depend on how people naturally choose to frame their thoughts when thinking about how good a system is compared to others. The experiments presented below also contribute to the general value integration literature by exploring framing biases in a spatial, rather than numerical, domain. Although one experiment in the value integration domain (Tsetsos et al., 2016) used a spatial task (participants were required to compare the heights of pairs of rectangles), it did not focus on the influence of framing biases. In addition to contributing to the value integration literature by examining framing biases in a more applied domain, the work presented in this chapter contributes to the

wider system trust literature by exploring additional psychological factors that might influence the evaluation of in-vehicle systems.

In Experiments 1 and 2, participants were presented with pairs of systems which were supposed to accurately indicate the amount of battery range left in an electric vehicle. In addition to the value indicated by each system, the true remaining range was also presented so that participants could see how accurate the systems were. After a series of trials participants were then asked to select the best or reject the worse system. Within a block of trials both systems indicated the same mean range, however, the estimates of range from one system had a higher standard deviation than those of the other. The influence of the overall mean value (how much battery range each system estimated on average) was also manipulated. Any given pair of systems either had a mean of 75% or 25% battery range, corresponding to the systems either consistently over- or under-estimating the true amount of battery remaining (which was always 50%). Half of all participants were always asked to select the best performing system, with the other half being asked to reject the worst performing system, providing a test of the effect of framing biases.

Experiments 3 – 5 also presented participants with high and low-variability pairs of systems and asked participants to either select the best or reject the worst performing system each block. However, Experiments 3 – 5 used a slightly different task in which the systems were aiming to maximally charge the battery of an electric vehicle, meaning that the system's success at the task need only be measured in one direction (more charge put into the battery corresponded to better performance), whereas the systems in Experiments 1 and 2 could perform poorly by either under- or over-estimating the true amount of battery life/range.

6.2. Experiment 1: Framing effects in evaluation of battery estimation systems

Experiment 1 was designed to build on previous findings (Glickman et al., 2018; Tsetsos et al., 2012) that participants were susceptible to framing biases when choosing between high and low-variability streams, and to examine whether these biases are present when evaluating the performance of computerised systems. In addition to testing for framing biases in a more applied setting, Experiment 1 also aimed to advance previous findings by: i) using a visual bar graph representation of quantities rather than Arabic numerals, and ii) examining the impact of mean value (whether the systems over or under-estimate the correct value) on how users evaluated system performance.

6.2.1. *Method*

Participants

Experiment 1 recruited sixty participants through a first-year Psychology undergraduate research panel, as a requirement for course credit (participant demographics for all five experiments are presented in Table 6.1). Sample sizes were based on previous value integration research. For instance, Tsetsos et al. (2012) recruited sixty-seven participants across three experiments, with Glickman et al. (2018) recruiting between fifteen and thirty-three participants per experiment, and Kunar et al. (2017) recruiting twenty in each experiment. In order to reduce the possibility of not finding any sensitivity to variability in participants' choices, each individual experiment recruited sixty participants, making the sample sizes substantially larger than those of comparable studies.

Table 6.1. Participant demographics for Chapter 6, Experiments 1 – 5.

Experiment	Number of participants (number female)	Age range, mean and standard deviation	Handedness (A = Ambidextrous)
1	60 (52 Female, 6 Male)	18 – 27 years M = 18.69 *SD* = 1.37	51 RH 8 LH
2	60 (49 Female, 9 Male)	17 – 22 years M = 18.43 *SD* = .82	52 RH 6 LH
3	60 (47 Female, 5 Male, 1 "Other")	18 – 35 years M = 19.38 *SD* = 2.85	47 RH 6 LH
4	60 (31 Female, 27 Male, 1 "Prefer not to say")	18 – 31 years M = 21.34 *SD* = 3.52	57 RH 2 LH
5	60 (40 Female, 15 Male, 1 "Other", 1 "Prefer not to say")	18 – 45 years M = 21.61 *SD* = 4.98	52 RH 4 LH 1 PNTS

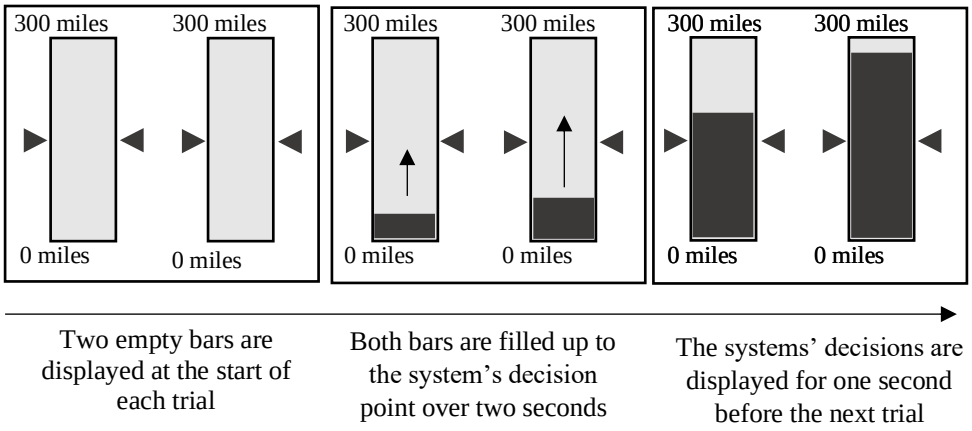
Stimuli and apparatus

The experiment was presented on 57 × 35cm LCD screens at 1920 × 1080px resolution in a full screen window. On each trial two systems were presented, one to the left and one to the right. The output of each system was indicated by a rectangle with a white outline filled with black space (such that they appeared empty on a black background). Each bar was 50 pixels wide and 600 pixels tall with 200 pixels in between the two rectangles. Over the course of each trial, both systems 'filled up' with a red rectangle which increased in size up to the amount of the estimated fuel range for that trial. The animated 'fill-up' time took 2s irrespective of the amount of range indicated. Text reading "300 miles" was present

at the top of each system to indicate the maximum possible amount of range remaining, with text presented on the bottom reading “0 miles”. Two grey arrows (on the left and right sides of each rectangle, pointing inwards) indicated the correct amount of battery range remaining (always 150 miles/50% full). After each block of trials, participants were asked to select either the best or reject the worst performing system (Left vs Right), by clicking one of two circular buttons presented in the middle of the screen (see *Figure 6.2A*).

Participants were presented with a paragraph of instructions on screen at the start of the experiment, explaining that they would be asked to choose the best or worst system after each block (depending on framing condition). Instructions in the positive framing condition were: “In each block you will see a sequence of trials where two systems will try to estimate the amount of battery range remaining. At the end of each block you will be asked to choose the BEST system out of the two you have just seen. Please press SPACE to start the first block when you are ready”. In the negative framing condition, the instructions were: “In each block you will see a sequence of trials where two systems will try to estimate the amount of battery range remaining. At the end of each block you will be asked to choose the WORST system out of the two you have just seen. Please press SPACE to start the first block when you are ready”.

A) Experiment 1



B) Experiment 2

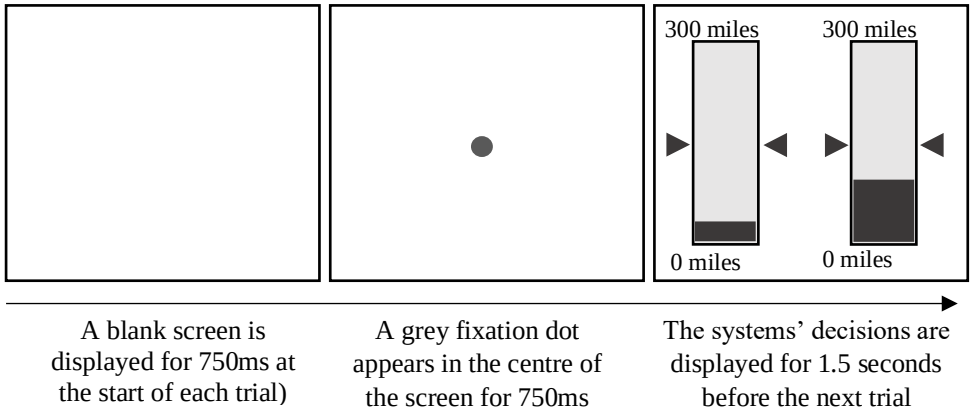


Figure 6.2. Schematic of a single trial in Experiments 1 and 2 of Chapter 6. Panel (A) represents Experiment 1, in which participants were given feedback on the systems reaching their decision. Panel (B) represents Experiment 2, in which system decisions were presented statically.

Design and procedure

Experiment 1 used a mixed 2 (Direction of Error: underestimate or overestimate actual range) × 2 (Framing: positive or negative) design with Framing as the between-subjects factor. All participants viewed twenty blocks of trials, in which two systems (presented side-by-side) aimed to estimate the amount of battery range remaining. The true range was always 150 miles (bar 50% full), and this was constantly indicated via two arrows (*Figure 6.2A*). In half of the blocks the systems underestimated the true range, indicating an average range of 75 miles (bar 25% full). In the other half they overestimated the true range, indicating an average range of 225 miles (bar 75% full). Both systems indicated the same mean, but one had a high variability (*SD* = 10), and one had a low variability (*SD* = 5). The values indicated by the two systems over each block were sampled from distributions with the required mean and standard deviation, with a minimum value of 2 and a maximum value of 48 (percentage of battery remaining) for the underestimating blocks, and minimum of 52 and a maximum of 98 for the overestimating blocks. Framing was randomly assigned between-subjects, such that half of all participants were asked to choose the *best* performing system after each block, and half were asked to choose the *worst* performing system. Within blocks, the placement of the two systems (whether the high-variability system was on the left or right side of the screen) was randomly paired with whether the block contained underestimating or overestimating systems.

Each block consisted of twelve trials in which both systems started as empty, and then began to fill up with a red bar to indicate the systems' estimate of the remaining battery range. The system took two seconds to indicate a value on each trial, the final value remained on screen for one extra second after which the bars disappeared, and the next trial began with both bars reset at zero. This resulted in a single trial length of three seconds. Participants were prompted to press the space key after each block, and then were required to click a button to choose one of the systems (left or right) to indicate which system to accept as better or reject as worse (depending on framing condition). The outcome variable was the number of times each participant chose the high variability system, divided by the total number of choices (giving a ratio between 0 and 1).

After providing informed consent, participants began the experiment by pressing the space key, and were presented with a paragraph of appropriately framed instructions - whether they would be asked to choose the best (positive-frame) or worst (negative-frame) performing system after each block. The framing instructions were presented once at the beginning of the experiment. Participants then completed the twenty blocks in a randomised

order (*Figure 6.3*), answering the question at the end of each block (either being asked to choose the best or the worst system depending on the framing condition). Participants were prompted to press the space key one more time after completing the question at the end of the final block, at which point they were given course credit, debriefed, and thanked for taking part. The experiment took approximately twenty minutes to complete.

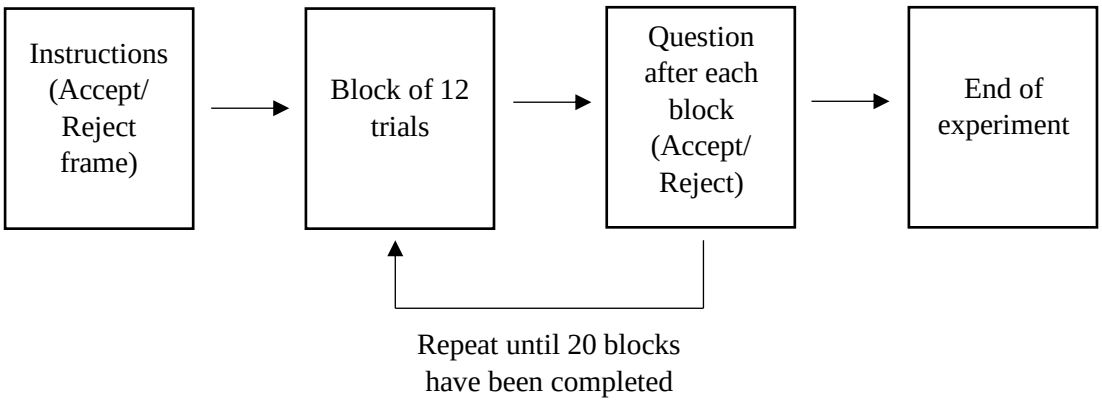


Figure 6.3. Overall structure of an experiment in Chapter 6, from presentation of the initial framing instructions (which applied to all twenty blocks of trials), to completion of the final block and accept/reject question.

