TECHNISCHE UNIVERSITÄT MÜNCHEN

Lehrstuhl für Ergonomie

Einfluss von Expertise auf die Kritikalitätswahrnehmung in kritischen Fahrsituationen von Fahrerassistenzsystemen

Patrick Matthias Galaske

Vollständiger Abdruck der von der Fakultät für Maschinenwesen
der Technischen Universität München
zur Erlangung des akademischen Grades eines

Doktor-Ingenieurs (Dr.-Ing.)

genehmigten Dissertation.

Vorsitzender: Prof. Dr.-Ing. Georg Wachtmeister

Prüfer der Dissertation

1. Prof. Dr. phil. Klaus Bengler
2. Prof. Dr. phil. Martin Baumann

Die Dissertation wurde am 10.11.2017 bei der Technischen Universität München eingereicht und durch die Fakultät für Maschinenwesen am 10.04.2018 angenommen.

Danksagung

Die vorliegende Dissertation entstand während meiner Tätigkeit als Doktorand im Bereich Gebrauchs- und Funktionssicherheit bei der BMW AG in München.

Mein besonderer Dank für die großzügige Förderung und die andauernde Unterstützung meiner Dissertation gilt Herrn Prof. Dr. phil. Klaus Bengler, dem Leiter des Lehrstuhls für Ergonomie an der Technischen Universität München. Bei Herrn Prof. Dr. phil. Martin Baumann, dem Leiter der Abteilung Human Factors des Instituts für Psychologie und Pädagogik an der Universität Ulm, bedanke ich mich ganz herzlich für die Übernahme des Korreferates. Für die Übernahme des Vorsitzes der Prüfungskommission danke ich Herrn Prof. Dr.-Ing. Georg Wachtmeister, dem Leiter des Lehrstuhls für Verbrennungskraftmaschinen an der Technischen Universität München.

An dieser Stelle möchte ich mich bei Mehdi Farid und Dr.-Ing. Alexander Huesmann für die Gelegenheit zu dieser Arbeit, aber auch die Herausforderung und Anleitung während meiner Zeit bei der BMW AG bedanken. Für die Unterstützung bei der Durchführung der Fahrversuche sowie die kritische Durchsicht meiner Dissertation bedanke ich mich bei Peter Exner, Ralf Reisenauer und Veronika Weinbeer. Besonderer Dank gilt den vielen Studienteilnehmern, deren Teilnahme diese Arbeit ermöglicht hat. Außerdem danke ich allen beteiligten Mitarbeitern des Lehrstuhls für Ergonomie sowie der BMW AG für den fachlichen Austausch und die zahlreichen Hilfestellungen.

Die Entstehung der Dissertation wäre ohne Unterstützung meiner Familie nicht möglich gewesen. Ganz besonders möchte ich mich bei meiner Frau Dr.-Ing. Nadia Galaske für die endlose Geduld, andauernde Motivation und liebevolle Unterstützung bedanken.

Kurzfassung

Die Beherrschbarkeit von Fahrerassistenzsystemen (FAS) beschäftigt sich mit den sicherheitskritischen Aspekten der Interaktion von Fahrer und FAS. In diesem Themengebiet treffen Human Factors-Aspekte und technische Überlegungen auf komplexe Weise aufeinander. Durch die wachsende Anzahl und Komplexität der in modernen Kraftfahrzeugen verfügbaren Assistenzsysteme gewinnt dieses Feld stetig an Bedeutung. Es besteht ein erheblicher Bedarf an einer Reduktion des Aufwands bei der Beherrschbarkeitsbewertung von Fahrerassistenzsystemen, den bestehende Methoden nur begrenzt erfüllen können.

Durch den RESPONSE3 Code of Practice ist ein Rahmen für die Bestimmung der Beherrschbarkeit von FAS festgelegt worden. Dort wird neben aufwendigen Probandenstudien auch die Ermittlung der Beherrschbarkeit durch Expertengruppen verankert, nicht jedoch spezifiziert. In anderen Themengebieten wie etwa Nuklearsicherheit, Wirtschaft oder Medizin wurden bereits viele Erkenntnisse über die Eignung von Expertenbewertungen gewonnen, die jedoch ein gemischtes Bild darstellen. Eine Übertragung der Aussagen auf die Beherrschbarkeitsbewertung von FAS ist deswegen nicht zulässig.

Die vorliegende Dissertation beschäftigt sich mit dem Einfluss von Beherrschbarkeitsexpertise auf die Bewertung kritischer Fahrsituationen während der Nutzung von Fahrerassistenzsystemen. Dazu wird der relevante Stand der Technik zu den Themen Fahrerassistenzsysteme, der Beherrschbarkeitsbewertung solcher Systeme und Expertenbewertungen dargestellt und daraus wissenschaftliche Fragestellungen entwickelt. Anschließend werden vergleichende Studien zum Einfluss von Beherrschbarkeitsexpertise auf die Bewertung kritischer Fahrsituationen beschrieben und die Ergebnisse der Studien dargestellt. Der Effekt der Beherrschbarkeitsexpertise wird bezüglich objektiver und subjektiver Maße beschrieben und quantifiziert. Es werden einfache Modelle zur Beschreibung des Effekts präsentiert und konkrete Handlungsempfehlungen für die Durchführung von Expertenstudien gegeben. Weiterhin werden Aussagen zur Reliabilität von Expertenurteilen getroffen sowie inter-individuelle Unterschiede zwischen einzelnen Experten quantifiziert.

Insgesamt ergeben sich aus dieser Dissertation konkrete Hinweise für die frühzeitige Bewertung kritischer Fahrsituationen durch Experten bezüglich der Art der Aussagen, die gemacht werden können, der Methoden sowie der Validität der Ergebnisse.

Abstract

The controllability of advanced driver assistance systems (ADAS) is focused on the safety relevant aspects of the interaction of driver and ADAS. In this area, human factors and technical aspects interact in a complex manner. Due to the increasing number and complexity of available ADAS in modern passenger vehicles, this field is continuously growing in relevance. There is relevant demand for a reduction in the effort required for the assessment of controllability of advanced driver assistance systems, which can't be accomplished fully through existing methods.

The RESPONSE 3 Code of Practice established a framework for the assessment of controllability of ADAS. Besides naïve test subject studies it also formulated the assessment of controllability through expert studies but did not specify them. In other areas such as nuclear safety, economy or medicine studies already allowed much insight into the suitability of expert studies, but a mixed picture remains. Therefor it is not possible to carry over those assertions to the assessment of controllability of ADAS.

This thesis studies the influence of expertise of controllability on the assessment of critical driving situations during the usage of advanced driver assistance systems. For this purpose, the relevant state of the art on the topics of driver assistance systems, controllability assessment and expert studies is summarized and from this the scientific goals are developed. Afterwards comparative studies on the influence of expertise of controllability on the assessment of critical driving situations and the results of these studies are described. The effect of expertise of controllability is quantified with respect to subjective and objective measures. Simple models for the effects are presented and recommendations for the design of expert studies are given. Furthermore, the reliability of expert judgments is discussed and inter-individual differences between experts are quantified.

Overall this thesis gives concrete indications on the early assessment of critical driving situations using experts regarding the type of assertions that can be made, the methods that should be used and the validity of the results.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

ACEA	Association des Constructeurs Européens d'Automobiles
AG	Aktiengesellschaft
ASIL	Automotive Safety Integration Level
AUROC	Area Under Receiver Operator Curve
BASt	Bundesanstalt für Straßenbau
BMW	Bayerische Motorenwerke
B(k | p, n)	Binomialverteilung: k Erfolge bei der Wahrscheinlichkeit p nach n Versuchen
C	Controllability (dt. Beherrschbarkeit)
CoP	Code of Practice (Response 3)
E	Exposure (dt. Auftretenswahrscheinlichkeit)
ESP	Elektronisches Stabilitätsprogramm
FAS	Fahrerassistenzsystem
FIDE	Fédération Internationale des Échecs (dt. Internationaler Schachverband)
FIZ	Forschungs- und Innovationszentrum
ISO	International Organisation for Standardization
HAF	Hochautomatisiertes Fahren
MMPI	Minnesota Multiphasic Personality Index
MSE	Mittlerer quadratischer Fehler (engl. „Mean Square Error“)
LKW	Lastkraftwagen
LR	Likelihood-Ratio

NB	Neubewertung
NDM	Naturalistic Decision Making
NHTSA	National Highway Traffic Safety Association
PKW	Personenkraftwagen
PS	Personenschäden
QM	Qualitätsmaßnahme
RPD	Recognition Primed Decision
ROC	Receiver Operator Curve
S	Severity (dt. Schadensschwere)
SAE	Society of Automotive Engineers
SBS	Störungsbewertungsskala
SET	Signalentdeckungstheorie
SuRT	Surrogate Reference Task
TAF	Teilautomatisiertes Fahren
THERP	Technique for Human Error-Rate Prediction
V&H	(kognitive) Verzerrungen und Heuristiken

1 Einleitung

„There are only two kinds of people who are really fascinating; people who know absolutely everything, and people who know absolutely nothing." Oscar Fingal O'Flahertie Wills Wilde

1.1 Motivation

Die Anzahl und Komplexität von Fahrerassistenzsystemen (FAS) in modernen Kraftfahrzeugen hat sich in den letzten zwei Dekaden deutlich erhöht. Moderne Fahrerassistenzsysteme beschränken sich nicht nur darauf, den Fahrer auf potentielle Gefahren in bestimmten Situationen hinzuweisen oder davor zu warnen, sondern unterstützen oder übernehmen Teile der Fahrzeugkontrolle auch für längere Zeiträume. Beispiele für solche Systeme, die bereits im Markt etabliert sind, sind Abstandstempomaten, eingreifende Spurhaltesysteme, selbstauslösende Notbremssysteme oder Parkmanöverassistenten. Es ist davon auszugehen, dass die Anzahl und Komplexität solcher Systeme in Zukunft sogar noch weiter wächst (Bengler et al., 2012; Continental AG, 2015).

Ein Aspekt der Entwicklung solcher eingreifenden Fahrerassistenzsysteme ist die Beherrschbarkeit dieser Systeme an den Systemgrenzen oder im Fehlerfall. Durch die wachsende Anzahl und die steigende Komplexität der agierenden Assistenzsysteme wird der Aufwand, die Fähigkeit des Fahrers, Schaden von sich oder seiner Umwelt abzuwenden – der Beherrschbarkeit – jedoch weiter erhöhen. Dieser Anstieg des Aufwands macht methodische Verbesserungen notwendig, um die Entwicklung zukünftiger Fahrerassistenzsysteme nicht einzuschränken. Probabilistische Methoden der Beherrschbarkeitsbewertung, die das Abwenden einer kritischen Fahrsituation durch den Fahrer als Verkettung stochastischer Zufallsereignisse modellieren, werden in Zukunft eine wichtige Rolle spielen, benötigen jedoch einen erheblichen a-priori Aufwand in der Entwicklung und Validierung dieser Methoden. Deswegen stellen sie auf mittelfristige Sicht noch keine geeignete Möglichkeit dar.