6.2.2. Results

The data for Experiment 1 are presented in *Figure 6.4*, and were first analysed using a 2 × 2 mixed ANOVA on the entire dataset, with Direction of Error (system Under- vs Over-estimating) as the within-subjects factor and Framing (Accept vs Reject) as the between-subjects factor, and the ratio of high variability choices (0-1) as the dependent variable. Neither the main effect of whether the system under- or over-estimated, the main effect of framing or their interaction were significant, F (1, 58) = 1.689, p = .199, η^2 = .014, F (1, 58) = 2.280, p = .136, η^2 = .019 and F (1, 58) = .457, p = .502, η^2 = .004, respectively. A one-sample, one-tailed Wilcoxon signed-rank t-test was conducted on the total proportion of times participants chose the high-variability system (across all twenty blocks), against the chance value of 10 (corresponding to a probability of .5). Participants did not choose the high-variability system at a level significantly above chance, W = 543.500, p = .589 (p = .831, two-tailed), effect size (Rank-Biserial correlation) = -.406, with a Bayes analysis indicating moderate evidence (8.55 times more likely) in favour of the null, BF_{+0} = 0.117 (Jarosz & Wiley, 2014). A two-tailed analysis also indicated moderate evidence in favour of the null (6.86 times more likely), BF_{+0} = 0.146. Bayesian analyses were conducted here (and

in subsequent experiments), due to the surprising finding of null results, given the substantially larger samples used in the present study, compared to similar previous research (Glickman et al., 2018; Kunar et al., 2017; Tsetsos et al., 2012). One-tailed tests were used because the results of previous studies (e.g. Tsetsos et al., 2012) supported a predicted direction of preference: Participants were expected to choose the more highly variable system. The results of the one-tailed tests were mirrored by their two-tailed equivalents in terms of significance, and are included here (and for subsequent experiments) for completeness.

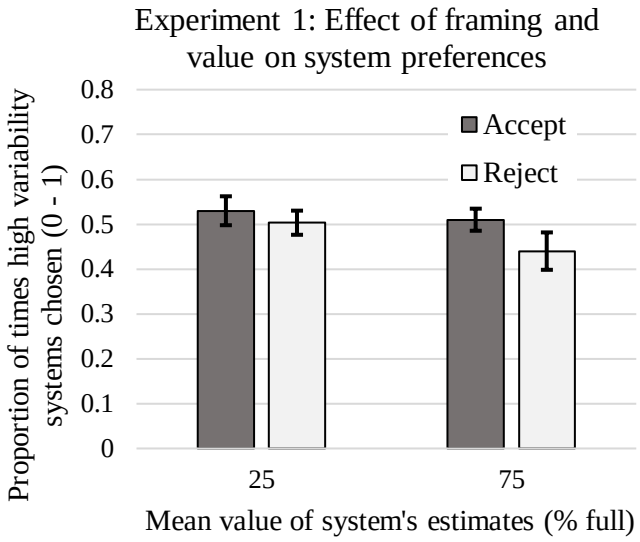


Figure 6.4. Proportion of times the high variability system was chosen in Experiment 1, as a function of mean value (25 vs 75%) and framing ("Accept" vs "Reject"). Error bars represent ±1 standard error of the mean. "Accept" refers to the positive framing condition ("please choose the best performing system), whereas "Reject" refers to the negative framing condition ("please choose the worst performing system").

6.2.3. Discussion

Experiment 1 aimed to build on the results of previous findings (Glickman et al., 2018; Tsetsos et al., 2012) by examining whether evaluations of battery range estimation systems are influenced by system variability and framing. Experiment 1 also extended previous work by examining the effect of direction of error (whether the system under- or over-estimated the amount of battery range remaining), by using a task in which performance can be measured in two directions (unlike Tsetsos et al., 2012, where higher numbers always

corresponded to better performance), and by giving users feedback on the system decisions through the addition of motion.

As in related earlier experiments that have used streams of digits (Glickman et al., 2018; Kunar et al., 2017; Tsetsos et al., 2012), it was predicted that participants would choose the highly variable system irrespective of whether asked whether to select the best performing system or reject the worse performing system. This is because of an increased bias for close solutions when asked to choose the best system and an increased bias towards far solutions when considering which system was the worst. However, there was no evidence of an above-chance preference for the high-variability system in either case, suggesting a lack of sensitivity to system variability in our task. It is possible that some specific aspect of this study prevented the bias and paradoxical decision making from emerging. Accordingly, Experiments 2 – 5 consisted of several variations on this task in order to test different possible explanations for this apparent lack of effect.

6.3. Experiment 2: Framing effects in evaluation of battery estimation systems

In Experiment 1 the bar-gauges of both systems initially started at zero (no fill) and gradually 'filled-up' to indicate their value over a period of 2s. This design choice was implemented to better mimic real-world instrumentation. However, it is possible that presenting the initial zero values at the start of each trial somehow interfered with processing of the final values. For example, perhaps the abrupt presentation of initial zero values at the start of each trial captured attention and influenced decision processes early on with subsequent 'gradually built-up values' less able to capture attention and so influencing decision processes to a lesser degree. Any such initial capture of attention might also reduce processing of subsequent values at those locations via processes related to 'inhibition of return' (e.g., Klein, 2000; Posner, 1980; Posner & Cohen, 1984; Posner, Rafal, Choate & Vaughan, 1985), whereby attentional processing is reduced for locations that have been previously attended.

Of note, previous studies have presented their numeric stimuli with an abrupt onset (which has been shown to capture attention: Jonides & Yantis, 1988; Yantis & Jonides, 1984; Yantis & Jonides, 1990), rather than building to up to a final value (e.g., Glickman et al., 2018; Kunar et al., 2017; Tsetsos et al., 2017). Thus, it is possible that this additional capture of attention due to abrupt onset may have contributed to the effects found in these studies. In order to test these possibilities, in Experiment 2 the final values of the two gauges were

presented abruptly in a single video frame – rather than the gauges starting at zero and gradually 'filling-up' (see *Figure 6.2B*).

6.3.1. *Method*

Participants

Experiment 2 recruited sixty new participants through a first-year Psychology undergraduate research panel, as a requirement for course credit (see Table 6.1).

Stimuli and apparatus

The stimuli and apparatus used in Experiment 2 were very similar to those used in Experiment 1 (see *Figure 6.2B*). However, rather than the red bars filling up gradually, the systems' decisions were presented instantaneously after a grey fixation dot appeared in the middle of the screen.

Design and procedure

The overall design was very similar to that of Experiment 1. However, the one key difference was that instead of viewing systems that started empty and filled up with a red bar to indicate their decision, participants in Experiment 2 were presented with the systems' decisions instantaneously. At the start of each trial, a blank screen was shown for 750ms before a grey fixation dot appeared. After another 750ms the dot disappeared and then the two systems were presented with the red bars already present to indicate each system's decision. The systems were presented on the screen for 1.5 seconds on each trial, in order to keep the overall trial and block length the same as in Experiment 1. The experiment took approximately twenty minutes to complete.

6.3.2. Results

The data for Experiment 2 are presented in *Figure 6.5*, and were first analysed using a 2 × 2 mixed ANOVA on the entire dataset, with Direction of Error (High vs Low) as the within-subjects factor and framing (Accept vs Reject) as the between-subjects factor, and the proportion of high-variability choices (0-1) as the dependent variable. As in Experiment 1, neither the main effects of Direction of Error, Framing, or their interaction were significant, F (1, 58) = 1.214, p = .275, η^2 = .011, F (1, 58) = 2.082, p = .154, η^2 = .016, F (1, 58) = 1.720, p = .195, η^2 = .016, respectively. A one-sample, one-tailed Wilcoxon signed-rank t-test showed that participants did not choose the high-variability system at a level significantly above (0.5) chance, W = 564.500, p = .397 (p = .794, two-tailed), effect size (Rank-Biserial Correlation) = -.383, with a Bayesian analysis indicating moderate support in favour of the null (5.58 times more likely, 6.81 two-tailed), BF_{+0} = 0.179 (0.147 two-tailed).

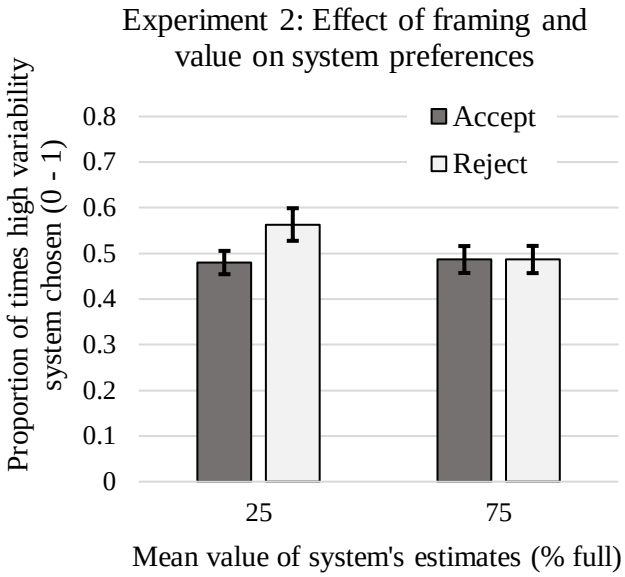


Figure 6.5. Proportion of times the high variability system was chosen in Experiment 2, as a function of mean value (25 vs 75%) and framing ("Accept" vs "Reject"). Error bars represent ±1 standard error of the mean. "Accept" refers to the positive framing condition ("please choose the best performing system), whereas "Reject" refers to the negative framing condition ("please choose the worst performing system").

6.3.3. Discussion

Experiment 2 aimed to test whether the lack of sensitivity to system variability in Experiment 1 could be explained by the presence of a zero start point. That is, in Experiment 1 both gauges started at zero and filled up to the desired value. In Experiment 2, the values were presented instantaneously which was closer to the presentation of numerical stimuli used in previous studies. However, even with this modification, participants were again not sensitive to differences in variability between the two systems. Experiments 3 to 5 test additional explanations in order to find the boundary conditions of paradoxical decision making in an instrumentation-based context.

6.4. Experiment 3: Framing effects in evaluation of battery charging systems

Experiment 3 made two changes to the paradigm. First, the difference between the low and high variability conditions was increased. This change was made due to null results from Experiments 1 and 2 (participants showed no preference in variability), and to bring the design more in line with previous value integration studies which used larger *SD*s (Kunar et

al., 2017: *SD* = ~16.5 or 20; Tsetsos et al., 2012; *SD* = 10 vs *SD* = 20), although Glickman et al. (2018) found effects of variability using smaller differences in variability (*SD* = 3.16 vs *SD* = 4.47). In Experiments 1 and 2, the low variability condition had a standard deviation of 5 and the high variability condition had a standard deviation of 10. However, when information is displayed in graphical format (as in the present work), larger differences in system variability might be needed than when information is presented in numerical format (as in Glickman et al., 2018; Kunar et al., 2017; Tsetsos et al., 2012; Usher et al., 2019). If this is the case, this would imply that bar-shaped gauges are more difficult to read and interpret or require greater processing capacity than simple numeric information which might lead to a reduced bias. Therefore, Experiment 3 used a standard deviation of 24 for the high variability systems (the highest possible standard deviation to generate a normal distribution with a mean of 50, with a minimum of 4 and a maximum of 96), and 5 for the low variability systems.