Expertenbewertungen sind eine mögliche Methode zur Beurteilung der Kritikalität von Fahrerassistenzsystemen. Diese Methode wurde bereits in anderen Domänen angewendet und kann auch auf diesen Bereich übertragen werden. Allerdings sind aus anderen

Domänen Einschränkungen der Methode Expertenbewertung bekannt, die eine erfolgreiche Anwendung für sicherheitskritische Studien verhindern können. Unterschiedliche Fachgebiete kommen dabei zu verschiedenen Aussagen zu der Eignung von Expertenstudien, sodass eine einfache Übertragung auf den Bereich Expertenbewertung der Beherrschbarkeit von Fahrerassistenzsystemen nicht sinnvoll erscheint.

1.2 Zielsetzung

Die Bewertung der Beherrschbarkeit kritischer Fahrsituationen, die bei der Verwendung von Fahrerassistenzsystemen entstehen können, stellt eine notwendige Komponente der Entwicklung dieser Systeme dar. Um eine Effizienzsteigerung der Entwicklung zu erzielen, wäre es wünschenswert, bereits frühzeitig im Entwicklungsprozess von Fahrerassistenzsystemen einen Indikator für die Eignung verschiedener Systemauslegungen zu haben, sodass eventuelle Defizite bei der Beherrschbarkeit in bestimmten Situationen früh adressiert werden können. Da zu diesem Zeitpunkt das Systemdesign Gegenstand häufiger Überarbeitung ist, besteht eine Forderung nach einer Methode mit einer geringen Iterationszeit, die auf verschiedenartige Fragestellungen wie Quer- und Längsdynamik in verschiedenen Umgebungen, aber auch Park- und Rangierszenarien, angewendet werden kann.

Die Expertenbewertung ist eine Methode, um Lösungen für eine große Bandbreite von Problemen zu erhalten. Jedoch wurden in verschiedenen Domänen Hinweise dafür gefunden, dass die Genauigkeit und Validität von Expertenbewertungen systematisch eingeschränkt ist.

Diese Dissertation beschäftigt sich mit der Prädiktion der subjektiven Störungsbewertung naiver Probenden im Fahrversuch mittels der subjektiven Urteile von Experten. Das Ziel besteht darin, Rahmenbedingungen zu identifizieren, unter denen eine quantifizierbare Aussage über die Genauigkeit der Aussagen gemacht werden kann, die durch in der Praxis durchführbare Expertenbewertungen gewonnen werden können.

Schließlich sollen die in dieser Dissertation geschilderten Erkenntnisse dazu beitragen, die frühen Phasen des Entwicklungsprozesses von Fahrerassistenzsystemen durch eine valide Methode zur frühzeitigen Identifikation von möglicherweise nicht-beherrschbaren Fahrsituationen zu unterstützen.

1.3 Aufbau der Dissertation

Diese Dissertation gliedert sich in insgesamt sieben Abschnitte. Nach der Einleitung werden zunächst Grundlagen aus dem Bereich Fahrerassistenzsysteme und Expertiseforschung zusammengefasst. Aus diesen wird im dritten Kapitel die konkrete Aufgabenstellung abgeleitet. Diesem Teil folgt die Beschreibung der empirischen Untersuchungen, die im Rahmen dieser Dissertation durchgeführt worden sind. Darin enthalten sind ein Vorversuch sowie die Beschreibungen der Zielsetzungen, Methoden und Ergebnisse von vier Einzelstudien.

Abbildung 1 Graphische Darstellung des Aufbaus der Dissertation.

Die Ergebnisse der Einzelstudien werden im gleichen Kapitel gesamthaft betrachtet und untersucht. Im fünften Kapitel werden die zuvor erarbeiteten Ergebnisse diskutiert und in dem Kontext der Zielsetzung der Dissertation betrachtet. Schließlich wird ein Ausblick auf mögliche weitere Forschungsfragen gegeben und eine Zusammenfassung der Dissertation dargestellt.

2 Grundlagen

Dieses Kapitel stellt grundlegende Zusammenhänge im Bereich Fahrerassistenzsysteme und der Beherrschbarkeitsbewertung dieser Systeme dar. Es führt von der Motivation solcher Systeme und ihrer Gestaltung in aktuellen Personenkraftwagen (PKW) zu den Randbedingungen für ihre Entwicklung. Es wird insbesondere auf die sogenannte Beherrschbarkeit von Fahrerassistenzsystemen eingegangen und die Bewertung der Beherrschbarkeit im Detail diskutiert. Schließlich wird die Expertenbewertung der Beherrschbarkeit von Fahrerassistenzsystemen erläutert und genauer auf den Expertenbegriff bzw. Expertise eingegangen sowie die Vorgehensweise der Expertenbefragung über verschiedene Fachbereiche hinweg verglichen. Auch Grundlagen zur probabilistische Risikoanalyse und der Signalentdeckungstheorie sowie Spieltheorie werden dargestellt.

2.1 Fahrerassistenzsysteme

Fahrerassistenzsysteme unterstützen den Fahrer bei der Bewältigung der Fahraufgabe auf verschiedene Arten. Moderne PKW verfügen schon heute über Systeme, die den Fahrer beispielsweise bei der Navigation unterstützen oder ihn vor nahenden Gefahren etwa durch Kreuzungsverkehr warnen. In vielen Segmenten gibt es bereits Fahrzeuge, die selbstständig Notbremsungen durchführen können, um Unfälle mit vorausfahrenden PKW oder Fußgängern zu vermeiden oder zumindest die Unfallfolgen zu verringern. Verfügbare Komfortsysteme unterstützen den Fahrer nicht nur bei der Längsführung des Fahrzeugs, sondern auch bei der Querführung. Im Bereich niedriger Geschwindigkeiten gibt es im Markt bereits Parksysteme, die die gesamte Längs- und Querführung des Fahrzeugs übernehmen können, jedoch weiterhin auf die Überwachung durch den Fahrer angewiesen sind.

Alle diese Systeme unterstützen den Fahrer bei einem oder mehreren Aspekten der Fahraufgabe. Diese lässt sich nach Donges (1982) in die drei Ebenen Navigation, Bahnführung und Stabilisierung aufteilen. Die Navigationsaufgabe besteht in der Auswahl einer geeigneten Verbindung zwischen dem Start und dem Ziel der Fahrt unter Berücksichtigung der Randbedingungen, wie etwa Fahrzeit und Verkehrsdichte. Unter Bahnführung wird die Auswahl einer möglichen Trajektorie über die geplante Route verstanden, die

die Verkehrsregeln und andere Verkehrsteilnehmer berücksichtigt. Die Stabilisierungsaufgabe schließlich besteht in der Führung des Fahrzeugs entlang der ausgesuchten Trajektorie. Auf der Navigations- und Stabilisierungsebene sind Assistenzsysteme sehr gut etabliert. Eine große Anzahl an PKW verfügt bereits über Navigationssysteme und das elektronische Stabilitätsprogramm (ESP) ist für die Zulassung von PKW innerhalb der europäischen Union vorgeschrieben (Europäisches Parlament und Rat, 2009). Erst moderne Fahrerassistenzsysteme sind jedoch auch in der Lage, den Fahrer auf Bahnführungsebene zu unterstützen.

Alle Fahrerassistenzsysteme lassen sich zusätzlich nach ihrer Wirkweise unterteilen (Gasser, Seeck & Smith, 2015). So gibt es informierende oder warnende Systeme, die nicht selbst in die Fahrdynamik eingreifen, aber den Fahrer eine Handlung vorschlagen oder ihn auf mögliche Gefahren aufmerksam machen. Notfallfunktionen greifen in die Fahrdynamik ein, beschränken sich jedoch auf solche Fälle, in denen der Eintritt eines Unfalls absehbar ist und eine konkrete Gegenmaßnahme ergriffen werden muss, um Schaden zu vermeiden oder zu vermindern. Solche Systeme sind beispielsweise Notbrems- oder Ausweichsysteme sowie Nothaltesysteme, die das Fahrzeug zum Stehen bringen können, wenn der Fahrer handlungsunfähig ist. In der dritten Kategorie sind Systeme, die kontinuierlich in der Fahrdynamik eingreifen und einen Teil der Fahraufgabe übernehmen. Sie sind dadurch gekennzeichnet, dass die Übergabe des betroffenen Aspekts der Fahraufgabe bewusst geschieht. Zumeist sind dies Komfortfunktionen wie Tempomatfunktionen oder Spurhaltesysteme.

Weiterhin ist es möglich, kontinuierlich eingreifende Fahrerassistenzsysteme nach der Stärke der Automatisierung zu kategorisieren. Die Nomenklatur der Bundesanstalt für Straßenwesen (BASt) kennt vier Automatisierungsstufen (Gasser, Arzt, Ayoubi & Bartels, 2012). Stufe 0 bezeichnet rein manuelles Fahren, bei dem der Fahrer sowohl die Längs- als auch Querführung übernimmt. Stufe 1 wird als assistiertes Fahren bezeichnet, bei dem das Fahrzeug den Fahrer bei der Längs- oder Querführung innerhalb seiner definierten Systemgrenzen unterstützt. Die restliche Fahraufgabe wird vom Fahrer übernommen. Stufe 2 bezeichnet das teilautomatisierte Fahren, bei dem das Fahrzeug sowohl die Längs- als auch Querführung übernimmt. Der Fahrer muss das System jedoch weiterhin überwachen und jederzeit die Fahraufgabe wieder übernehmen können. Stufe 3 wird als hochautomatisiertes Fahren bezeichnet, bei dem im Unterschied zu Stufe 2 das System nicht mehr dauerhaft durch den Fahrer überwacht werden muss. Er muss jedoch die Fahraufgabe innerhalb eines gewissen Zeitraums übernehmen können, wenn ihn das System dazu auffordert. Die vierte Stufe ist das vollautomatisierte Fahren, bei dem das

Fahrzeug die Längs- und Querführung während seiner Aktivierung vollständig übernimmt. Das bedeutet, dass auch keine Fahrerübernahmen während der Benutzung des Systems mehr notwendig sind.

Tabelle 1 enthält die Automatisierungsstufen nach der BASt Nomenklatur sowie die entsprechenden Automatisierungsstufen nach der Definition der SAE (SAE J3016). Die Stufen 0 bis 2 der BASt finden sich bei der SAE entsprechend wieder. Stufe 3 nach der Definition der SAE ist bedingte Automatisierung (engl. „Conditional Automation"), die Systeme beschreibt, die den Fall, dass der Fahrer einer Übernahmeaufforderung nicht nachkommt, nicht absichern. Dies ist nach der Definition der SAE erst ab Stufe 4 der Fall. Stufe 5 nach der Definition der SAE entspricht dann wieder der Vollautomatisierung aus der Definition der BASt. Eine weitere Definition der Automatisierungsstufen ist in (NHTSA, 2013) verfügbar. Auch hier stimmen die Stufen Null bis Zwei mit der Definition der BASt überein. Die dritte Automatisierungsstufe nach der Definition der NHTSA entspricht der Stufe 3 nach Definition der BASt, jedoch wird hier nur von einer Übernahme des Fahrers innerhalb eines ausreichend komfortablen Zeitraums gesprochen. Automatisierungsstufe 4 nach der Definition der NHTSA entspricht wieder der Vollautomatisierung nach der Definition der BASt.

Tabelle 1 Vergleich der Automationsstufen nach BASt (Gasser et al., 2012), SAE (SAE J3016) und NHTSA (NHTSA, 2013).

Stufe	BASt Nomenklatur	SAE	NHTSA
0	Fahrer only	No Automation	0
1	Assistiert	Assisted	1
2	Teilautomatisiert	Partial Automation	2
3	Hochautomatisiert	Conditional Automation	3
4	Vollautomatisiert	High Automation	3/4
5		Full Automation	

Zusätzlich lassen sich Fahrerassistenzsysteme auch nach den anwendbaren Geschwindigkeitsbereichen (z.B. Rangierbereich, Staubereich, Autobahngeschwindigkeiten), den unterstützten Fahrzeugtypen (z.B. PKW, Anhänger, LKW), verfügbaren Straßentypen und mehreren anderen Faktoren unterteilen.

2.2 Beherrschbarkeit

Fahrerassistenzsysteme, die in die Fahrzeugführung eingreifen können, bringen immer auch Risiken mit sich. So können Systeme, die die Längsdynamik des Fahrzeugs beeinflussen, im Fehlerfall auch zum ungewollten Selbstbeschleunigen des Fahrzeugs oder einer ungerechtfertigten Abbremsung des Fahrzeugs führen. Systeme, die die Querdynamik beeinflussen, können durch fehlerhafte Eingriffe ein Verlassen des Fahrstreifens verursachen. Selbst informierende Systeme, die fehlerhafte Hinweise an den Fahrer geben, können ihn oder andere Verkehrsteilnehmer in eine gefährliche Situation bringen. Die Verhinderung solcher Fehlfunktionen ist das Ziel der funktionalen Sicherheit von Fahrerassistenzsystemen.

Auch in dem Fall, dass diese Systeme keine Fehlfunktion haben und innerhalb ihrer Systemdefinition agieren, können dennoch Gefährdungen der Fahrzeuginsassen oder anderer Verkehrsteilnehmer eintreten. So können beispielsweise Fehler bei der Mode Awareness des Fahrers dazu führen, dass der Fahrer in einer kritischen Fahrsituation erst später eingreift, als er es andernfalls getan hätte. Die Gebrauchssicherheit beschäftigt sich mit der Bewertung solcher Risiken, die auch beim bestimmungsgemäßen Gebrauch von Fahrerassistenzsystemen auftreten können.

Sowohl bei der Gebrauchs- als auch funktionalen Sicherheit spielen drei Faktoren eine wichtige Rolle für die Bewertung von Risiken. Um das Risiko des Eintritts eines Schadens in einer Fahrsituation zu bestimmen, muss zunächst bekannt sein, wie häufig die potentiell gefährliche Ausgangssituation auftritt. Dieser Faktor wird als Auftretenswahrscheinlichkeit (engl. „exposure“) bezeichnet und mit dem Buchstaben E abgekürzt. Ein weiterer Faktor ist das Ausmaß der potentiellen Schädigung für die Insassen des Ego-Fahrzeugs oder anderer Verkehrsteilnehmer. Dieses wird als Schadensschwere (engl. „severity“) bezeichnet und mit dem Buchstaben S abgekürzt. Schließlich steht zwischen der potentiell gefährlichen Ausgangssituation und dem Eintritt der prognostizierten Schädigung eine Reihe von notwendigen Handlungen der beteiligten Verkehrsteilnehmer. Die Wahrscheinlichkeit der Verhinderung des Eintritts der Schädigung durch die relevanten Verkehrsteilnehmer wird als Beherrschbarkeit bezeichnet. Dieser Faktor wird englisch Controllability genannt und mit dem Buchstaben C abgekürzt (ISO 26262-3). Er stellt ebenfalls einen wesentlichen Teil der Entwicklung neuer Fahrerassistenzsysteme dar (Neukum & Reinelt, 2005).

Die Bewertung einer potentiell gefährlichen Fahrsituation erfolgt, indem jeder der Faktoren Exposure (E), Severity (S) und Controllability (C) für diese Situation klassifiziert wird. Die möglichen Klassen können Tabelle 2 entnommen werden.

Tabelle 2 Definitionen der Klassen für Unfallschwere (a), Häufigkeit (b) und Beherrschbarkeit (c) (ISO 26262-3).

Klasse	Beschreibung
S0	Keine Verletzungen.
S1	Leichte bis mittelschwere Verletzungen. Geringe Sterbewahrscheinlichkeit.
S2	Schwere Verletzungen. Todesfolge unwahrscheinlich.
S3	Lebensgefährliche Verletzungen mit möglicher Todesfolge.

Klasse	Beschreibung
E0	Unglaubwürdig.
E1	Sehr niedrige Häufigkeit. Seltener als einmal im Jahr.
E2	Niedrige Häufigkeit. Mehrmals jährlich.
E3	Mittlere Häufigkeit. Einmal monatlich.
E4	Große Häufigkeit. Bei fast jeder Fahrt.

Klasse	Beschreibung
C0	Im Allgemeinen Beherrschbar.
C1	Einfach Beherrschbar. Eintritt des Schadens in weniger als 1% der Fälle.
C2	Schwierige Beherrschbarkeit. Schadenseintritt in weniger als 10% der Fälle.
C3	Sehr schwierig Beherrschbar. Schadenseintritt häufiger als in 10% der Fälle.

Das Gesamtrisiko der zu bewertenden Situation ergibt sich aus der Kombination der Werte der drei Variablen E, S und C. Für bestimmte Risikoniveaus legt die ISO 26262 Maßnahmen fest, die mindestens getroffen werden müssen, um das Gesamtrisiko zu verringern. Diese reichen von einer einfachen Qualitätsmaßnahme (QM) bis zu den ASIL (Automotive Safety Integration Level) A bis D, wobei A die geringste Stufe und D die höchste Stufe darstellt. Das jeweils notwendige Niveau der zu treffenden Maßnahmen kann Tabelle 3 entnommen werden.

Tabelle 3 ASIL Bestimmungsmatrix (Wilhelm, Ebel & Weitzel, 2015).

		Beherrschbarkeit („controllability“)		
Schadensschwere („severity“)	Häufigkeit („exposure“)	C1 (einfach)	C2 (normal)	C3 (schwierig)
S1 (leicht/mittel)	E1 (sehr niedrig)	QM	QM	QM
	E2 (niedrig)	QM	QM	QM
	E3 (mittel)	QM	QM	A
	E4 (hoch)	QM	A	B
S2 (schwer)	E1 (sehr niedrig)	QM	QM	QM
	E2 (niedrig)	QM	QM	A
	E3 (mittel)	QM	B	B
	E4 (hoch)	A	B	C
S3 (lebensgefährlich)	E1 (sehr niedrig)	QM	QM	A
	E2 (niedrig)	QM	A	B
	E3 (mittel)	A	B	C
	E4 (hoch)	B	C	D

Die Bewertung der Beherrschbarkeit ist also eine der drei Komponenten für die Feststellung der funktionalen Sicherheit von Fahrerassistenzsystemen und damit für den Nachweis der Sicherheit der Systeme notwendig. Da eingreifende FAS immer weitere Verbreitung finden und moderne FAS die Fahrdynamik von PKW entscheidend beeinflussen können, gewinnt die Bewertung der Beherrschbarkeit immer weiter an Bedeutung (Continental AG, 2015). Dies gilt sowohl im Falle von Fehlfunktionen der Systeme als auch an deren Systemgrenzen. Die voranschreitende Entwicklung von FAS und die steigende Verbreitung von teilautomatisiertem Fahren werden diesen Bedeutungsgewinn der Beherrschbarkeitsbewertung weiter antreiben. Auch beim hochautomatisierten Fahren wird es weiterhin Unfallszenarien geben, in denen die Beherrschbarkeit durch den übernehmenden Fahrer oder den betroffenen Fremdverkehr ermittelt werden muss. Die folgenden Abschnitte beschäftigen sich mit der Methode der Bewertung der Beherrschbarkeit im Entwicklungsprozess von FAS.

2.3 Code of Practice

Der Code of Practice (CoP) für die Evaluation von Fahrerassistenzsystemen ist aus dem Forschungsprojekt PReVENT hervorgegangen und enthält Empfehlungen für die Bewertung der Beherrschbarkeit von Fahrerassistenzsystemen (PReVENT, 2009). Durch die Billigung des Code of Practice durch die ACEA (Association des Constructeurs Européens d’Automobiles) hat dieses Dokument den Status eines Quasi-Standards erhalten

(ACEA, 2009). Da kein anderes vergleichbares Dokument existiert, stellt es somit eine wichtige Grundlage für die Beherrschbarkeitsbewertung dar.

Der Code of Practice unterteilt sich in zwei relevante Segmente. Erstens definiert er einen Prozess zur Identifikation von Risiken, die mit dem Gebrauch oder einer Fehlfunktion eines Fahrerassistenzsystems verbunden sein können. Dieser Teil wurde mit der ISO 26262 teilweise überarbeitet und steht damit in einem aktuelleren Dokument zur Verfügung. Zweitens gibt er Empfehlungen zur Methode der Beherrschbarkeitsbewertung, die in keinem anderen gleichrangigen Dokument existieren. Dabei verwendet der CoP bereits die gleichen Beherrschbarkeitslevel (C0 bis C3), wie sie später in die ISO 26262 eingeflossen sind.

Zusätzlich zu der Definition der Beherrschbarkeit als statistische Wahrscheinlichkeit Schäden oder Verletzungen zu vermeiden stellt der CoP Beherrschbarkeit auch als Ergebnis der Wahrnehmbarkeit (perceptability) der kritischen Situation, der Fähigkeit des Fahrers zur Entscheidung zu einer effektiven Gegenmaßnahme (driver decision) sowie des Vermögens des Fahrers, diese Maßnahme erfolgreich umzusetzen (driver action) dar. Zu der Wahrnehmbarkeit der Kritikalität der zu prüfenden Fahrsituation zählen laut CoP unter anderem die Vorhersagbarkeit der Grundsituation, das Vertrauen des Fahrers in das Fahrerassistenzsystem, die zu erwartende Vigilanz der jeweiligen Verkehrsteilnehmer und die physische sowie physiologische Möglichkeit zur Wahrnehmung der relevanten Informationen. Zum Entscheidungsverhalten der Fahrer tragen die Verständlichkeit des bedienten Systems, die Möglichkeit zum Erlernen des Systemverhaltens sowie gegebenenfalls Verhaltensanpassungen des Fahrers an das Fahrerassistenzsystem bei. Die Fähigkeit, die gewählte Gegenmaßnahme durchzuführen, wird beispielsweise beeinflusst durch intuitives Systemdesign, das zu erwartende Niveau an Erfahrung des Fahrers mit dem System sowie die gegebenenfalls notwendige Fahrfertigkeit zum Durchführen der Gegenmaßnahme. Am Beispiel einer Auffahrwarnung während des Betriebs eines Abstandstempomaten muss der Fahrer einen Warnton, ein Warnsymbol in den Instrumenten sowie das bremsende Vorderfahrzeug wahrnehmen. Anschließend muss er sich zu einer Maßnahme entscheiden, um die kritische Situation zu lösen. Eine mögliche Maßnahme ist Bremsen. Schließlich muss der Fahrer diese Maßnahme erfolgreich durchführen, das heißt die Bremse erfolgreich und in ausreichendem Maße betätigen. Jede dieser drei Schritte wird von einem aufmerksamen und fahrtüchtigen Fahrer vermutlich mit Erfolg abgeschlossen, was die gute Beherrschbarkeit dieser Situation begründet.