The second change related to the direction of error. Experiments 1 and 2 presented systems that estimated the amount of battery remaining in an electric car. The values indicated by the systems were manipulated to have a mean value of either 25% or 75% whilst the true range was 50%. As such they would either under- or over-estimate the amount of battery range remaining. Hence, both high and low indicated values could be inaccurate relative to the true value of 50%. This contrasts with previous work (Glickman et al., 2018; Kunar et al., 2017; Tsetsos et al., 2012) which used streams of numbers, in which higher values always corresponded to positive outcomes (i.e. task performance was measured in a single direction). Of note, each participant viewed both ten blocks of the systems overestimating and ten blocks of the systems underestimating the true amount of battery range remaining. It is possible that having to judge error by comparing indicated high and low values of a system with an intermediate true value removes judgement bias leading to no effect of system variability. It may be the case that variability-based framing bias only occurs when values are judged in one direction, for example, when larger always means better. Relatedly, participants may only be sensitive to variability if they only see the system either over- or under-estimating the target value, rather than viewing systems that miss the target value in both directions, as in Experiments 1 and 2. Such effects have been found in other domains. For example, in visual search, detecting a target of medium length amongst distractors of longer and shorter length is much more difficult than finding either the longest or the shortest target (Hodsoll & Humphreys, 2001). Similar effects are found with colour search in which searching for targets that are non-linearly separable from their distractors is

particularly difficult (Bauer, Jolicoeur & Cowan, 1996; Daoutis, Pilling & Davies, 2006). To test this possibility, in Experiment 3 the task was changed so that higher values were always associated with better performance. Specifically, the systems in Experiment 3 aimed to fully charge the battery of an electric car (to 100%). The success of the charging system was thus judged by how close to 100% charge the system achieved – thus higher values always indicated better performance.

6.4.1. *Method*

Participants

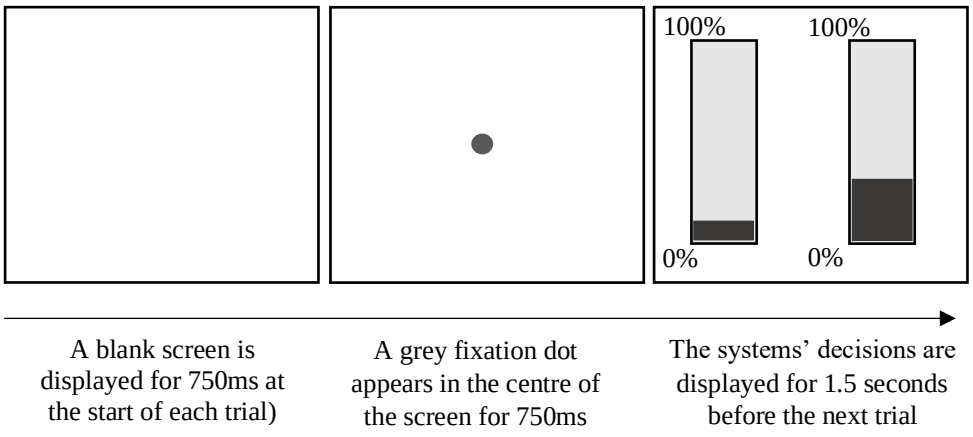
Experiment 3 recruited sixty new participants, through a mixture of a paid online recruitment panel and a first-year Psychology undergraduate research panel for course credit (see Table 6.1). Paid participants were paid £3 for taking part.

Stimuli and apparatus

Stimuli and apparatus were similar to those of Experiment 2, including instantaneous presentation of the systems' decisions following a fixation dot. However, there were several important differences. First, there were no arrows presented next to the systems (as in Experiments 1 & 2) to indicate the target for each trial because the target was always to achieve 100% battery charge. The instructions at the start of the experiment were modified to reflect this. In the positive framing condition, the instructions were: "In each block you will see a sequence of trials where two systems will try to fully charge the battery. At the end of each block you will be asked which system was the BEST out of the two you have just seen. Please press SPACE to start the first block when you are ready". In the negative framing condition, the instructions were: "In each block you will see a sequence of trials where two systems will try to fully charge the battery. At the end of each block you will be asked which system was the WORST out of the two you have just seen. Please press SPACE to start the first block when you are ready".

In addition, the text at the bottom and top of each system was '0%' and '100%' respectively (see *Figure 6.6A*).

A) Experiments 3 & 5



B) Experiment 4

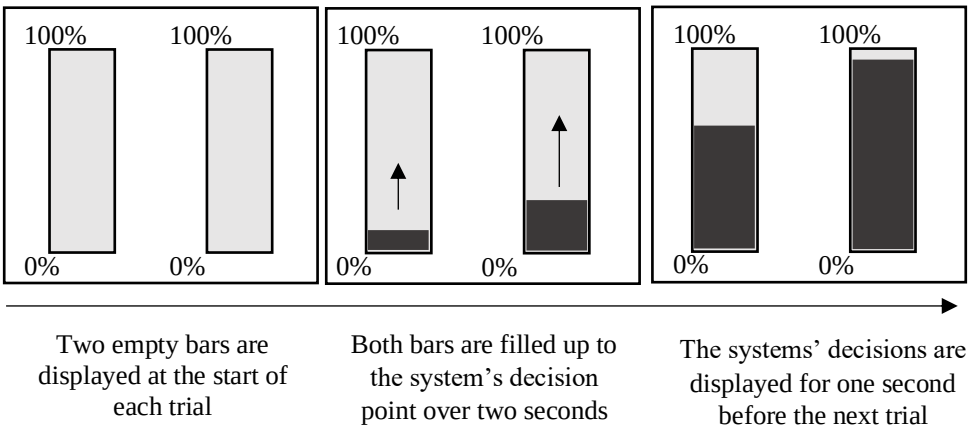


Figure 6.6. Schematic of a single trial in Experiments 3 – 5. Panel (A) represents Experiments 3 and 5, in which system decisions were presented statistically. Panel (B) represents Experiment 4, in which participants were given feedback on the systems reaching their decision.

Design and procedure

Experiment 3 presented participants with pairs of systems with high and low variability (as in Experiments 1 & 2). As there was no manipulation of direction of error in Experiment 3, the only independent variable was Framing (Accept vs Reject), randomly assigned between-subjects. All participants viewed twenty blocks of trials, in which two systems aimed to fully charge the battery of an electric car (to 100%). Participants viewed two systems side-by-side in each block. Both systems had the same mean (50% charged), but one had a higher variability (*SD* = 24), and one had a lower variability (*SD* = 5). Framing was randomly assigned between-subjects, such that half of all participants were asked to choose the *best* performing system after each block, and half were asked to choose the *worst* performing system. On each block the location of the low and high variability system (left and right) was assigned randomly but remained constant throughout a block of trials. Similar to Experiment 2, the decisions of the systems were instantaneously presented in each trial, with the same timings as in Experiment 2. The overall procedure was the same as the first two experiments, with participants giving informed consent, completing all twenty blocks in a randomised order, and then being debriefed, thanked, and being paid/granted course credit. The experiment took approximately twenty minutes to complete.

6.4.2. Results

The results for Experiment 3 are shown in *Figure 6.7.* A between-subjects *t*-test (two-tailed) on the effect of framing was conducted on the whole dataset as a first step in the analysis, with the proportion of times participants chose the highly variable system as the outcome variable (between 0 and 1, as in Experiments 1 & 2). As before, there was no significant effect of framing on the proportion of times participants chose the high-variability system, t (58) = 1.242, p = .219, d = .321. A one-sample, one-tailed *t*-test was conducted on the total proportion of times participants chose the high-variability system (across all twenty blocks), against a chance value of 10 (corresponding to a probability of .5). Participants chose the high variability system at a level significantly above chance, t (59) = 2.254, $p < .05$ (p = .014, .028 two-tailed), d = 3.107. An associated one-tailed Bayesian analysis provided moderate evidence in favour of the alternative over the null hypothesis, BF_{+0} = 2.88 (1.46 two-tailed).

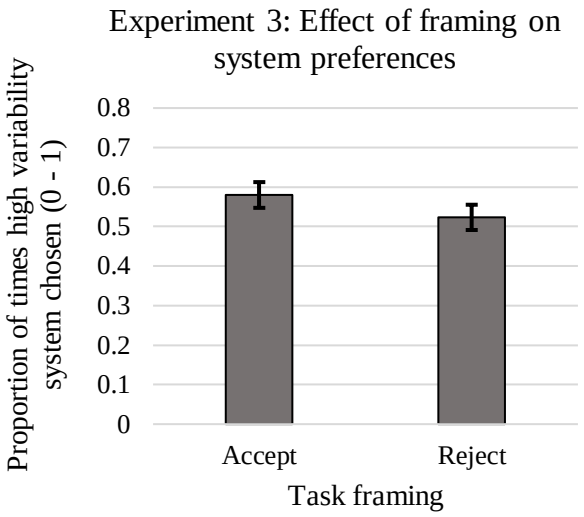


Figure 6.7. Proportion of times the high variability system was chosen in Experiment 3 as a function of framing ("Accept" vs "Reject"). Error bars represent ±1 standard error of the mean. "Accept" refers to the positive framing condition ("please choose the best performing system), whereas "Reject" refers to the negative framing condition ("please choose the worst performing system").

6.4.3. Discussion

Experiment 3 increased the difference in variability between the systems and used a task in which a single direction of magnitude was associated with a positive quality - a battery charging task where higher values always corresponded to better performance. This brought the system evaluation task in this study more in line with previous work in which framing biases have been found (Glickman et al., 2018; Kunar et al., 2017; Tsetsos et al., 2017).

In contrast to Experiments 1 and 2, the results showed that people were influenced by the variability of the systems. Specifically, overall, participants were more likely to choose the highly variable charging system when asked to select the best system and also when asked to select the worst performing system (in line with Tsetsos et al., 2012). This finding suggests that participants showed similar paradoxical choice preferences to Tsetsos et al. (2012), with their attention being drawn to higher values when asked to select the best system, and to lower values when asked to select the worst (both leading to selection of the high-variability system).

As there were two key differences in the design between Experiment 3 and Experiments 1 and 2 (larger differences in variability and a task with a single direction of

performance), it is not clear from these results to what extent each of these two factors influenced the results. Therefore, Experiments 4 and 5 were conducted in order to test these two factors independently, in order to examine what conditions are necessary for the paradoxical choice effects found in Experiment 3.

6.5. Experiment 4: Framing effects in evaluation of battery charging systems

Experiment 4 was very similar to Experiment 3 with one modification – participants viewed battery charging systems that started empty and filled up to indicate each attempt at charging the battery (whereas each attempt was presented instantaneously in Experiment 3). This was done to test the possibility that abrupt, instantaneous (as used previously by Glickman et al., 2018; Kunar et al., 2017; Tsetsos et al., 2012) presentation of system decisions may be required for users to be sensitive to system variability. Importantly, Experiment 4 used the same variability differences as Experiment 3 in which a variability-based bias was found.

6.5.1. *Method*

Participants

Experiment 4 recruited sixty new participants through a paid online recruitment panel (see Table 6.1). All participants were paid £3 for taking part.

Stimuli and apparatus

The stimuli and apparatus used in Experiment 4 were similar to those used in Experiment 3, with the exception that the systems' decisions were presented gradually as in Experiment 1. That is, participants viewed the red bar for each system increasing in size over a 2s period until it reached the decision for that trial (see *Figure 6.6B*).

Design and procedure

The design of Experiment 4 was similar to that of Experiment 3. The one key difference was that, rather than being presented instantaneously with decisions as in Experiment 3, participants in Experiment 4 were presented with two empty systems which filled up with a red bar to indicate their decisions (as in Experiment 1). The overall procedure was the same as in the three previous experiments, with participants giving informed consent, completing all twenty blocks in a randomised order, and then being paid, debriefed, and thanked. The experiment took approximately twenty minutes to complete.