Nach dem Code of Practice gibt es drei Methoden, um die Beherrschbarkeit einer Fahrsituation zu bewerten. In einfachen Fällen kann die Beherrschbarkeit durch eine Rationalisierung bewertet werden. Dieser Fall tritt etwa dann ein, wenn durch eine Argumentation klar die Aussage gemacht werden kann, dass eine Situation sehr einfach oder sehr schwer beherrschbar ist. Beispielsweise führen fehlerhafte Anzeigen nicht sicherheitskritischer Bedienelemente in der Regel nicht zu Personenschäden, da die Situation für den Fahrer im Allgemeinen als beherrschbar gilt. Starke, automatische Bremseingriffe wiederum gelten als wenig beherrschbar, da ein bewusstes und zeitnahes Übersteuern, beispielsweise durch die Betätigung des Gaspedals, unter Einwirkung einer starken Verzögerung ohne vorherige Übung oder erweiterte Fahrausbildung schwierig ist.

Für Fragestellungen, bei denen es nicht möglich ist, eine zwingende Argumentation für ein bestimmtes Beherrschbarkeitsniveau zu finden, bietet der CoP Probandenstudien entweder im Fahrsimulator oder im Realfahrzeug an. Fahrsimulatorstudien werden für die frühen Entwicklungsphasen empfohlen, Realfahrzeugversuche auf der Teststrecke oder im Realverkehr lediglich für spätere Phasen. Der CoP hält dabei fest, dass für eine Probandenstudie naive Teilnehmer aus dem Kollektiv der späteren Fahrer ausgewählt werden müssen. Mit naiv wird dabei gemeint, dass die Probanden über keine überdurchschnittlichen Kenntnisse über Fahrerassistenzsysteme im Allgemeinen, das spezifisch zu bewertende Fahrerassistenzsystem oder über ein erweitertes Fahrertraining verfügen dürfen. Der CoP verlangt weiterhin, dass für alle zu prüfenden Fahrsituationen objektiv messbare Kriterien für die Nicht-Beherrschbarkeit aufgestellt werden, sodass für jeden Probanden und jeden Durchlauf der Testsituation eine binäre Entscheidung getroffen werden kann, ob das Szenario bestanden oder nicht bestanden wurde. Bei der Bewertung der Beherrschbarkeit eines fehlerhaften Lenkmoments können beispielsweise das Überschreiten der Markierungen des Fahrstreifens des Egofahrzeugs ohne vorherige Absicherung durch den Fahrer oder eine Kollision mit einer Fahrstreifenbegrenzung Nicht-Beherrschbarkeitskriterien darstellen (Neukum, 2010).

Für eine Studie zur Bestätigung der C2-Beherrschbarkeit einer Fahrsituation, die im CoP als Beherrschbarkeit in 85% der Fälle definiert wird, werden mindestens 20 Probanden empfohlen. Für eine Bestätigung der Beherrschbarkeit ist es dabei notwendig, dass alle 20 Teilnehmer die betreffende Fahrsituation bestehen, also kein einziger Proband irgendeins der Nicht-Beherrschbarkeitskriterien erfüllt.

Das Kriterium, dass 20 von 20 Teilnehmern die Fahrsituation bestehen müssen, entsteht aus der Zielsetzung, dass nach einer bestandenen Beherrschbarkeitsstudie zum 5% Signifikanzniveau ausgeschlossen werden kann, dass das System in Wirklichkeit unbeherrschbar ist. Dies entspricht einem Bernoulli-Experiment mit 20 Wiederholungen und der Wahrscheinlichkeit p = 100% – 85% = 0,15.

$$p_{FalscheAnnahme,C2} = B(0|0{,}15; 20) = 0{,}0388 = 3{,}88\% < 5\%$$

Wird das gleiche Kriterium auf die C1-Beherrschbarkeit angewendet, bei der das Nicht-Beherrschbarkeitskriterium nur in 1% der Fälle ausgelöst werden darf, so führt dies zu einer minimalen Anzahl an Probanden von 299.

$$p_{FalscheAnnahme,C1} = B(0|0{,}01; 299) = 0{,}0495 = 4{,}95\% < 5\%$$

Damit wird klar, dass es nicht sinnvoll ist, die C1-Beherrschbarkeit durch Probandenstudien experimentell zu ermitteln. Selbst wenn es logistisch möglich wäre, eine so große Anzahl an naiven Teilnehmern an einer Studie teilnehmen zu lassen, könnten Einflüsse wie missverstandene Instruktionen oder Übermüdung eines Probanden zu einem Verletzen der Nicht-Beherrschbarkeitskriterien bei zumindest einem der Teilnehmer führen, sodass der experimentelle Nachweis der C1-Beherrschbarkeit schwierig erscheint (Weitzel & Winner, 2012).

Bei der Betrachtung von Probandenstudien für die Beherrschbarkeitsbewertung ist zu beachten, dass durch die wachsende Komplexität der Fahrerassistenzsysteme auch die Anzahl der beherrschbarkeitsrelevanten Fahrsituationen wächst. Dies führt zu einem entsprechenden Wachstum beim Aufwand der Beherrschbarkeitsbewertung. Weiterhin muss bei diesen Studien die Validität von Realfahrzeugstudien gegen das Sicherheitsrisiko von Versuchen im Realfahrzeug mit naiven Probanden abgewogen werden. Da das Ziel der Studie in der Bewertung der Beherrschbarkeit besteht, kann a-priori nicht ausgeschlossen werden, dass es tatsächlich zu einer Verletzung der Nicht-Beherrschbarkeitskriterien kommt, was in bestimmten Fahrsituationen zu einer unzulässigen Gefährdung der Studienteilnehmer führen kann. Die alternative Durchführung der Studien im Fahrsimulator führt wiederum zu Einschränkungen bei der externen Validität der Ergebnisse.

Die dritte im Code of Practice genannte Methode zur Bewertung der Beherrschbarkeit von Fahrerassistenzsystemen ist die Expertenbewertung. Der CoP definiert, dass eine Expertenrunde für die Beherrschbarkeitsbewertung sowohl aus Funktionsentwicklern als auch zusätzlichen Spezialisten besteht. Es wird explizit empfohlen, Mitarbeiter aus

anderen Abteilungen als der eigentlichen Fahrerassistenzsystementwicklung einzuschließen, um eine unabhängige Sichtweise auf die Fragestellung zu erhalten. Es wird vorausgesetzt, dass die Experten das zu bewertende System und sein Verhalten in den relevanten Situationen kennen. Sollte es Zweifel an der eigenen Fähigkeit zu einem Urteil geben, empfiehlt der CoP zusätzliche Informationen zu beschaffen oder eine Probandenstudie durchzuführen (PReVENT, 2009).

Insgesamt gibt der Code of Practice für die Evaluation von Fahrerassistenzsystemen nur wenige Richtlinien für die Durchführung von Expertenbewertungen zur Bewertung der Beherrschbarkeit von Fahrerassistenzsystemen. Die nächsten Abschnitte beschäftigen sich deswegen mit Experten, ihrem Einsatz in Expertenbewertungen und den Eigenschaften solcher Expertenstudien.

2.4 Der Expertenbegriff

Es gibt in der Literatur keine allgemein akzeptierte Definition des Begriffs Experte. Insbesondere durch die häufige Verwendung der Bezeichnung Experte in populären Medien kommt es häufig zu in Konflikt stehenden Interpretationen dieser Bezeichnung. Im Folgenden sollen bestehende Definitionen diskutiert und anschließend der Expertenbegriff für den Rahmen dieser Dissertation festgelegt werden.

Meyer und Booker (1990) beschreiben Experten als Personen, die die gewünschte Tiefe an fachlichem Hintergrundwissen in dem zu untersuchenden Fachbereich haben und die durch ihre Fachkollegen oder Studiendurchführende als ausreichend qualifiziert angesehen werden, um die Fragestellung zu beantworten.

O'Hagan et al. (2006) nennen zwei mögliche Definitionen. Zum einen könne der Begriff Experte lediglich die Person bezeichnen, deren Urteile zu einer Fragestellung entnommen werden sollen. Zum anderen könne der Begriff aber auch bedeuten, dass die Person großes Wissen über den Untersuchungsgegenstand hat, wobei für den Expertenstatus auch die Wissensorganisation und die Art der Nutzung dieses Wissens relevant sind.

Ferrell (1994) beschrieb Experten als Personen mit substantiellem Wissen über die Ereignisse, deren Unsicherheit adressiert werden soll.

In der Domäne Schach werden Experten anhand verschiedener Merkmale identifiziert. So spielen die Funktion und der Inhalt des Gedächtnisses eine Rolle. Auch ein Vorteil bei der Spielfeldwahrnehmung durch erlerntes Vorwissen, das die Analyse der Situation

und der Verhältnisse der Spielfiguren erleichtert, ist relevant. Höhere Expertise kann auch anhand einer verringerten nonverbalen Suchaktivität nach sinnvollen Spielzügen erkannt werden, da bessere Experten effizientere Suchstrategien besitzen (Vasyukova, 2012).

Hoffman (1996) schlägt vor, dass Experten anhand der Faktoren kognitive Entwicklung, Wissensstruktur und der verwendeten Urteilsprozesse definiert werden können. Unter kognitiver Entwicklung wird hier die Stufe der Kompetenz in der relevanten Domäne verstanden, die zu einer qualitativen Änderung der Leistung führt. Unter Wissensstruktur wird die Menge und Organisation von domänenspezifischem Wissen verstanden. Experten zeichnen sich dadurch aus, dass ihr Wissen konzept-, kontext- und inhaltsspezifisch ist. Unter Urteilsprozessen werden sowohl die Perzeptionsleistung, die Fähigkeit zur flexiblen Anpassung der Urteilslogik an Spezialfälle als auch die unbewusste Verinnerlichung des domänenspezifischen Wissens verstanden. Der letzte Punkt ist dabei mit dem fähigkeitsbasierten Fertigkeitsniveau in Rasmussens Drei-Ebenen-Modell vergleichbar (Rasmussen, 1983).

Phillips, Klein und Sieck (2004) schränken den Expertenbegriff auf solche Individuen ein, die eine herausragende Fähigkeit in einer bestimmten Domäne erreicht haben. Deswegen ist eine Nominierung durch Fachkollegen für den Expertenstatus unverzichtbar. Bei der Rekrutierung von Experten nutzten sie regelmäßig Fragen wie „Wer ist die Person, die alles [über diese Domäne] weiß?“. Sie sehen hohe Selektivität als notwendig an, um die von ihnen sogenannten Professionellen, die jedoch keine „wahren“ Experten sind, auszusieben, und nur solche Individuen zuzulassen, deren Leistung in der Domäne ohne erkennbaren Mangel ist.

Camerer und Johnson (1991) wenden hingegen eine vergleichsweise lockere Definition an und bezeichnen als Experten alle, die Erfahrung damit haben, Vorhersagen in der zu untersuchenden Domäne zu treffen, und zumindest irgendwelche professionelle oder soziale Anerkennung für ihre Tätigkeit genießen. Sie merken jedoch an, dass die mit diesen Kriterien rekrutierten Experten in der Regel nicht unterqualifiziert sind. Während sie zwar keine Unterscheidung zwischen außerordentlichen und nicht-außerordentlichen Experten machen, vermuten sie, dass Ergebnisse von einer Gruppe auf die andere übertragbar sind. Phillips et al. (2004), deren Definition im vorherigen Absatz beschrieben wurde, nehmen explizit Bezug auf diese Definition und bezeichnen sie als eine Definition von Professionellen. Sie heben den Unterschied zwischen diesen sogenannten Professionellen und Experten nach ihrer eigenen Definition hervor.

Shanteau (1992) operationalisiert den Expertenbegriff als Individuen, die die Fähigkeit besitzen, in ihrer Domäne auf höchstem Leistungsniveau zu handeln.

Die ISO 5492 definiert Experten in Bezug auf die sensorische Analyse von Nahrungsmitteln und Ernährungsprodukten als Personen, die durch Wissen oder Erfahrung die Kompetenz besitzen, ein Urteil in der betrachteten Domäne abzugeben. Dabei wird weiter spezifiziert: Ein sogenannter Expertengutachter („expert assessor“) ist ein aufgrund seiner Fähigkeit ausgewählter Gutachter mit einem hohen Grad an Sinnessensitivität und Erfahrung in der Methode, der fähig ist, konsistente und wiederholbare sinnliche Gutachten verschiedener Produkte zu erstellen.

Hart (1986) fordert von Experten, dass sie eine ausreichende Erfolgsrate besitzen, Lösungen zeiteffizient finden und gleichzeitig bereit sind, ihre eigenen Grenzen zu benennen.

Eine weitere mögliche Einteilung besteht in der Unterscheidung von absoluter und relativer Expertise. Absolute Experten zeichnen sich dadurch aus, dass sie in einer bestimmten Aufgabe messbar allen anderen Personen überlegen sind und die gegenwärtige Spitze menschlicher Leistungsfähigkeit auf diesem Gebiet markieren. Sogenannte relative Experten werden durch den Vergleich ihrer Fähigkeiten mit einer bestimmten Vergleichsgruppe definiert, wobei der Expertenstatus lediglich von dem Nachweis abhängt, bezüglich der Expertise höher als die Vergleichsgruppe zu liegen. Ob die Expertise für eine zuverlässige Beantwortung aller Fragen in diesem Fachbereich ausreicht, ist dabei unerheblich (Chi, 2006). Relative Experten lassen sich beispielsweise anhand einer Rangliste bestimmen, also im Beispiel Schach anhand des FIDE Ratings, das bestimmte Spieler als besser als andere ausweist, ohne dass eine Zugehörigkeit zur obersten Klasse notwendig wäre. Der Rang des Großmeisters ist jedoch den absolut gesehen besten Spielern vorbehalten, sodass diese als absolute Experten bezeichnet werden können.

In Burgman, Fidler, McBride, Walshe und Wintle (2006) wird weiterhin unterschieden zwischen Experten, die ihren Status a-priori erhalten, und solchen, deren Expertenstatus a-posteriori bestimmt wird. A-priori Experten werden aufgrund ihrer Erfahrung oder Qualifikation als solche identifiziert. Aufgrund der Eigenschaften der Person wird induziert, dass die Urteile dieser Person glaubhaft sind. Beispiele für dieses Vorgehen finden sich beispielsweise vor Gericht, wenn die Zulässigkeit eines Experten aufgrund dessen Lebenslauf oder ähnlicher formeller Faktoren beurteilt wird (Timmerbeil, 2003). Nach der Zulassung eines Experten fehlt den anderen Teilnehmern häufig die eigene Expertise, um möglicherweise berechtigte Zweifel an den Schlussfolgerungen des Experten

zu äußern, sodass hier ein Anreiz entsteht, mehr auf formelle Kriterien eines Experten als auf tatsächliche Expertise zu achten (Australian Law Reform Commission, 2000). Den Gegensatz dazu stellen a-posteriori Experten dar, deren Status erst nach einer objektiven Messung der Genauigkeit der Vorhersagen eines Expertenkandidaten vergeben wird. Meehl (1954) führte wegweisende Studien zur Güte von Expertenurteilen durch und stellte dabei fest, dass traditionelle formelle Kriterien für die Ernennung von Experten kein geeigneter Prädiktor für die tatsächliche Genauigkeit bei der Lösung von Problemen darstellt. Es ist jedoch nicht in allen Bereichen möglich, die Genauigkeit der Urteile von Experten mit vertretbarem Aufwand objektiv zu messen. Dies kann finanzielle, organisatorische oder auch rechtliche Gründe haben und führt dazu, dass der Expertenstatus oft mehr auf Vertrauen als auf Empirie basiert (Burgman et al., 2006).

Zusammenfassend lässt sich folgern, dass in unterschiedlichen Domänen eine Vielzahl an Definitionen für Experten entwickelt wurde (Farrington-Darby & Wilson, 2006). Obwohl es keinen generell akzeptierten Standard für eine allgemein gültige Definition von Experten gibt, muss aus praktischen Gründen für diese Dissertation eine Festlegung getroffen werden. Um den Untersuchungsgegenstand nicht unzulässig einzuengen, soll der Begriff Experte im Rahmen dieser Dissertation lediglich die Rolle der Person als qualifizierter Urteiler in einer Fragestellung bezeichnen. Damit bezeichnet der Begriff Experte nicht beliebige Urteiler, sondern nur solche, denen die Expertenrolle aufgrund von Kriterien zugewiesen wurde. Die Kriterien werden später festgelegt. Diese Definition des Expertenbegriffs lehnt sich an die weniger strenge Definition nach (O'Hagan et al., 2006) an.

Die hier getroffene Definition des Expertenbegriffs führt dazu, dass keine Aussage über das tatsächliche Vorhandensein der Expertise bei dem betroffenen Personenkreis gemacht wird. In diesem Modell ist die Expertise der als Experten bezeichneten Personen abhängig von den Auswahlkriterien für die Expertenrolle. Experten nach dieser Definition können also im Spezialfall über sehr große Expertise in einem bestimmten Bereich oder auch über keine besondere oder sogar unterdurchschnittliche Expertise verfügen. Das Ausmaß der Expertise und auch die Vertrauenswürdigkeit der Urteile ist keine direkte Konsequenz aus der Zuweisung der Bezeichnung Experte, sondern muss erst durch weitere Untersuchungen festgestellt werden.

2.5 Expertenbewertungen in anderen Bereichen

Expertenbewertungen unterschiedlichster Arten werden in vielen Domänen eingesetzt. Mit der Entwicklung probabilistischer Risikobewertung für Nuklearkraftwerke in dem WASH-1400 Bericht an die U.S. Nuclear Regulatory Commission wurden in dieser Domäne subjektive Expertenschätzungen für ansonsten unbekannte Variablen eingeführt (Rasmussen & et al., 1975). Der darauf folgende Review-Bericht kritisierte die Methodik des WASH-1400, begrüßte aber explizit die Verwendung von subjektiven Expertenurteilen (Lewis et al., 1979). Trotz dauerhafter Kritik an der Methode subjektiver Expertenschätzung bleibt sie ein wichtiger Bestandteil in der Risikoanalyse auch von Atomkraftwerken (Cooke & Goossens, 2000; World Health Organization, 2013). Weitere Beispiele für die Anwendung von Expertenbewertungen finden sich unter anderem in den Bereichen:

- Ökonomie (Braun & Yaniv, 1992; De Bondt, 1991),
- Weinbewertung (Ashton, 2012; Sauvageaot, Urdapilleta & Peyron, 2006),
- Rechtswissenschaften (Englich & Mussweiler, 2001),
- Klimaforschung (Morgan, 2014),
- Meteorologie (Stewart, Moninger, Grassia, Brady & Merrem, 1989),
- probabilistisches Risikoassessment (Clemen & Winkler, 1999; Ouchi, 2014),
- biologische Sicherheit (Burgman et al., 2006),
- Luftfahrt (Harper & Cooper, 1986; Helmreich, Merrit & Sherman, 1996),
- Versicherungen (Cabantous, Hilton, Kunreuther & Michel-Kerjan, 2011; Kunreuther, Pauly & McMorrow, 2013),
- Medizin (Ayanian, Landrum, Normand, Guadagnoli & McNeil, 1998; Chapman, 2004) und
- Politik (Morgan, 2014; Tetlock, 2005).

Alle diese Bereiche haben gemein, dass ihre Untersuchungsgegenstände aufgrund wirtschaftlicher Voraussetzungen, Sicherheitsaspekten oder ihrer Komplexität nicht für eine analytische oder empirische Bewertung geeignet sind. Deswegen gibt es eine große

Bandbreite an Beiträgen zum Thema Expertiseforschung. Diese spiegelt sich in den Methoden wider, die in diesem Bereich angewendet werden, von denen einige allgemein anwendbare im nächsten Unterkapitel diskutiert werden.

2.6 Methode der Expertenbefragung

Expertenbefragungen mit mehreren Experten sind ein komplexer Vorgang. Nicht nur die eigentliche Fragestellung muss hier betrachtet werden, sondern auch die Methode, wie die Urteile der verschiedenen Experten entnommen werden und wie diese dann miteinander kombiniert werden, falls diese nicht vollständig miteinander übereinstimmen. Die Vielfalt an Fachbereichen, in denen Expertenbewertungen durchgeführt werden, spiegelt sich auch in einer Vielfalt an Methoden bei der Durchführung von Expertenbewertungen wider. Dabei ist das ganze Gebiet Expertenbefragungen durch Methoden gekennzeichnet, die individuell für spezielle Fragestellungen entwickelt und angewandt werden. Etablierte Methoden, die weitläufig angewendet werden und ein allgemein akzeptiertes Fundament für Expertenbewertungen bilden, existieren nicht. In diesem Abschnitt werden folgende Inhalte diskutiert:

- Grundlagen der Durchführung von Expertenbewertungen
    - Reproduzierbarkeit
    - Zurechenbarkeit
    - Empirische Kontrolle
    - Neutralität
    - Fairness
- Mathematische Urteilszusammenführung
    - Nicht-bayes’sche Methoden
    - Bayes’sche Methoden
    - Psychologische Skalierungsmethoden
- Verhaltensbasiere Urteilszusammenführung
    - Delphi-Methode
    - Nominal-Group-Technik

Cooke (1991) nennt Richtlinien für das Design und die Durchführung von Expertenbewertungen, die als Grundlage für die Entwicklung von Expertenbefragungen verwendet werden können. Er definierte fünf erforderliche Prinzipien, die er aus dem Zielbild der Rationalität ableitet. Sein Ziel besteht darin, subjektive Urteile als Ressource für die

Wissenschaft nutzbar zu machen. Die von ihm empfohlenen Richtlinien sind Reproduzierbarkeit, Zurechenbarkeit, empirische Kontrolle, Neutralität und Fairness.