6.5.2. Results

The results for Experiment 4 are displayed in *Figure 6.8.* A between-subjects Mann-Whitney *U*-test (two-tailed) on the effect of framing was conducted on the whole dataset as a first step in the analysis, with the proportion of times participants chose the high-variability system as the outcome variable (between 0 and 1, as in Experiments 1 – 3). There was no significant effect of framing on the proportion of times participants chose the high variability system, $U = 504.000$, $p = .427$, effect size (Rank Biserial Correlation) = .120. A one-sample, one-tailed Wilcoxon signed ranked *t*-test was conducted on the number of times participants chose the high variability system (across all twenty blocks), against a chance value of 10 (corresponding to a proportion of .5). Participants did not choose the high variability system at a level significantly above chance, $W = 832.000$, $p = .302$ (.605 two-tailed), effect size (Rank-Biserial correlation) = -.060, with a Bayes analysis indicating moderate evidence (6.34 times more likely, 7.02 two-tailed) in favour of the null, $BF_{+0} = 0.158$ (0.143 two-tailed).

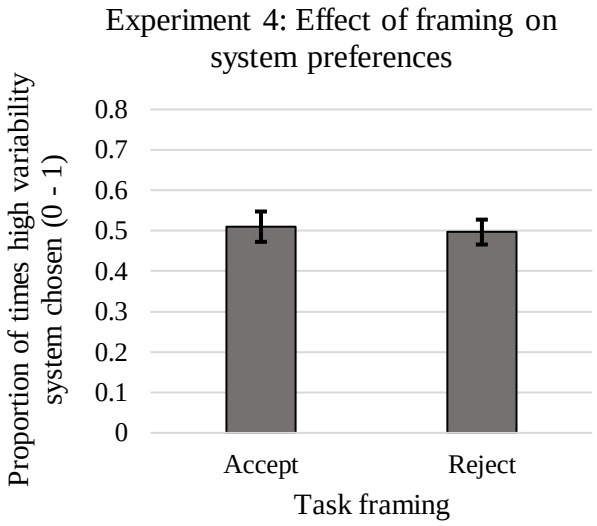


Figure 6.8. Proportion of times the high variability system was chosen in Experiment 4 as a function of framing ("Accept" vs "Reject"). Error bars represent the standard error. Error bars represent ±1 standard error of the mean. "Accept" refers to the positive framing condition ("please choose the best performing system), whereas "Reject" refers to the negative framing condition ("please choose the worst performing system").

6.5.3. Discussion

Experiment 4 was designed as a follow-up to Experiment 3, to examine the effect of using moving systems (where participants were given information about the time course of the systems' decisions) in contrast to static presentation of system decisions. Participants in

Experiment 4 did not show any above-chance preference for high-variability systems, in contrast to Experiment 3 where the high variability systems were selected in both framing conditions. This suggests that static and instantaneous presentation of system decisions may be necessary for participants to be sensitive to differences in variability and framing effects. However, recall that Experiment 2 also presented the abruptly presented, 'final decision', displays and yet also did not find a variability-based bias. However, Experiment 2 had systems that: i) both over- and underestimated the target value, and ii) used a smaller difference in variability between systems than Experiment 4. It is possible that a bias would occur with smaller variability differences so long as only a single direction of error is possible (e.g., the systems can only underestimate rather than being able to both under- and over-estimate) and display information is present instantly rather than gradually. To test this possibility, Experiment 5 presented instant display information (as in Experiment 3) using the 'single error' direction battery charging task (as in Experiments 3 and 4) but used the relatively smaller difference in variability between systems as used in Experiments 1 and 2.

6.6. Experiment 5: Framing effects in evaluation of battery charging systems

Experiment 5 used a combination of conditions from the earlier experiments to test if a variability-based bias would occur with: i) instantaneous display of information, ii) a uni-directional task (battery charging), and iii) a relatively small difference in variability between systems (as used in Experiments 1 and 2).

6.6.1. *Method*

Participants

Experiment 5 recruited sixty new participants through a paid online recruitment panel (see Table 6.1). All participants were paid £3 for taking part.

Stimuli and apparatus

The stimuli and apparatus were identical to those used in Experiment 3, but the differences in variability between the two systems in each block was smaller. See *Figure 6.6* for a schematic of a single trial in Experiments 3 – 5.

Design and procedure

The design of Experiment 5 was very similar to that of Experiment 3, in that participants were presented with the two systems' decisions on charging the battery instantaneously after viewing a fixation dot. However, the key difference with Experiment 3 was that the difference in variability were smaller (low SD = 5 and high SD = 10 as in Experiments 1 and 2). The overall procedure was the same as in the three previous

experiments, with participants giving informed consent, completing all twenty blocks in a randomised order, and then being paid, debriefed and thanked. The experiment took approximately twenty minutes to complete.

6.6.2. Results

The results for Experiment 5 are displayed in *Figure 6.9.* A between-subjects *t*-test (two-tailed) on the effect of framing was conducted on the whole dataset as a first step in the analysis, with the proportion of times participants chose the high-variability system as the outcome variable (between 0 and 1, as in Experiments 1 – 4). There was no significant effect of framing on the proportion of times participants chose the high-variability system, t (58) = .946, p = .348, d = .244. A one-sample, one-tailed *t*-test was conducted on the total proportion of times participants chose the high-variability system (across all twenty blocks), against a chance value of .5. Participants did not choose the high-variability system at a level significantly above chance, t (59) = .394, p = .347 (.695 two-tailed), d = .051, with a Bayes analysis indicating moderate evidence (5.06 times more likely, 6.57 two-tailed) in favour of the null, BF_{+0} = 0.198 (0.152 two-tailed).

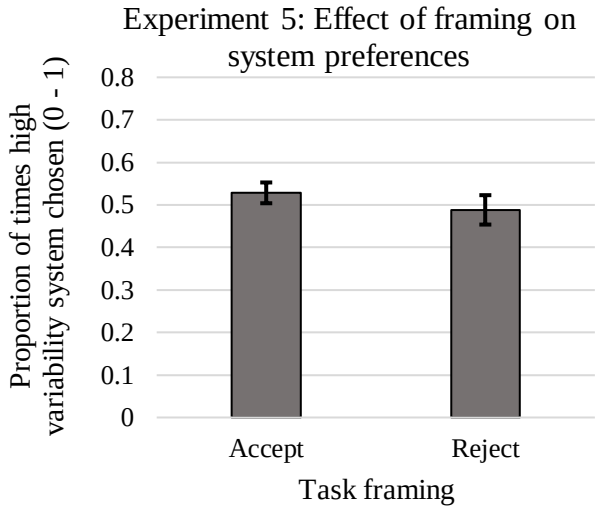


Figure 6.9. Proportion of times the high variability system was chosen in Experiment 5 as a function of framing ("Accept" vs "Reject"). Error bars represent ±1 standard error of the mean. "Accept" refers to the positive framing condition ("please choose the best performing system), whereas "Reject" refers to the negative framing condition ("please choose the worst performing system").

6.6.3. Discussion

Experiment 5 examined the effect of using smaller differences in variability between systems on participants' choices between instantaneously presented, unidirectional battery charging systems. Experiment 5 did not find any above-chance preferences for high-variability systems, in contrast to Experiment 3 in which participants selected high-variability systems in both framing conditions. This suggests that, in addition to static presentation of system decisions (as in previous research such as Glickman et al., 2018; Kunar et al., 2017; Tsetsos et al., 2012), larger differences in variability between systems were required in order to discover sensitivity to variability and framing effects (as were found in Experiment 3).

6.7. General discussion

Previous research into value integration has shown that participants asked to judge the overall value of a stream of digits can be biased by values that fall at the edges of the distribution (Glickman et al., 2018; Kunar et al., 2017; Tsetsos et al., 2012; Usher et al., 2019) and this can result in paradoxical choices depending on how the question is framed (Glickman et al., 2018; Tsetsos et al., 2012; Usher et al., 2019). The experiments presented in this chapter aimed to explore choice biases when applied to more real-world instances of information presentation and system evaluation. Here this more real-world context took the form of a hypothetical motor manufacturer who was interested in evaluating the performance of systems designed to estimate battery range (Experiments 1 & 2) and charge the battery of an electric vehicle (Experiments 3 – 5). The research is important from a practical viewpoint because purchase and use intentions of systems are influenced by judgements of reliability and trust (Buckley, Kaye & Pradhan, 2018; Choi & Ji, 2015; Vrkljan & Anaby, 2011). Therefore, if the framing of a question can lead to paradoxical choices then clearly this needs to be accounted for and controlled against. At the extreme, identical systems could be either accepted as the best or rejected as the worst depending on how the evaluations are framed. The present study also extends previous literature of system trust, which has shown that users are sensitive to system accuracy (Chancey et al., 2013; Chavaillaz & Sauer, 2017; Chavaillaz, et al., 2016; Madhavan & Phillips, 2010; Yu et al., 2017), by investigating whether users are also sensitive to variability in the system's accuracy and/or framing biases in how the system is evaluated.

From a theoretical view point the work examined whether variability-based biases occur in more general settings and are not restricted to the relatively artificial task of judging the value of streams of rapidly presented digits. In other words, does the phenomena exist

beyond the typical lab-type psychophysical style testing paradigms? Based on previous work, in each experiment, it was predicted that participants would choose the highly variable system at a level above chance, in line with previous findings (Glickman et al., 2018; Tsetsos et al., 2012; Usher et al., 2019).

Participants in Experiments 1 and 2 were asked to choose between pairs of systems that presented the same mean value, but which differed in the variability of their responses. One system produced results that had a relatively high variability and the other system produced responses that had a lower variability. Hence although both systems produced the same average value over a range of trials, one was noisier than the other. After gaining experience with both systems over a block of trials, participants were asked either to pick the best performing system or the worst performing system. The systems also either consistently under- or over-estimated the amount of battery range remaining within a block. Contrary to the expectations based on prior research, participants in Experiment 1 did not choose the high variability system at a level above chance. Participants in Experiment 2 were presented with instantaneous system decisions (rather than building up over time), in order to better match previous methodologies (Glickman et al., 2018; Kunar et al., 2017, Tsetsos et al., 2012). However, again there was no effect of system variability on choice preference. Neither experiment provided any evidence of an effect of whether the system consistently under- or overestimated the true value.

One possible reason for this lack of effects is that performance at the task could be evaluated in multiple directions (i.e. the system could over- or under-estimate the true amount of battery remaining). This is in contrast to experiments conducted by, for example, Tsetsos et al. (2012) and Kunar et al. (2017) where higher values always corresponded to better outcomes. Furthermore, the differences in standard deviation between the high- and low-variability systems were relatively small in Experiments 1 and 2 (*SD* = 10 for the high-variability condition, *SD* = 5 for the low-variability condition). To compare the design with previous work, Tsetsos et al. (2012) used a mean of 45-55 and standard deviations of 10 and 20 in the experiment in which they compared to two distributions with the same mean but different levels of variability. While Kunar et al. (2017) did not directly focus on differences between the variability of two streams, their first three experiments used an *SD* of 20, with subsequent experiments using an average of 16.5 – both larger than either of those used in the present study (Experiments 1 & 2). However, Glickman et al. (2018) used a mean of 40-50 and variances of 10 (*SD* = 3.16) and 20 (*SD* = 4.47) for the low and high-variability conditions, respectively, resulting in a substantially smaller difference in variability in

comparison to the present study. Overall, the smaller standard deviations used in the present study (although larger than those in Glickman et al., 2018) may have contributed to participants' lack of sensitivity to variability. Experiments 3 – 5 were designed to address these possibilities and bring the paradigm closer to the previous work mentioned above.

Experiments 3 – 5 presented participants with systems which aimed to fully charge an electric car battery. This approach was more similar to the previous research in this area because the system output was directly related to performance in a linear fashion. That is, larger bars/values always indicated better performance. Experiment 3 found that participants now chose the highly variable system as both the best and the worst option (i.e. it was selected above chance in both framing conditions), in line with Tsetsos et al. (2012). This suggests that the framing of the question was successful in directing participants' attention towards different values when evaluating the performance of the presented instruments. Specifically, being asked to choose the best system made higher values more salient, with attention being drawn towards lower values when participants were asked to select the worst system. This mimics previous findings that have used streams of numbers rather than bar graphs and demonstrates the existence of a paradoxical choice in this new context.