Das erste von Cooke genannte Prinzip ist Reproduzierbarkeit. Obwohl dies eigentlich ein selbstverständliches Element der empirischen Forschung ist, betont er, dass sich in der Literatur dennoch eine große Anzahl an Expertenstudien finden, die tatsächlich nicht reproduzierbar sind. Dies kann daran liegen, dass die angewendete Methode nicht ausreichend dokumentiert ist oder aufgrund fehlerhaften Versuchsdesigns systematisch nicht reproduzierbar ist. Insbesondere bei sicherheitsrelevanten Studien ist also auf eine vollständige Reproduzierbarkeit aller Bestandteile der angewendeten Methode zu achten.

Das zweite Prinzip ist die Zurechenbarkeit. Auch diese ist ein Grundpfeiler wissenschaftlicher Arbeit. Cooke wendet den Begriff jedoch nicht nur auf den Durchführer der Studie an, der sich für die Ergebnisse verantwortlich zeigen muss, sondern verlangt auch, dass alle in die Studie eingehenden Expertenurteile auf ihren Urheber zurückverfolgt werden können. Dies ist gerade deswegen von Bedeutung, da Experten in einer Domäne oft auch ein eigenes – direktes oder indirektes – Interesse in dem Bereich haben, das im Konflikt mit dem Untersuchungsgegenstand stehen kann, selbst wenn dieser Konflikt zum Untersuchungszeitpunkt noch nicht öffentlich bekannt ist. Expertensysteme, in denen subjektive Urteile vieler Experten auf nicht-trivialer Art und Weise zu gänzlich neuen Aussagen synthetisiert werden, führen zu einer mangelhaften Zurechenbarkeit, die später eine Rückverfolgung von Fehlentscheidung unmöglich macht. Dabei betont Cooke, dass für ausreichende Zurechenbarkeit eine Veröffentlichung der Namen der Datenquellen nicht zwingend notwendig ist.

Auch das Prinzip der empirischen Kontrolle ist in der wissenschaftlichen Praxis nicht neu. Cooke stellt fest, dass dieses Prinzip in der Anwendung auf Expertenbewertungen bedeutet, dass es zumindest theoretisch möglich sein muss, die Ergebnisse einer Expertenbewertung objektiv zu widerlegen, selbst wenn sich der Untersuchungsgegenstand nur schlecht für andere Untersuchungsmethoden eignet. Durch die Verankerung der Möglichkeit empirischer Kontrolle im Design von Expertenbewertungen wird sichergestellt, dass die teilnehmenden Experten ein Interesse haben, das wahrscheinlichste Urteil abzugeben, und dass Fehlurteile korrigiert werden können.

Unter dem Prinzip der Neutralität versteht Cooke, dass das Design einer Expertenbewertung so gestaltet sein muss, dass die teilnehmenden Experten ein dominierendes Interesse daran haben, die ihrer Meinung nach wahrscheinlichste Antwort zu nennen.

Cooke kritisiert damit explizit Methoden, die ein erhöhtes Gewicht auf diejenigen Experten legen, deren Urteile konsequent nahe an den durchschnittlichen Urteilen der restlichen Experten liegen. Durch eine solche Gestaltung verschiebe sich das Ziel der Studie zum Schätzen der Mehrheitsmeinung und weg vom Schätzen des eigentlichen Zielparameters.

Das fünfte und letzte Prinzip von Cooke ist Fairness. Damit wird bezeichnet, dass vor Beobachtung der Leistung der Experten alle Teilnehmer gleich behandelt werden müssen, sofern keine empirischen Daten existieren, die eine Unterscheidung der Urteiler rechtfertigen. Dies sei insbesondere bei bayes'schen Methoden der Urteilssynthese zu beachten, da hier schon vor Untersuchungsbeginn eine a-priori Einschätzung des Vertrauens in die Experten notwendig ist (Cooke, 1991).

Sofern mehrere Experten für die Beurteilung einer Fragestellung zur Verfügung stehen, stellt sich die Frage, wie anschließend die Mehrzahl an Urteilen zu einem Ergebnis zusammengeführt werden kann. Clemen und Winkler (1999) unterscheiden mathematische und verhaltensbezogene (behavioral) Methoden. Zu den letzteren zählen beispielsweise die Delphi-Methode oder die Nominal-Group-Technik. Ouchi (2014) unterscheidet mathematische Modelle weiter in drei verschiedene Klassen: Nicht-bayes'sche Methoden, bayes'sche Methoden und psychologische Skalierungsmethoden. Nicht-bayes'sche Methoden bestehen beispielsweise in einer gewichteten Mittelwertbildung mittels des geometrischen, logarithmischen oder arithmetischen Mittels. Ein Kernproblem bei dieser Methode besteht darin, die Gewichtungen für die Kombination der Urteile ideal zu wählen. Eine detaillierte Betrachtung dieser Methoden und zur Wahl der Gewichte findet sich bei Bedford und Cooke (2001).

In der Literatur finden sich jedoch Hinweise auf spezifische Nachteile von Gruppenbewertungen durch Experten. Wenn von rationalen Entscheidern ausgegangen wird, die aufgrund ihrer unterschiedlichen Erfahrungen und Wissensständen Urteile treffen, die symmetrisch um die wahre Lösung verteilt sind, ist zu erwarten, dass der Wert, auf den sich mehrere Experten einigen können, weniger von der wahren Lösung abweicht, als die ursprünglichen Einzelurteile. Da jedoch der Urteilsbildungsprozess innerhalb von Gruppen auch eine soziale Komponente hat und von der Gruppendynamik der individuellen Teilnehmer abhängt, ist die Annahme der Rationalität solcher Expertengruppen zu bezweifeln (Burgman et al., 2006). Es wurden zwar Verfahren, wie beispielsweise die Delphi-Methode, entwickelt, um Gruppenentscheidungen, welche nicht auf Basis mathematischer Überlegungen zusammengeführt werden können, zu strukturieren, diese

sind aber nicht immer effektiv und auch ihrerseits mit Nachteilen verbunden (Green, Armstrong & Graefe, 2007).

Expertenbewertungen werden häufig dann durchgeführt, wenn der Betrachtungsgegenstand für andere Untersuchungsmethoden nicht zur Verfügung steht. Dies ist beispielsweise dann der Fall, wenn eine experimentelle Analyse durch Sicherheitsbedenken eingeschränkt wird. Dennoch gibt es auch bei der Expertenbewertung gegebenenfalls mehrere mögliche Versuchsumgebungen mit unterschiedlichen Graden an ökologischer Validität. Beim sogenannten Hazard Perception Task, bei dem Urteiler entscheiden müssen, ob Fahrsituationen im Straßenverkehr kritisch sind oder nicht, stellt sich oft die Wahl zwischen einer interaktiven Darstellung der Fahrszene mit Markierung kritischer Elemente in der Fahrszene oder einer nicht-interaktiven Darstellung mit Beantwortung eines Fragebogens. Malone und Brünken (2015) untersuchten diese zwei Varianten und fanden keinen signifikanten Interaktionseffekt zwischen der Versuchsumgebung und der Expertise.

2.7 Naturalistic Decision Making

Der Bereich Naturalistic Decision Making (NDM) ist ein Zweig der Expertiseforschung und konzentriert sich auf eine praxisnahe Untersuchung von Entscheidungsprozessen. In ihm wird die Übertragung von Effekten, die unter Laborbedingungen gemessen werden, auf reale Anwendungssituationen von Expertise als nicht ausreichend valide gesehen. Im Bereich NDM wird nicht bestritten, dass die unter Laborbedingungen gemessenen Effekte in den untersuchten Populationen tatsächlich existieren. Allerdings wird die externe Validität dieser Effekte in Frage gestellt, da reale Entscheidungen nicht unter Laborbedingungen getroffen werden und kritische Entscheidungen, die von großer Bedeutung sind, von Personen mit sehr hoher Expertise getroffen werden, was bei vielen Laboruntersuchungen nicht der Fall ist (Klein, 2008; Lipshitz, Klein, Orasanu & Salas, 2001).

Mit dem Ziel der Erhöhung der externen Validität der Untersuchungsergebnisse konzentrieren sich Untersuchungen im Bereich NDM auf reale Untersuchungsumgebungen in der beruflichen Praxis der zu untersuchenden Experten. Der damit verbundene Nachteil an Kontrolle der Versuchsbedingungen und die oft geringe Stichprobengröße sind Nachteile, die im NDM jedoch in Kauf genommen werden (Kahneman & Klein, 2009).

Da im NDM Experten anhand ihrer hohen Leistungsfähigkeit definiert werden, ist eine Untersuchung der Gründe für mangelhafte Leistung der betrachteten Experten rein logisch ausgeschlossen. In diesem Bereich wird diese Frage ersetzt durch die Frage, warum auch hocherfahrene Fachleute nicht den Expertenstatus erreichen oder wie aus Fachleuten wahre Experten werden. Eine andere Kernfrage im NDM ist die Übertragbarkeit des Domänenwissens von Experten, wobei die Domäne des Experten per Definition der Bereich ist, in dem keine Abweichung von hervorragender Leistung erkennbar ist (Klein, 2008).

Im NDM wird angenommen, dass der Entscheidungsprozess eines Experten zunächst auf Basis eines Vergleichs der aktuellen Situation mit einer großen Datenbank an Erfahrungen durchgeführt wird, wobei dieser Vorgang aufgrund der großen Erfahrung des Experten nicht zwingend ein willentlicher Akt ist sondern auch unwillkürlich stattfinden kann. Erst bei einer empfundenen Abweichung der realen Situation vom Erfahrungsschatz wird geprüft, ob die identifizierten Abweichungen eine Anpassung der Handlungsstrategie notwendig macht oder ob sogar neue Strategien entwickelt werden müssen. Dieses Modell der Urteilsfindung wird als Recognition-Primed-Decision (RPD) Modell bezeichnet. Eine detaillierte Darstellung findet sich in (Klein, 1998). Während zwar auch eine Reihe anderer Modelle existieren, ist das RPD Modell im Bereich der NDM dominant (Lipshitz et al., 2001).

Ein wichtiges Distinktionsmerkmal der NDM ist, dass Experten eine hohe Urteilsqualität erreichen können, ohne explizites Wissen über die Wirkzusammenhänge in der Domäne ihrer Expertise zu besitzen. Ein Beispiel dafür sind etwa Feuerwehrmänner, die die Entwicklung eines Hausbrandes vorhersagen können, ohne explizite Kenntnis des Grundrisses des Gebäudes und der Brennbarkeitseigenschaften der darin enthaltenen Materialien zu haben (Klein, 2008). Damit differenziert sich die NDM beispielsweise von verwandten Bereichen wie dem Rationalistic Decision Making, in dem Entscheidungen als Folge eines bewussten Denkvorgangs mit rationalen Argumenten aufgefasst werden (Hassard, 2009).

Obwohl zwar nicht vorausgesetzt wird, dass erfolgreiche Experten explizites Wissen über den Bereich ihrer Expertise verfügen, ist es auch im Kontext der NDM für die Entwicklung von Expertise notwendig, Erfahrungen in einer validen Umgebung zu sammeln und direktes Feedback über die Leistung des Urteilenden zu erhalten. Umgebungen, in denen nicht alle relevanten Einflussfaktoren beobachtbare Größen sind oder bei

denen das Ergebnis der Arbeit nicht vollständig nachvollzogen werden kann, eignen sich nicht für die Entwicklung von Expertise (Kahneman & Klein, 2009).

2.8 Kognitive Verzerrungen und Heuristiken

Der Ansatz kognitiver Verzerrungen und Heuristiken (V&H) unterscheidet sich von den Prinzipien der NDM durch grundsätzliche Zweifel an der Leistungsfähigkeit menschlicher Expertise. Er lässt sich zurückverfolgen auf Betrachtungen von Meehl (1954), der in einer 20 Studien umfassenden Meta-Analyse feststellte, dass selbst einfache statistische Modelle eine höhere Genauigkeit bei der Vorhersage von Patientenverhalten als Einschätzungen von professionellen Psychologen hatten. Später stellten Grove, Zald, Lebow, Snitz und Nelson (2000) in einer größeren Meta-Analyse fest, dass nur in 6 von 136 betrachteten Studien subjektive Methoden einer rein statistischen Analyse überlegen waren.

Goldberg (1970) untersuchte das Diagnoseverhalten von 29 Psychologen, indem er ihre Diagnosen auf Basis des Minnesota Multiphasic Personality Index (MMPI) anhand eines linearen Modells nachbildete. In der Validierungsstichprobe erzielten die linearen Modelle des Urteilsverhaltens der Psychologen überraschenderweise eine größere Genauigkeit als die Psychologen selbst. Da das lineare Modell nur mit einem Teil der verfügbaren Daten über die Patienten gebildet wurde, war eigentlich davon auszugehen, dass die Urteile der Psychologen diesen überlegen sein müssten. Goldberg folgerte daraus, dass sich die Psychologen durch irrelevante Faktoren in ihrer Diagnose beeinflussen lassen. Da das lineare Modell davon nicht betroffen ist, ergibt sich die bessere Leistung der statistischen Methode.

Der allgemeine Zweifel an der Validität subjektiver Urteile führt zum Ansatz der kognitiven Verzerrungen und Heuristiken, in dem versucht wird, die Abweichungen subjektiver Urteile von normativen Lösungen zu identifizieren und zu klassifizieren. Aus diesen Forschungen ergaben sich eine große Anzahl sogenannter kognitiver Verzerrungen und Heuristiken. Heuristiken bezeichnen „mentale Abkürzungen“, die es Menschen erlauben, auch komplexe Probleme zeit- und aufwandseffizient in einer großen Anzahl an Fällen zu lösen. Während diese Strategien häufig zu nahezu optimalen Lösungen führen, wird im Rahmen der V&H impliziert, dass sie in Sonderfällen, die mitunter praxisrelevant sein können, von der normativen Lösung abweichen. Kognitive Verzerrungen bezeichnen die systematischen Abweichungen von der wahren Lösung, die aus der fehlerhaften Anwendung von Heuristiken resultieren. Im Rahmen der V&H wird dabei

explizit betont, dass Heuristiken prinzipiell nützlich sind, um schnell und effizient Lösungen für auch normativ äußerst komplexe Probleme zu finden. Der Fehler besteht lediglich in der fehlerhaften Anwendung der gelernten Heuristiken in Fällen, in denen sie nicht gültig sind. Es ist im Rahmen der V&H jedoch Konsens, dass Heuristiken sowohl bei Laien als auch bei Fachleuten zu häufig angewendet werden und dadurch das subjektive Urteilsvermögen auch von hochqualifizierten Experten als nicht vertrauenswürdig zu gelten hat (Kahneman, 2012).

In der Literatur wurde eine Vielzahl an Heuristiken identifiziert. Caputo (2013) listet sechs häufig untersuchte Heuristiken auf:

- Verfügbarkeitsheuristik: Beim Abschätzen von Wahrscheinlichkeiten oder Häufigkeiten lassen sich Probanden durch die mentale Verfügbarkeit von Ereignissen irreführen.

- Repräsentativitätsheuristik: Beim Treffen von Urteilen über ein Individuum oder Ereignis werden Informationen über einen zuvor geformten Stereotyp verwendet, der dem Individuum oder Ereignis zugeordnet wurde.

- Bestätigungsheuristik: Beim Testen einer Hypothese werden vorzugsweise solche Informationen ausgewählt, die die Hypothese bestätigen.

- Affektheuristik: Urteile werden teilweise auf Grundlage der emotionalen Einstellung zu einer Person, Gruppe oder Sache getroffen, noch bevor eine logische Analyse auf höherer Ebene durchgeführt wird (engl. „integrative affect heuristic“). Tritt auch auf, wenn die Stimmungslage des Urteilers unabhängig von der Einstellung zum Betrachtungsgegenstand das Urteilsverhalten beeinflusst (engl. „incidental affect heuristic“).

- Begrenzte Aufmerksamkeit (engl. „bounded awareness“): Bei der unbewussten Auswahl der zu berücksichtigenden Informationen können eigentlich relevante Informationen unterdrückt werden, sodass diese dann beim Entscheidungsprozess nicht mehr zur Verfügung stehen.

- Risikoaversion: Ein Risiko, einen bestimmten Verlust zu erleiden, wird anders bewertet, als die gleiche Wahrscheinlichkeit, einen ebenso großen Gewinn zu erhalten.

Diese Heuristiken resultieren in einer großen Anzahl an kognitiven Verzerrungen. Caputo (2013) führt 21 kognitive Verzerrungen auf. Eine der häufig untersuchten kognitiven Verzerrungen ist die Verankerung („anchoring“). Sie bezeichnet den Einfluss irrelevanter Faktoren auf den Urteilsprozess einer Person. In einer frühen Studie zu diesem Thema sollten Teilnehmer den Anteil von Nationen vom afrikanischen Kontinent in den Vereinigten Nationen schätzen (Tversky & Kahneman, 1974). Bevor sie ihr Urteil abgaben wurde jedoch ein Glücksrad gedreht, das entweder 10% oder 65% ergab. Obwohl dieses Ergebnis nichts mit der eigentlich zu bewertenden Frage zu tun hat und auch im Versuchskontext kein weiterer Bezug hergestellt wurde, ergab sich eine Korrelation zwischen dem Ergebnis des Glücksrads und dem Urteil der Probanden. Dieser Effekt ist sehr robust gegenüber Manipulationen und wurde für unterschiedliche Aufgaben und Rahmenbedingungen nachgewiesen. Eine Übersicht ist in (Chapman & Johnson, 2002; Epley & Gilovich, 2004; Mussweiler, Englich & Strack, 2004) dargestellt.

Dieser Effekt ist jedoch nicht nur auf Urteile von Personen mit geringer Expertise in dem betroffenen Fachgebiet beschränkt. So wurde bei einer Untersuchung des Urteilsverhaltens von Richtern festgestellt, dass dieses durch offensichtlich irrelevante Faktoren beeinflusst werden kann. In einer Studie wurde das Urteilsverhalten praktizierender Richter in Deutschland in Form von Fallstudien untersucht. Dabei füllten die Richter das vom Staatsanwalt zu fordernde Strafmaß selbst aus, nachdem sie es eigenhändig durch Würfeln ermittelt hatten. Obwohl den Richtern also offenkundig bekannt war, dass das geforderte Strafmaß alleine auf Zufall basierte, beeinflusste es ihr Urteil. Jene Richter, die ein hohes Strafmaß gewürfelt hatten, urteilten im Mittel höher als jene, die ein geringeres Strafmaß würfelten (Englich, Mussweiler & Strack, 2006).

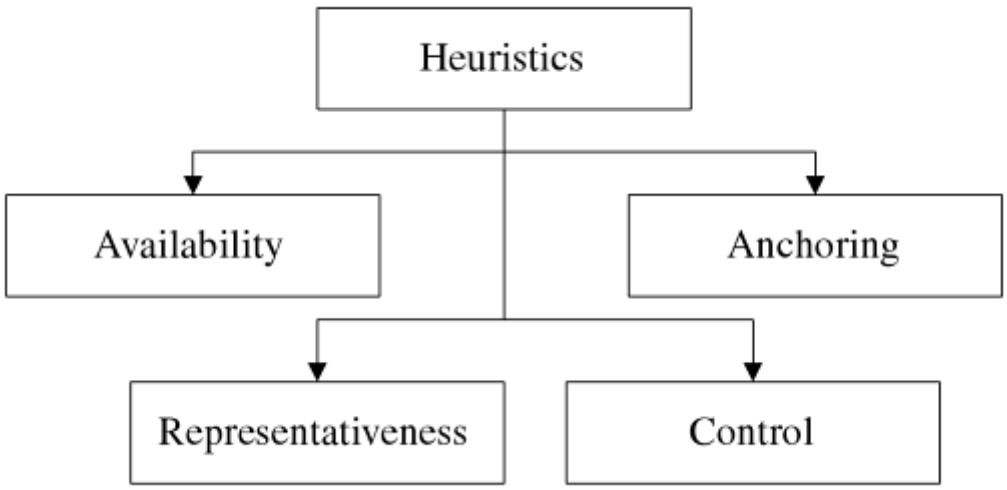

Abbildung 2 Einflussfaktoren auf Heuristiken (Ayyub, 2001).

Verschiedene Versuche, um die negativen Konsequenzen von Heuristiken zu beheben oder zu beschränken, werden unter dem Begriff Debiasing (von engl. „bias“ für kognitive Verzerrung) diskutiert. Ein Kernprinzip des Debiasing besteht in der Aktivierung bewusster mentaler Prozesse, um die unbewusste Verwendung von Heuristiken in Fällen, in denen diese nicht zu einer guten Lösung führen, zu verhindern. Die Hoffnung besteht darin, dass bewusstes Lösen des zu betrachtenden Problems zu einer Lösung führt, die näher an der normativen Lösung liegt. Ein anderes Prinzip besteht darin, die Wirkung der Heuristik zu akzeptieren und ihren Effekt durch Training auszugleichen. Für beide Prinzipien existieren mehrere Verfahren. Darunter fallen beispielsweise Restrukturierung der Aufgabe, Aufklärung über Art und Wirkung relevanter kognitiver Verzerrungen oder Veränderung der Antwortskala. Auch eine Erhöhung der Motivation zur Vermeidung von Heuristiken wurde als Methode zur Verbesserung der Urteilsqualität untersucht. Das allgemeine Fazit einer großen Anzahl an Studien zum Thema Debiasing ist jedoch, dass kognitive Verzerrungen in einer großen Bandbreite an Situationen robust auftreten und sich nicht durch allgemeine Manipulationen beheben lassen (Fischoff, 1981; Larrick, 2004; Wilson, Houston, Etling & Brekke, 1996).