The findings of Experiment 3 are consistent with the theory of selective integration (Usher et al., 2019), which posits that judgments of overall value are made by combining the results of multiple serial comparisons, and that people attend to different features of the decision space depending on the goal of the task (e.g. Glickman et al., 2018; Kunar et al., 2017). In the case of Experiment 3, participants demonstrated this type of behaviour by selecting the higher-variability system as both the best and the worst-performing systems, suggesting that the goal of the task itself (how the question was framed) influenced their attention to different aspects of the systems' behaviour (higher numbers when asked to select the best system, lower when asked to select the worst). This provides some evidence of framing biases in value integration in the domain of computerised systems.

The findings of Experiment 3 (as well as the focus of the experiments presented in this chapter) relate to the focus of Chapter 5, in terms of the focus on distance from the correct solution. In Experiments 3 – 5 of the present chapter, the high-variability systems were frequently closer to the correct solution (100% battery charge) compared to the low-variability systems, but were also frequently further away from the correct solution (on trials where the system put in a very low amount of charge). The results of Experiment 3 suggest that being asked to select the best-performing system biased participants' attention towards trials where the high-variability system was very close to fully charging the battery, whereas

being asked to select the worst-performing system biased attention towards trials where the high-variability system put very little charge into the battery. This resulted in the high-variability system being chosen as both the best and worst alternative, dependant on framing, and suggested that participants were sensitive to distance from the correct solution. Similarly, the last four experiments in Chapter 5 (Experiments 3a-b & 4a-b) focused on the interaction between system accuracy and distance from the correct solution, by examining trust in systems aiming to reach a target position on a dial. Across these four experiments, there was some evidence that participants were slightly more forgiving of systems that were less accurate if they were closer to the correct solution when they could not reach it. The findings of Chapter 5 build on previous research showing that users of computerised systems are sensitive to accuracy (Chancey et al., 2013; Chavaillaz & Sauer, 2017; Chavaillaz, et al., 2016; Madhavan & Phillips, 2010; Yu et al., 2017), while the work presented in this chapter builds on this further by providing (some) evidence that this sensitivity to distance from the correct solution can be influenced by framing biases.

However, the positive finding of Experiment 3 must be considered within the context of the other four experiments presented in this chapter, which did not find any above-chance influence of variability on participants' preferences. Indeed, Bayes analyses consistently indicated evidence in favour of the null hypothesis. As discussed earlier in this section, Experiments 1 and 2 departed from previous research by using a task in which the system could over or under-estimate battery life, rather than using a task where performance can be measured in a single direction (e.g. Glickman et al., 2018; Kunar et al., 2017; Tsetsos et al., 2012). Experiment 3 increased the difference in variability between the high- and low-variability systems, and also changed the task to one where performance could be measured in a single direction. Experiments 4 and 5 aimed to clarify the picture further by testing the influence of the amount of feedback provided (Experiment 4: Participants saw the system charging the battery, rather than just the outcome) and the size of the differences in variability between the two systems (Experiment 5 used the same differences as Experiments 1 and 2). Notably, the differences in variability (SD = 5, SD = 24, for the low- and high-variability conditions, respectively) in Experiments 3 and 4 were larger those used in comparable value psychophysics experiments (Glickman et al., 2018: SD = 3.16 vs SD = 4.47); Kunar et al., 2017: SD = ~16.5 or 20; Tsetsos et al., 2012; SD = 10 vs SD = 20). Neither Experiment 4 nor Experiment 5 found any sensitivity to variability in participants' choices.

Overall, these results suggest that three conditions are required for users of battery-related systems in electric vehicles to display framing biases and sensitivity to variability: i)

relatively large differences in variability between systems (which Experiments 1, 2 & 5 did not have); ii) system performance measured in a single direction (which Experiments 1 & 2) did not have); iii) instant presentation of system decisions (which Experiments 1, 2 & 4 did not have). Only Experiment 3 met all of these conditions and revealed sensitivity to system variability and susceptibility to framing biases. It is also important to note that, although participants' bias towards selecting the highly variable system was significantly above chance, the size of the bias was fairly small. Specifically, participants selected the highly variable system as the best system 58% of the time in the positive frame condition, and those in the negative frame condition selected it as the worst system 52% of time (where chance is 50%). However, this is reasonably comparable to other similar studies. For instance, participants in Glickman et al. (2018) preferred the more highly variable stream of numbers approximately 60% of the time, with Tsetsos et al. (2012) finding a 62% preference for the broader, more highly variable stream.

The lack of sensitivity to variability and framing biases in four out of five experiments suggests that the phenomena might be quite fragile, as least when considered in the context of the evaluation of more real-world systems thus questioning its generalisability (Glickman et al., 2018; Kunar et al., 2017; Tsetsos et al., 2012; Usher et al., 2019). For instance, the framing effects found in experiments conducted by Tsetsos et al. (2012), which used streams of numbers in context of sampling payoffs from different slot machines, may not apply (or only apply in limited settings) to more 'real-world' contexts such as evaluating the performance of in-vehicle computerised systems. However, the experiments presented in this chapter make several departures from previous value integration work (tasks with multiple directions of performance; using small differences in variability between systems, and providing dynamic information on system decisions), in addition to testing whether framing biases occur in the context of system trust. It may be helpful for future research to replicate the results of Experiment 3, to test whether more 'direct' extensions of findings such as Tsetsos et al. (2012) can reliably uncover the same framing effects. It is also worth noting that the lack of effects in four of the experiments is unlikely to have resulted from issues of sample size. Each experiment presented in this chapter recruited sixty participants (total N = 300), whereas comparable studies have used much smaller samples and have found positive results (e.g. Kunar et al., 2017 recruited 20 participants per study, with Tsetsos et al., 2012 recruiting 67 across three experiments).

Another key difference between previous value psychophysics experiments (Glickman et al., 2018; Kunar et al., 2017; Tsetsos et al., 2012) and the work presented in this

chapter is the timeframe over which users evaluated the systems. In previous studies, participants have been asked to compare rapidly-presented streams of digits (e.g. one digit every 250ms in Kunar et al., 2017), making very quick decisions on overall value. In contrast, the experiments presented participants with longer blocks of trials, in which each trial was three seconds long, giving users a much larger timeframe over which to compare the performance of the two systems and consider their decisions. This is represented as participants making a decision per trial (Glickman et al., 2018; Kunar et al., 2017; Tsetsos et al., 2012), versus making a decision per block of trials in the present study. It is possible that these different timeframes tap into separate processes for comparing the values of two streams/systems, which may have contributed to the differences in results. This could reflect the distinction between 'System 1' and 'System 2' thinking (Kahneman, 2011), with previous value psychophysics studies requiring a faster, more automatic judgment (in line with System 1), and the design of the present study providing participants with more time to reflect and consider their decision on the systems' performance (in line with System 2). This distinction is important because, in many 'real-world' applications (e.g. choosing between two cars), decision makers usually have more time to compare and consider in depth, in a more 'System 2' method of judgment. The longer timeframe over which to make value judgments in the present study, therefore, is arguably more reflective of typical real-world decisions relevant to system trust, compared to the rapid decisions required by previous value-psychophysics experiments. Future research would benefit from exploring this distinction in more depth, to shed light on how System 1 and 2 thinking relates to the psychological processes behind value integration and evaluation of computerised systems, as well as how well classic value psychophysics designs reflect real-world decision making in this domain.

6.8. Summary and conclusions

Against the backdrop of previous research demonstrating that value judgments are susceptible to biases in framing (Glickman et al., 2018; Kunar et al., 2017; Tsetsos et al., 2012; Usher et al., 2019), this chapter five experiments testing the effect of framing on evaluation of computerised systems. Participants across all five experiments were asked to compare pairs of systems either estimating (1 & 2) or generating (3 – 5) battery charge, where both systems produced the same mean value but different standard deviations, and were asked to select either the best or worst-performing system. Experiment 3 found that participants were more likely to select the highly variable system as both the best and the worst option, supporting the paradoxical framing effects found by previous studies such as

Tsetsos et al. (2012). However, none of the other four experiments revealed any preference for either the high or low-variability system. The results of several experiments in this chapter suggest that value integration effects may only occur in the domain of system trust when i) system performance can be measured in a single direction, ii) the differences in system variability are large, and iii) system decisions are presented instantaneously. These findings provide some limited support for the existence of value integration biases in evaluation of computerised systems, but the potentially stringent conditions for these effects might suggest some limitations of the generalisability of framing biases in value integration. Future work that further clarifies the conditions under which users of computerised systems are susceptible to biases in value integration would be valuable.

7. General discussion

This thesis presented a total of twenty-three experiments which investigated the psychological factors influencing evaluations of autonomous vehicles and in-vehicle instrumentation, using a combination of driving simulator and laboratory-based methodologies. The thesis is divided into two complimentary streams, the first focusing on evaluation of autonomous driving styles (Chapters 3 & 4), with the second focusing on trust of in-vehicle display systems (Chapters 5 & 6). This chapter will summarise the scope and findings of both of these streams and will consider how these relate to both the existing literature and practical considerations.

Chapter 1 (General introduction) discussed previous work relating to both research streams, although specific sections of the literature are discussed in more depth within the introductory sections of the empirical chapters. Overall, previous work into evaluations of autonomous vehicles has found variables such as demographics of users, appearance of vehicles, information presented to occupants and autonomous driving style (e.g. choice of following headway) to be influential. However, less work has been conducted on interactions between autonomous vehicles and other traffic, as well as the role of traffic context, which led Chapters 3 and 4 to focus on overtaking scenarios involving autonomous and human-driven vehicles. Previous work on trust of computerised systems has revealed factors such as user characteristics, information provision and objective system accuracy to be influential, with research largely focusing on evaluations of specific technologies. Chapter 5 built on previous work by evaluating system trust using a more context-neutral methodology and by investigating the influence of relative ranking, while Chapter 6 tested for the existence of framing-based decision biases in value integration in relation to in-vehicle instrumentation.

Chapter 3 presented six experiments (three simulator, three video-based) investigating the influence of pull-in distance, vehicle perspective, presence of a third vehicle, following distance and level of immersion on acceptance and evaluations of autonomous overtaking manoeuvres (Ritchie et al., 2019). Overall, the findings revealed a sharp increase in the positivity of ratings with increasing pull-in distance up to approx. 28m, plateauing and beginning a downward trend beyond this point, irrespective of whether the participant was in an autonomous vehicle performing the overtake or in a human-driven vehicle being overtaken. The presence of a third vehicle following the overtaking vehicle did not appear to influence perceptions of the overtake (from the perspective of being overtaken), but participants showed a trend of rating longer pull-ins more negatively (and shorter pull-ins more positively) if a third

vehicle was closely following the overtaking vehicle. Comparison of the simulator studies with video-based analogues revealed that the sharpest pull-ins were not rated as negatively when the level of immersion was lower. Physiological and behavioural variables in the simulator experiments also suggested the generation of a threat response to being overtaken very sharply, as well as some evidence that drivers decreased their speed in response to being overtaken at the shortest pull-in distances. These results suggest that a threshold of acceptability may exist for overtaking manoeuvres, rather than perceptions varying linearly with increasing pull-in distance. Another important implication from an implementation perspective is the influence of a third vehicle's following distance, as it suggests that whether a given manoeuvre is acceptable to other road users depends not only on the autonomous vehicle's behaviour itself, but also on contextual factors over which the occupant or vehicle would likely have no control. This suggests that the processing of local traffic context might be useful for modulating the behaviour of at least certain manoeuvres that are performed autonomously.

Chapter 4 also focused on autonomous overtaking, but more on the influence of continually being overtaken by or overtaking other vehicles in a stream of traffic. Three driving simulator-based experiments investigated the effects of lane position, vehicle speed, information presentation and traffic context on occupant perceptions of highway journeys in a fully autonomous vehicle. In the first two experiments, occupants showed a strong preference for overtaking compared to being overtaken, even when travelling at the maximum legal speed limit (and therefore being overtaken by vehicles violating the rules of the road). This reduction in satisfaction did not disappear when occupants were given additional nudges to increase the salience of the fact their vehicle could not legally increase speed. A third experiment revealed that occupants preferred overtaking to being overtaken when following another car, but not when following a lorry, possibly due to reduced satisfaction from following a slow-moving vehicle eliminating the benefit of overtaking other vehicles. These findings are line with those of Chapter 3, in that evaluation of journeys in autonomous vehicles appear to be sensitive to traffic context, and that rule-violating driving behaviour on the part of human drivers can indirectly influence how the behaviour of an autonomous vehicle is evaluated.

The findings of this first stream add to the literature by shedding light on the psychological and contextual factors influencing perceptions of autonomous vehicles, as well as interactions between human-driven and autonomous vehicles. This is an important contribution because concerns have been raised around integration of autonomous vehicles into mixed traffic conditions, particularly that they may struggle to understand human driving conventions and could be 'bullied' by other drivers (Hancock, 2019; Hancock, Nourbakhsh &

Stewart, 2019; Chater, Misyak, Watson, Griffiths & Mouzakitis, 2018; Tennant, Howard, Franks, Bauer & Stares, 2016; Tennant, 2015). However, despite these concerns, research into evaluations of autonomous vehicles has not tended to focus on interactions with human drivers. The findings of both Chapters 3 and 4 suggest that evaluations of autonomous driving styles (both from an occupant perspective and from that of other road users) can be influenced by the behaviour of the surrounding traffic, alarmingly in situations where human drivers are in violation of the rules of the road (e.g. tailgating, exceeding the speed limit). From a practical perspective, it would be valuable for autonomous vehicles to be programmed with some awareness of how perceptions of a given driving style might be influenced by traffic context. It may also be helpful for autonomous vehicles to provide information to occupants on how the traffic context influences the range of options the autonomous vehicle has available (e.g. if the vehicle cannot enter the right-hand lane on a dual carriageway without substantially exceeding the speed limit), in order to improve occupant trust. While providing simple prompts did not appear to improve occupants' satisfaction in Chapter 4, more detailed/precise descriptions from the autonomous vehicle on the impact of context could be a promising area for future research. It is also important to note that the experiments in Chapters 3 and 4 focused on motorway and dual carriageway environments, used largely student populations (young adults), and either used a fully human-driven vehicle (Chapter 3) or full automation, with no intermediary modes of automation examined. While beyond the scope of this thesis, it would be valuable for future research to examine whether these findings apply to other road settings, age groups, and different levels of automation.

Chapter 5 presented nine experiments on the influence of accuracy, relative rank position, task complexity and closeness to solution on trust of in-vehicle instrumentation. The research expanded on previous work by using a more context-neutral methodology (in order to avoid any possible confounds of domain on trust ratings), and also by testing whether findings regarding the wider influence of rank position on decision making (Stewart, 2009; Stewart, Chater & Brown, 2006; Walasek & Stewart, 2015; Walasek & Stewart, 2019) apply to the domain of system trust. Participants viewed systems that either aimed to generate a target colour or reach a target position on a dial and were asked to indicate their trust in the system. Overall, participants showed a sharp reduction in trust below 100% accuracy, with further drops in trust then becoming much shallower below this initial drop. This low tolerance of error is comparable to the pattern of ratings over pull-in distance in Chapter 3 and was also in line with previous work on system trust (Chavaillaz & Sauer, 2017; Chavaillaz, Wastell & Sauer, 2016; Yu et al., 2017). Surprisingly, this pattern was not sensitive to relative rank position,

with systems with low levels of objective accuracy always being rated poorly even if they were the best option available, in contrast with previous work demonstrating a substantial role of rank position in judgement and decision making more widely. Trust ratings were not sensitive to task complexity but did show a shallower drop with decreasing accuracy when the system was closer to the correct solution on trials where it did not make the correct decision. These findings suggest a substantial influence of objective measures of performance (in this case, accuracy) on trust of instrumentation, with a smaller, more limited mediating role of some contextual factors (closeness to solution), analogous to the limited influence of following distance (compared to the larger effect of pull-in distance) in Chapter 3.

Chapter 6 also investigated factors influencing evaluations of in-vehicle instrumentation, but from a more theoretical perspective, testing for framing biases in judgments of value when comparing the performance of battery charging and battery estimation systems. Across five experiments, participants were presented with pairs of systems that achieved the same overall performance (e.g. % of battery charged), but with different standard deviations, such that one system was more variable than the other. These manipulations were based on previous findings that, when asked to select the best value stream out of two rapidly presented streams of digits that had the same mean but different standard deviations, participants choose the more variable system, but also choose the same stream when asked to reject the lower value stream. This represents a paradoxical choice because people selected the very same stream as being the highest or lowest value depending on how the task was phrased (i.e. select the best or reject the worst: Glickman, Tsetsos & Usher, 2018; Tsetsos, Chater & Usher, 2012; Usher, Tsetsos, Glickman & Chater, 2019). Out of five experiments in Chapter 6, only one found evidence of such framing biases. Overall, the findings of Chapter 6 suggested that framing biases in value integration either do not apply to the domain of in-vehicle display systems, or only apply under a specific set of experimental conditions.

This second, more theoretical stream of the thesis aimed to extend previous research by testing whether evaluation of computerised systems is influenced by rank-based effects (Stewart, 2009; Stewart et al., 2006; Walasek & Stewart, 2015; 2019) and framing-based paradoxical choices (Glickman et al., 2018; Tsetsos et al., 2012; Usher et al., 2019), which have been shown to be influential in decision making more broadly. Overall the results suggest that, contrary to heavily rank-based perspectives, trust of computerised systems was very sensitive to absolute accuracy, with no mediating effect of relative rank position. It is possible that, in the case of systems attempting a relatively simple task (e.g. generating a target colour), users' error tolerance is so low that less accurate systems are not seen more favourably due to

ranking well against alternatives. Future research could clarify this by examining possible interactions between relative rank position and task complexity. In addition, participants in the experiments of Chapter 5 were always comparing systems of the same domain – it would be interesting to examine whether rank-based effects occur when comparing systems with multiple different domains/with different goals (as users of systems are often required to do in everyday life). While the findings of Chapter 6 suggest that framing-based choice biases might apply to battery charging systems to some extent, this area of research would benefit from further clarifying the exact conditions under which such biases occur. If these effects of framing on system evaluation are shown to be reliable by further research, then this could have practical applications for areas such as product evaluation, as it would suggest that the framing of questions can influence users' preferences when asked to compare multiple systems. However, given the general lack of framing bias obtained across most of the experiments presented here, further research might also examine the extent to which framing biases have impact in real-world systems and contexts.

7.1. Summary and conclusions

In summary, the work presented in this thesis examined psychological factors influencing evaluations of autonomous vehicles and computerised systems. Using a mixture of driving simulator and laboratory-based methodologies, two complimentary streams examined the influence of various factors involving both the system's performance and the context of the task on users' evaluations and levels of trust. While the first stream of thesis focused on the more applied question of how autonomous vehicles can interact successfully with human-driven traffic, the second stream focused on applying more theoretical findings from the psychology of decision making in general to evaluations of system trust, such as the influence of relative rank and framing biases. Overall, findings across the two streams suggested very strong levels of sensitivity to the performance/behaviour of autonomous vehicles and computerised systems (e.g. pull-in distance in overtaking, choice of lane position in traffic, system accuracy), with a more limited influence of contextual factors (e.g. traffic context, framing effects) on evaluations of the behaviour of autonomous vehicles and computerised systems. While this work offers the benefit of a wide range of methodological approaches (utilising both higher-immersion simulator settings and context neutral laboratory-based tasks), it is necessarily limited in that it did not examine evaluations of autonomous vehicles and systems in a more real-world setting. Future research could benefit from examining how the effects of autonomous driving behaviours, system performance, and task/traffic context apply

to more immersive settings (e.g. test-track studies) and comparing this to the results of the research presented within this thesis.

References

Abe, G., Sato, K., & Itoh, M. (2017). Driver trust in automated driving systems: The case of overtaking and passing. *IEEE Transactions on Human-Machine Systems, 48* (1), 85 – 94.

Abe, G., Sato, K., Uchida, N., & Itoh, M. (2019). Effect of changes in levels of automated driving on manual control recovery. *IFAC PapersOnLine, 52-19,* 79 – 84.

Angrilli, A., Cherubini, P., Pavese, A., & Manfredini, S. (1997). The influence of affective factors on time perception. *Perception & Psychophysics*, *59*(6), 972 – 982.

Asaithambi, G., & Shravani, G. (2017). Overtaking behaviour of vehicles on undivided roads in non-lane based mixed traffic conditions. *Journal of Traffic and Transportation Engineering (English Edition), 4*(3), 252 – 261.

Balk, S. A., Jackson, S., & Philips, B. (2017). Preferred following distance and performance in an emergency event while using cooperative adaptive cruise control. *Proceedings of the Ninth International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design,* pp. 30 – 36.

Bansal, P., & Kockelman, K. M. (2018). Are we ready to embrace connected and self-driving vehicles? A case study of Texans. *Transportation, 45,* 641 – 675.

Bar-Gera, H., & Shinar, D. (2004). The tendency of drivers to pass other vehicles. *Transportation Research Part F, 8,* 429 – 439.

Barg-Walkow, L. H., & Rogers, W. A. (2016). The effect of incorrect reliability information on expectations, perceptions and use of automation. *Human Factors, 58*(2), 242 – 260.

Barthou, A., Kemeny, A., Reymond, G., Merienne, F., & Berthoz, A. (2010). Driver trust and reliance on a navigation system: Effect of graphical display. *Driving Simulator Conference* (pp. 199 – 210). September 2010, Paris, France.

Basu, C., Yang, Q., Hungerman, D., Singhal, M., & Dragan, A. D. (2017). Do you want your autonomous car to drive like you? *HRI '17, March 06 – 09 2017,* Vienna, Austria.

Bauer, B., Jolicoeur, P., & Cowan, W. B. (1996). Visual search for colour targets that are or are not linearly separable from distractors. *Vision Research, 36*(10), 1439 – 1465.

Becker, F., & Axhausen, K. W. (2017). Literature review on surveys investigating the acceptance of automated vehicles. *Transportation, 44,* 1293 – 1306.

Beggiato, M., Hartwich, F., & Krems, J. (2019). Physiological correlates of discomfort in automated driving. *Transportation Research Part F, 66,* 445 – 458.

Beller, J., Heesen, M., & Vollrath, M. (2013). Improving the driver-automation interaction: An approach using automation uncertainty. *Human Factors, 55* (6), 1130 – 1141.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B (Methodological), 57*(1), 289 – 300.

Boucsien, W. (2012). *Electrodermal Activity (2nd Ed).* New York: Springer.

Braithwaite, J. J., Watson, D. G., Jones, R., & Rowe, M. (2015). A guide for analysing electrodermal activity (EDA) & skin conductance responses (SCRs) for psychological experiments (revised version: 2.0). Technical report, 2nd version: Selective Attention & Awareness Laboratory (SAAL), Behavioural Brain Sciences Centre, University of Birmingham, UK.

Brown, G. D. A., Gardner, J., Oswald, A. J., & Qian, J. (2008). Does wage rank affect employees' well-being? *Industrial Relations, 47*(3), 355 – 389.

Buckley, L., Kaye, S., & Pradhan, A. K. (2018). Psychosocial factors associated with intended use of automated vehicles: A simulated driving study. *Accident Analysis and Prevention, 115,* 202 – 208.

Chancey, E. T., Proaps, A., & Bliss, J. P. (2013). The role of trust as a mediator between signalling system reliability and response behaviours. *Proceedings of the Human Factors and Ergonomics Society 57th Annual Meeting,* 285 – 289.

Chancey, E. T., Yamani, Y., Brill, J. C., & Bliss, J. P. (2017). Effects of alarm system error bias and reliability on performance measures in a multi-tasking environment: Are false alarms really worse than misses? *Proceedings of the Human Factors and Ergonomics Society 2017 Annual Meeting,* 1621 – 1625.

Chater, N., Misyak, J., Watson, D. G., Griffiths, N, & Mouzakitis, A. (2018). Negotiating the traffic: Can cognitive science help make autonomous vehicles a reality? *Trends in Cognitive Sciences, 22*(2), 93 – 95.

Chavaillaz, A., & Sauer, J. (2017). Operator adaptation to changes in system reliability under adaptable automation. *Ergonomics, 60*(9), 1261 – 1272.

Chavaillaz, A., Schwaninger, A., Michel, S., & Sauer, J. (2020). Some cues are more equal than others: Cue plausibility for false alarms in baggage screening. *Applied Ergonomics, 82,* 102916.

Chavaillaz, A., Wastell, D., & Sauer, J. (2016). System reliability, performance and trust in adaptable automation. *Applied Ergonomics, 52,* 333 – 342.

Cheng, F. F., & Wu, C. S. (2010). Debiasing the framing effect: The effect of warning and involvement. *Decision Support Systems, 49,* 328 – 334.

Choi, J. K., & Ji, Y. G. (2015). Investigating the importance of trust on adopting an autonomous vehicle. *International Journal of Human-Computer Interaction, 31,* 692-702.

Cramer, H., Evers, V., Kemper, N., & Wielinga, B. (2008). Effects of autonomy, traffic conditions and driver personality traits on attitudes and trust towards in-vehicle agents. *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology,* pp. 477 – 482.

Critchley, H. D. (2002). Electrodermal responses: What happens in the brain. *The Neuroscientist, 8*(2), 132 – 142.

Daoutis, C. A., Pilling, M., & Davies, I. R. L. (2006). Categorical effects in visual search for colour. *Visual Cognition, 14*(2), 217 – 240.

Daviaux, Y., Bonhomme, E., Ivers, H., de Sevin, E., Micoulaud-Franchi, J., Bioulac, S., Morin, C. M., Philip, P., & Altena, E. (2020). Event-related electrodermal responses to stress: Results from a realistic driving simulator scenario. *Human Factors, 62*(1), 138 – 151.

Dewe, H., Watson, D. G., & Braithwaite, J. J. (2016). Uncomfortably numb: new evidence for suppressed emotional reactivity in response to body-threats in those predisposed to sub-clinical dissociative experiences. *Cognitive Neuropsychiatry,* DOI: 10.1080/13546805.2016.1212703.

Dixon, S. R., Wickens, C. D., & McCarley, J. S. (2007). On the independence of compliance and reliance: Are automation false alarms worse than misses? *Human Factors, 49*(4), 564 – 572.

Driverless car market watch. http://www.driverless-future.com/?page_id=384. Last accessed 18/08/2020.

Ekman, F., Johansson, M., Bligård, L., Karlsson, M., & Strömberg, H. (2019). Exploring automated vehicle driving styles as a source of trust information. *Transportation Research Part F, 65,* 268 – 279.

Ert, E., & Erev, I. (2013). On the descriptive value of loss aversion in decisions under risk: Six clarifications. *Judgment and Decision Making, 8*(3), 214 – 235.

Faerevaang, C. L., Nguyen, B. A., Jimenez, C. A., & Jentsch, F. (2017). Attitudes toward unreliable diagnostic aiding in dangerous task environments. *Proceedings of the Human Factors and Ergonomics Society 2017 Annual Meeting,* 1161 – 1165.

Fleetwood, J. (2017). Public health, ethics, and autonomous vehicles. *American Journal of Public Health, 107*(4), 532 – 538.

Frith, C. D., & Allen, H. A. (1983). The skin conductance orienting response as an index of attention. *Biological Psychology, 17*(1), 27 – 39.

Glickman, M., Tsetsos, K., & Usher, M. (2018). Attentional selection mediates framing and risk-bias effects. *Psychological Science, 29*(12), 2010 – 2019.

Gouy, M., Wiedemann, K., Stevens, A., Brunett, G., & Reed, N. (2014). Driving next to automated vehicle platoons: How do short time headways influence non-platoon drivers' longitudinal control? *Transportation Research Part F, 27,* 264 – 273.

Haboucha, C. J., Ishaq, R., & Shiftan, Y. (2017). User preferences regarding autonomous vehicles. *Transportation Research Part C, 78,* 37 – 49.

Hancock, P. A. (2019). Some pitfalls in the promises of automated and autonomous vehicles, *Ergonomics, 62*(4), 479 – 495.

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors, 53*(5), 517 – 527.

Hancock, P. A., Nourbakhsh, I., & Stewart, J. (2019). On the future of transportation in an era of automated and autonomous vehicles. *PNAS, 116*(16), 7684 – 7691.

Harper, C. D., Hendrickson, C. T., Mangones, S., & Samaras, C. (2016). Estimating potential increases in travel with autonomous vehicles for the non-driving, elderly and people with travel-restrictive medical conditions. *Transportation Research Part C, 72,* 1 – 9.

Hartwich, F., Witzlack, C., Beggiato, M., & Krems, J. F. (2018). The first impression counts – A combined driving simulator and test track study on the development of trust and acceptance of highly automated driving. *Transportation Research Part F, 98,* 207 – 220.

Heikoop, D. D., de Winter, J. C. F., van Arem, B., & Stanton, N. (2017). Effects of platooning on signal-detection performance, workload, and stress: A driving simulator study. *Applied Ergonomics, 60,* 116 -127.

Helldin, T., Falkman, G., Riveiro, M., & Davidsson, S. (2013). Presenting system uncertainty in automotive UIs for supporting trust calibration in autonomous driving. *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (Automotive UI, 13),* October 28 – 30, Eindhoven, The Netherlands, pp. 210 – 217.

Hjälmdahl, M., Krupenia, S., & Thorslund, B. (2017). Driver behaviour and driver experience of partial and fully automated truck platooning – a simulator study. *European Transportation Research Review, 9*(8), DOI: 10.1007/s12544-017-0222-3.

Hodsoll, J., & Humphreys, G. W. (2001). Driving attention with the top down: The relative contribution of target templates to the linear separability effect in the size dimension. *Perception & Psychophysics, 63*(5), 918 – 926.

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors, 57*(3), 407 – 434.

Hohenberger, C., Spörrle, M., & Welpe, I. M. (2016). How and why do men and women differ in their willingness to use automated cars? The influence of emotions across different age groups. *Transportation Research Part A, 94,* 374-385.

Howard, D., & Dai, D. (2014). Public perceptions of self-driving cars: the case of Berkeley, California. *Proceedings of the 93rd Transportation Research Board Annual Meeting, Washington, D. C.,* pp. 1 – 21.

Huerta, E., Glandon, T., & Petrides, Y. (2012). Framing, decision-aid systems, and culture: Exploring influences on fraud investigations. *International Journal of Accounting Information Systems, 13,* 316 – 333.

Hulse, L. M., Xie, H., & Galea, E. R. (2018). Perceptions of autonomous vehicles: Relationships with road users, risk, gender and age. *Safety Science, 102,* 1 – 13.

Hutchinson, T. P. (2008). Tailgating (CASR046), Centre for Automotive Safety Research, The University of Adelaide, Australia.

Igliński, H., & Babiak, M. (2017). Analysis of the potential of autonomous vehicles in reducing the emissions of greenhouse gases in road transport. *Procedia Engineering, 192,* 353 – 358.

Jarosz, A., F., & Wiley, J. (2014). What are the odds? A practical guide to computing and reporting Bayes factors. *Journal Of Problem Solving, 7,* 2 – 29.

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

Jones, E. C., Leibowicz, B. D. (2019). Contributions of shared autonomous vehicles to climate change mitigation. *Transportation Research Part D, 72,* 279 – 298.

Jonides, J., & Yantis, S. (1988). Uniqueness of abrupt visual onset in capturing attention. *Perception & Psychophysics, 43*(4), 346 – 354.

Kahneman, D. (2011). *Thinking, fast and slow.* Allen Lane.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*(2), 263 – 292.

Kaltenbach, E., & Dolgov, I. (2017). On the dual nature of transparency and reliability: Rethinking factors that shape trust in automation. *Proceedings of the Human Factors and Ergonomics Society 2017 Annual Meeting.* (pp. 308 – 312). DOI: 10.1177/1541931213601558.

Kennedy, R. S., Lane, N. E., Berbaum, K. S., & Lilienthal, M. G. (1993). Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *The International Journal of Aviation Psychology, 3*(3), 203 – 220.

Kinnear, N., Helman, S., Wallbank, C., & Grayson, G. (2015). An experimental study of factors associated with driver frustration and overtaking intentions. *Accident Analysis and Prevention, 79,* 221 – 230.

Klein, R. M. (2000). Inhibition of return. *Trends in Cognitive Sciences, 4*(4), 138 – 147.

Kopelias, P., Demiridi, E., Vogiatzis, K., Skabardonis, A., & Zafiropoulou, V. (2020). Connected & autonomous vehicles – Environmental impacts – A review. *Science of the Total Environment, 712: 135237*, 1 – 7.

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

Kunar, M. A., Watson, D. G., Tsetsos, K., & Chater, N. (2017). The influence of attention on value integration. *Attention, Perception, & Psychophysics, 79,* 1615 – 1627.

Kyriakidis, M., Happee, R., & de Winter, J. F. C. (2015). Public opinion on automated driving: Results of an international questionnaire among 5000 respondents. *Transportation Research Part F, 32,* 127-140.

Lacson, F. C., Wiegmann, D. A., & Madhavan, P. (2005). Effects of attribute and goal framing on automation reliance and compliance. *Proceedings of the Human Factors and Ergonomics Society 49th Annual Meeting* (pp. 482 – 486). DOI: 10.1177/154193120504900357.

Lee, J., Kim, K. J., Lee, S., & Shin, D. (2015). Can autonomous vehicles be safe and trustworthy? Effects of appearance and autonomy of unmanned systems. *International Journal of Human-Computer Interaction, 31,* 682 – 691.

Lewis-Evans, B., De Waard, D., & Brookhuis, K. A. (2010). That's close enough – a threshold effect of time headway on the experience of risk, task difficulty, effort, and comfort. *Accident Analysis and Prevention, 42,* 1926 – 1933.

Lim, B.Y., Dey, A. K., & Avrahami, D. (2009). Why and why not explanations improve the intelligibility of context-aware intelligent systems. *CHI 2009, Studying Intelligent Systems.* (pp. 2119 – 2128). April 9th, 2009. Boston, MA, USA.

Litman, T. (2020). Autonomous vehicle implementation predictions: Implications for transport planning. Report for the Victoria Transport Policy Institute, 2020.

Look Ma, No Hands! 05/09/2012. IEEE News Releases. https://www.ieee.org/about/news/2012/5september-2-2012.html. Last accessed 18/08/2020.

Lu, D., Li, Z., & Huang, D. (2017). Platooning as a service of autonomous vehicles. 2017 IEEE 18th International Symposium on "World of Wireless, Mobile and Multimedia Networks" (WoWMoM), 1 – 6.

Madhavan, P., & Phillips, R. R. (2010). Effects of computer self-efficacy and system reliability on user interaction with decision support systems. *Computers in Human Behaviour, 26,* 199 – 204.

Madhavan, P., Wiegmann, D. A., & Lacson, F. C. (2006). Automation failures on tasks easily performed by operators undermine trust in automated decision aids. *Human Factors, 8*(2), 241 – 256.

Madigan, R., Nordhoff, S., Fox, C., Ezzati Amini, E., Louw, T., Wilbrink, M., Schieben, A., & Merat, N. (2019). Understanding interactions between Automated Road Transport Systems and other road users: A video analysis. *Transportation Research Part F, 66,* 196 – 213.

Marsh, S., & Dibben, M. R. (2003). The role of trust in information science and technology. *Annual Review of Information Science and Technology, 37,* 465 – 498.

McGuirl, J. M., & Sarter, N. B. (2006). Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. *Human Factors, 48*(4), 656-665.

Meng, M., Rau, A., & Mahardhika, H. (2018). Public transport travel time perception: Effects of socioeconomic characteristics, trip characteristics and facility usage. *Transportation Research Part A, 114,* 24 – 37.

Merat, N., Jamson, A. H., Lai, F. C. H., Daly, M., & Carsten, O. M. J. (2014). Transition to manual: Driver behaviour when resuming control from a highly automated vehicle. *Transportation Research Part F, 27,* 274 – 282.

Morando, M. M., Tian, Q., Truong, L. T., & Vu, H. L. (2018). Studying the safety impact of autonomous vehicles using simulation-based surrogate safety measures. *Journal of Advanced Transportation, 2018,* 1 – 11, doi: 10.1155/2018/6135183.

Mullett, T. L., & Tunney, R. J. (2013). Value representations by rank order in a distributed network of varying context dependency. *Brain and Cognition, 82,* 76 – 83.

National Office of Statistics. (2018). Table SPE0111. Free flow vehicle speeds by road type and vehicle type in Great Britain, 2018. https://www.gov.uk/government/statistical-data-sets/vehicle-speed-compliance-statistics-data-tables-spe. Last accessed 18/08/2020.

Nowakowski, C., O'Connell, J., Shladover, S. E., & Cody, D. (2010). Cooperative adaptive cruise control: Driver acceptance of following gap settings less than one second. *Proceedings of the Human Factors and Ergonomics Society 54th Annual Meeting,* pp. 2033 – 2037.

Ozturk, O., Shayan, S., Liszkowski, U., & Majid, A. (2013). Language is not necessary for color categories. *Developmental Science, 16*(1), 111 – 115.

Pak, R., Fink, N., Price, M., Bass, B., & Sturre, L. (2012). Decision support aids with anthropomorphic characteristics influence trust and performance in younger and older adults. *Ergonomics, 55*(9), 1059-1072.

Pak, R., Rovira, E., McLaughlin, A. C., & Baldwin, N. (2017). Does the domain of technology impact user trust? Investigating trust in automation across different consumer-oriented domains in young adults, military, and older adults. *Theoretical Issues in Ergonomics Science, 18*(3), 199 – 220.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors, 39*(2), 230 – 253.

Payre, W., Cestac, J., & Delhomme, P. (2016). Fully automated driving: Impact of trust and practice on manual recovery. *Human Factors, 58*(2), 229 – 241.

Penmetsa, P., Adanu, E. K., Wood, D., Wang, T., & Jones, S. L. (2019). Perceptions and expectations of autonomous vehicles – A snapshot of vulnerable road user opinion. *Technological Forecasting & Social Change, 143,* 9 – 13.

Pettigrew, S., Talati, Z., & Norman, R. (2018). The health benefits of autonomous vehicles: public awareness and receptivity in Australia. *Australian and New Zealand Journal of Public Health, 42*(5), 480 – 483.

Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology, 32,* 3 – 25.

Posner, M. I., & Cohen, Y. (1984). Components of visual orienting. *Attention and Performance X: Control of Language Processes, 32,* 531 – 556.

Posner, M. I., Rafal, R. D., Choate, L. S., & Vaughan, J. (1985). Inhibition of return: Neural basis and function. *Cognitive Neuropsychology, 2*(3), 211 – 228.

Rakotonirainy, A., Schroeter, R., & Soro, A. (2014). Three social car visions to improve driver behaviour. *Pervasive and Mobile Computing, 14,* 147 – 160.

Rice, S., & McCarley, J. S. (2011). Effects of response bias and judgment framing on operator use of an automated aid in a target detection task. *Journal of Experimental Psychology: Applied, 17*(4), 320 – 331.

Ritchie, O. T., Watson, D. G., Griffiths, N., Misyak, J., Chater, N., Xu, Z., & Mouzakitis, A. (2019). How should autonomous vehicles overtake other drivers? *Transportation Research Part F: Traffic Psychology and Behaviour, 66,* 406 – 418.

Rovira, E., McLaughlin, A. C., Pak, R., & High, L. (2019). Looking for age differences in self-driving vehicles: Examining the effects of automation reliability, driving risk, and physical impairment on trust. *Frontiers in Psychology, 10* (article 800), 1 – 13.

Rovira, E., Pak, R., & McLaughlin, A. (2017). Effects of individual differences in working memory on performance and trust with various degrees of automation. *Theoretical Issues in Ergonomics Science, 18*(6), 573 – 591.

Rupp, M. A., Michaelis, J. R., McConnell, D. S., & Smither, J. A. (2016). The impact of technological trust and self-determined motivation on intentions to use wearable fitness technology. *Proceedings of the Human Factors and Ergonomics Society 2016 Annual Meeting,* 1434 – 1438.

SAE International (2018). J3016_201806. Taxonomy and definitions for terms relating to driving automation systems for on-road motor vehicles. https://www.sae.org/standards/content/j3016_201806/. Last accessed 18/08/2020.

Schaefer, K. E., Chen, J. Y. C., Szalma, J. L., & Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human Factors, 58*(3), 377 – 400.

Schoettle, B., & Sivak, M. (2014). A survey of public opinion about autonomous and self-driving vehicles in the U.S., the U.K., and Australia. *Report No. UMTRI – 2014 – 21. University of Michigan Transportation Research Institute.*

Schwark, J., Dolgov, I., Graves, W., & Hor, D. (2010). The influence of perceived task difficulty and importance on automation use. *Proceedings of the Human Factors and Ergonomics Society 54th Annual Meeting – 2010,* 1503 – 1507.

Siebert, F. W., Oehl, M., Bersch, F., & Pfister, H. (2017). The exact determination of subjective risk and comfort thresholds in car following. *Transportation Research Part F, 46,* 1 – 13.

Siebert, F. W., Oehl, M., & Pfister, H. (2014). The influence of time headway on subjective driver states in adaptive cruise control. *Transportation Research Part F, 25,* 65 – 73.

Skottke, E., Debus, G., Wang, L., & Huestegge, L. (2014). Carryover effects of highly automated convoy driving on subsequent manual driving performance. *Human Factors, 56*(7), 1272 – 1283.

Stewart, N. (2009). Decision by sampling: The role of the decision environment in risky choice. *Quarterly Journal of Experimental Psychology, 62,* 1041 – 1062.

Stewart, N., Chater, N., & Brown, G. D. A. (2006). Decision by sampling. *Cognitive Psychology, 53,* 1 – 26.

Taib-Maimon, M., & Shinar, D. (2001). Minimum and comfortable driving headways: Reality vs perception. *Human Factors, 43* (1), 159 – 172.

Tennant, C., Howard, S., Franks, B., Bauer, M. W., & Stares, S. (2016). *Autonomous vehicles negotiating a place on the road: A study on how drivers feel about interacting with autonomous vehicles on the road. Executive summary.* London School of Economics and Goodyear. https://www.lse.ac.uk/business-and-consultancy/consulting/assets/documents/autonomous-vehicles-executive-summary.pdf. Last Accessed 18/08/2020.

Tennant, C., Howard, S., Franks, B., Hall, M., Bauer, M. W., & Stares, S. (2015). *The ripple effect of drivers' behaviour on the road: A study on drivers' behaviour. Executive summary.* London School of Economics and Goodyear. https://www.lse.ac.uk/business-and-consultancy/consulting/assets/documents/the-ripple-effect-of-drivers-behaviour-on-the-road-exec-summary.pdf. Last accessed 18/08/2020.

Thaler, R., & Sunstein, C. (2008). *Nudge: Improving decisions about health, wealth, and happiness.* Yale University Press.

The Highway Code. URL: https://www.highwaycodeuk.co.uk/. Last accessed 18/08/2020.

The Motor Vehicles (Approval) Regulations (2001). Retrieved from: http://www.legislation.gov.uk/uksi/2001/25/made. Last accessed 18/08/2020.

Tipples, J. (2008). Negative emotionality influences the effects of emotion on time perception. *Emotion, 8*(1), 127 – 131.

TRL (2017). GATEway Project Report PPR807: *Driver responses to encountering automated vehicles in an urban environment.* URL: https://gateway-project.org.uk/wp-content/uploads/2017/02/D4.6_Driver-responses-to-encountering-automated-vehicles-in-an-urban-environment_PPR807.pdf. Last accessed 18/08/2020.

Tsetsos, K., Chater, N., & Usher, M. (2012). Salience driven value integration explains decision biases and preference reference. *PNAS, 109*(24), 9659 – 9664.

Tsetsos, K., Moran, R., Moreland, J., Chater, N., Usher, M., & Summerfield, C. (2016). Economic irrationality is optimal in noisy decision making. *PNAS, 113*(11), 3102 – 3107.

UK speed camera tolerances revealed: is your car's speedo accurate? Auto Express, 24/04/2019. https://www.autoexpress.co.uk/car-news/106674/uk-speed-camera-tolerances-revealed-is-your-cars-speedo-accurate. Last accessed 18/08/2020.

Usher, M., Tsetsos, K., Glickman, M., & Chater, N. (2019). Selective integration: An attentional theory of choice biases and adaptive choice. *Current Directions in Psychological Science, 28*(6), 552 – 559.

Venturer Trial 2: Interactions between autonomous vehicles and other vehicles on links and at junctions. Trial 2 findings. November 2017. URL: http://www.venturer-cars.com/trial-2-results/. Last accessed 18/08/2020.

Verame, J. K. M., Costanza, E., & Ramchurn, S. D. (2016). The effect of displaying system confidence information on the usage of autonomous systems for non-specialist applications: A lab study. *CHI 2016, Robot Personalities.* (pp. 4908 – 4920). May 7th – 12th, 2016. San Jose, CA, USA.

Verberne, F. M. F., Ham, J., & Midden, C. J. H. (2012). Trust in smart systems: Sharing driving goals and giving information to increase trustworthiness and acceptability of smart systems in cars. *Human Factors, 54* (5), 799 – 810.

Vlaev, I., Chater, N., Stewart, N., & Brown, G. D. A. (2011). Does the brain calculate value? *Trends in Cognitive Sciences, 15*(11), 546 – 554.

Vrkljan, B. H., & Anaby, D. (2011). What vehicle features are considered important when buying an automobile? An examination of driver preferences by age and gender. *Journal of Safety Research, 42*, 61 – 65.

Walasek, L., & Stewart, N. (2015). How to make loss aversion disappear and reverse: Tests of the decision by sampling origin of loss aversion. *Journal of Experimental Psychology: General, 144*(1), 7 – 11.

Walasek, L., & Stewart, N. (2019). Context-dependent sensitivity to losses: Range and skew manipulations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 45*(6), 957 – 968.

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology, 52*, 113 – 117.

Wiczorek, R., & Meyer, J. (2016). Asymmetric effects of false positive and false negative indications on the verification of alerts in different risk conditions. *Proceedings of the Human Factors and Ergonomics Society 2016 Annual Meeting*, 289 – 292.

Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., & Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *PNAS, 104*(19), 7780 – 7785.

Yantis, S., & Jonides, J. (1984). Abrupt visual onsets and selective attention: Evidence from visual search. *Journal of Experimental Psychology: Human Perception and Performance, 10*(5), 601 – 621.

Yantis, S., & Jonides, J. (1990). Abrupt visual onsets and selective attention: Voluntary versus automatic allocation. *Journal of Experimental Psychology: Human Perception and Performance, 16*(1), 121 – 134.

Yu, K., Berkovsky, S., Taib, R., Conway, D., Zhou, J., & Chen, F. (2017). User trust dynamics: An investigation driven by differences in system performance. *IUI 2017, Trust.* (pp. 307 – 317). March 13-16, 2017, Limassol, Cyprus.