2.9 Entwicklung von Expertise

Ein wichtiger Bestandteil der Expertiseforschung widmet sich dem Verständnis der Entwicklung von Expertise. Ericsson, Krampe und Tesch-Romer (1993) entwickelten anhand von Interviews mit Violinisten unterschiedlicher Fähigkeitsniveaus die Theorie bewusster Anstrengung („deliberate effort“). Sie beobachteten zwar qualitativ unterschiedene Niveaus der Expertise, folgerten aus ihren Beobachtungen jedoch, dass diese nicht durch ein inhärentes Talent zustande kommt, sondern durch große Mengen bewusster Anstrengung zur Aneignung von Expertise. Insbesondere stellten sie fest, dass Individuen mit hoher Expertise anstrengende Übungen als besonders relevant für die Verbesserung der gewünschten Fähigkeiten betrachteten und sie im Vergleich zu in dem untersuchten Feld weniger fähigen Individuen besonders häufig ausübten. Die Theorie bewusster Anstrengung wurde später in einer Vielzahl anderer Domänen überprüft und bestätigt, darunter in (Hodge & Deakin, 1998) und (Ward, Hodges, Starkes & Williams, 2007). In verschiedenen Bereichen hat sich herausgestellt, dass die beobachtete Leistung nicht durch große professionelle Erfahrung vorhergesagt werden kann, jedoch auf bewusstes Üben zurückgeführt werden kann (Ericsson, 2008).

Im Bereich mentaler Tätigkeiten, die nicht auf einer besonders hohen Fähigkeit zur Körperbeherrschung basieren, sondern auf geistigen Assoziationen basieren, ist das Prinzip der Gewinnung von Expertise ähnlich. Hier wird auch von befähigter Intuition („skilled intuition“) gesprochen. Diese setzt voraus, dass die Situation einen Hinweis gibt, der es dem Experten erlaubt, zusätzliche Informationen aus dem Gedächtnis abzurufen und damit die Antwort für die gegebene Problemstellung zu finden. Diese Tätigkeit kann ebenso geübt werden wie manuelle Tätigkeiten (Simon, 1992). So wurde die Theorie bewusster Anstrengung auch für mentale Tätigkeiten erfolgreich getestet, beispielsweise in der Mathematik (Buttersworth, 2006).

Expertise, die nicht auf reiner Routine, sondern einem breiten Verständnis der Domäne basiert, setzt weiterhin voraus, dass der Lernende Übungen durchführt, bei denen eigene Lösungsstrategien entwickelt werden können. Solche Übungen haben ein hohes Potential für Fehler und können zu einer langsameren Entwicklung messbarer Verbesserung führen, fördern jedoch das Vernetzen der kausalen Zusammenhänge in der Domäne (Carbonell, Stalmeijer, Könings, Segers Mien & van Marrienboer, 2014; Chi, Glaser & Rees, 1982).

2.10 Probabilistische Risikoanalyse

Alternativ zur Risikobeurteilung durch empirische Versuche oder subjektiven Abschätzungen gibt es weiterhin die Möglichkeit, Systeme, deren innere Mechanismen zumindest teilweise bekannt sind, als Verkettung mehrerer mit statistischer Unsicherheit behafteter Vorgänge aufzufassen. So wurde etwa im Bereich der Nuklearsicherheit die THERP-Methode entwickelt, um die Rate menschlicher Fehler beim Betrieb von Nuklearkraftwerken zu bestimmen (Swain & Guttmann, 1983). Bei dieser Methode wird von einem Systemfehler ausgegangen und in einem Ereignisbaum alle darauf folgenden menschlichen und technischen Handlungsschritte notiert. Das Resultat der Handlungsabfolgen wird abgeschätzt und jede Handlung, bei welcher mehrere möglich sind, mit Wahrscheinlichkeiten belegt. Dadurch kann die Eintrittswahrscheinlichkeit aller aus dem ursprünglichen Fehlerfall entstehenden Schäden abgeschätzt werden. Die Methode ist vergleichbar mit einer Fehlerbaumanalyse und daher gleichfalls darauf angewiesen, dass der Baum an möglichen Handlungen vollständig ist und die mit jeder Aktion verbundenen Wahrscheinlichkeiten die Realität zufriedenstellend abbilden (Kirwan, 1994). Fehlerbaumanalysen werden im Zusammenhang mit der ISO 26262 bereits bei der Be-

wertung der funktionalen Sicherheit von Fahrerassistenzsystemen eingesetzt und können auch auf die Bewertung der Beherrschbarkeit übertragen werden. Dieses Modell kann dann verwendet werden, die Wahrscheinlichkeit einer Abfolge von Bedienhandlungen zu bestimmen, wodurch die Eingangswahrscheinlichkeiten für eine Beherrschbarkeitsbewertung von Fahrerassistenzsystemen bestimmt werden können. Eine Dekomposition einer Tätigkeit, wie etwa das Fahren eines Lastkraftwagens (LKW), in Form einer hierarchischen Aufgabenanalyse ist möglich, führt jedoch zu einer großen Anzahl an Elementaraufgaben, die jeweils mit Erfolgs- und Misserfolgswahrscheinlichkeiten belegt werden müssen (Bedinger, Walker, Piecyk, Greening & Krupenia, 2015).

Dennoch werden derartige Methoden bereits im Automobilsektor eingesetzt, um den menschlichen Beitrag zu Unfällen zu bestimmen (Rangra, Sallak, Schön & Vanderhaegen, 2015). Auf dem gleichen Ansatz aufbauend präsentierte Helmer (2015) eine numerische Simulationsmethode, um nicht nur die Grundwahrscheinlichkeit der Ereignisse im Ereignisbaum, sondern auch die jeweiligen Verteilungsfunktionen und die aus den Aktionen der Verkehrsteilnehmer resultierenden kinematischen Zusammenhänge zu berücksichtigen. Da diese Methoden auf validierbaren Grundannahmen (Wahrscheinlichkeiten) basieren und ihre Ergebnisse zumindest im Prinzip extern validiert werden können, sind diese Konzepte in Zukunft möglicherweise verlässliche Werkzeuge für die Fahrerassistenzsystementwicklung (Kompaß, Helmer, Wang & Kates, 2015).

2.11 Kritikalität von Fahrsituationen

Durch Probandenstudien lässt sich in gewissen Rahmen eine realistische Abschätzung der objektiven Beherrschbarkeit von Fahrerassistenzsystemen in bestimmten Situationen gewinnen. Auch probabilistische Methoden haben zum Ziel, den Bereich der objektiven Beherrschbarkeit möglichst genau zu bestimmen. Dabei bleibt jedoch unklar, welche Störungen der Fahraufgabe durch Fahrerassistenzsysteme vom Fahrer als hinnehmbar oder inakzeptabel empfunden werden. Diese ist jedoch für eine Auslegung von Systemen ebenfalls von zentraler Bedeutung (Neukum & Krüger, 2003).

Die sogenannte Störungsbewertungsskala wurde durch Neukum und Krüger (2003) entwickelt, um das subjektive Empfinden der Störung des Fahrers für den Fall von Lenkwinkelfehlern zu operationalisieren. Es handelt sich um eine eindimensionale, hierarchische Skala, die im ersten Schritt fünf Grobkategorien für die Auswirkung der Störung beschreibt. Diese lauten „nichts bemerkt“, „die Störung wurde bemerkt“, „das Fahren

wurde gestört“, „die Störung war gefährlich“ und „das Fahrzeug war nicht mehr kontrollierbar“ Die drei mittleren Kategorien werden in jeweils drei Unterstufen unterteilt, sodass sich insgesamt eine 11-stufige Skala ergibt. Die Störungsbewertungsskala (SBS) wird in Abbildung 3 dargestellt (Neukum & Reinelt, 2005).

Abbildung 3 Störungsbewertungsskala (Neukum & Krüger, 2003).

Die Kategorien werden wie folgt definiert (Neukum & Krüger, 2003): In der Kategorie „Spürbar“ werden Störungen erfasst, die vom Fahrer als solche bemerkt werden, jedoch nur geringfügige oder keine Kompensation durch den Fahrer erfordern. Komforteinbußen sind möglich, (subjektive) Einschränkungen der Sicherheit jedoch nicht. In die Kategorie „Störung des Fahrens“ fallen Fehler, bei denen der Fahrer eine erhebliche, jedoch noch als vertretbar eingeschätzte Kompensation der Störung durchführen muss. Auch Situationen dieser Kategorie dürfen nicht mit Einschränkungen der Sicherheit verbunden sein. Die Kategorie „gefährlich“ betrifft Eingriffe, bei denen der Kompensationsaufwand als inakzeptabel eingestuft wird oder bei denen sicherheitskritische Situationen auftreten. Die beiden Randkategorien „nichts bemerkt“ und „unbeherrschbar“ schließlich bezeichnen die Fälle, in denen die Störung entweder überhaupt nicht festgestellt wurde oder in denen der Fahrer den Eindruck hatte, einen Kontrollverlust des Fahrzeugs nicht mehr verhindern zu können (Neukum & Krüger, 2003).

Die Störungsbewertungsskala erfasst nicht die Beherrschbarkeit selbst, sondern lediglich die subjektiv empfundene Störung des Fahrers. Es ist also für die Bewertung der

Beherrschbarkeit in einer kritischen Fahrsituation immer notwendig, auch objektive Maße der Beherrschbarkeit heranzuziehen. So gesehen stellt die subjektive Bewertung der Störung ein zusätzliches, konservatives Maß dar, das etwa genutzt werden kann, um genauere Untersuchungen auffälliger Fahrsituationen zu motivieren (Neukum & Reinelt, 2005).

Die Skala nimmt insgesamt keinen Bezug zu Störungen von Lenksystemen, obwohl sie für diese entwickelt worden ist. Dadurch ist einerseits auch für Normalfahrer ein Vergleich der empfundenen Störung mit anderen alltäglichen Störungen möglich, andererseits kann die Skala damit auch für einen Vergleich verschiedener Situationen verwendet werden (Neukum & Reinelt, 2005).

Die SBS wurde in verschiedenen Varianten bereits für mehrere Studien erfolgreich eingesetzt (Wesp, 2011). Entwickelt wurde die Skala für eine Studie von Lenksystemen, bei der festgestellt wurde, dass die SBS ein wesentlich sensibleres Kriterium für die Bewertung der Beherrschbarkeit als binäre pass-fail-Kriterien wie etwa Spurverletzungen darstellt (Neukum & Krüger, 2003). Bereits bei dieser Studie wurden die Urteile von Normal- und Profifahrern verglichen. Hierbei wurde anhand einer kleinen Stichprobe festgestellt, dass Profifahrer kritischere Urteile als Normalfahrer abgeben. Auch für längsdynamischen Szenarien wurde die Skala bereits validiert (Neukum, Lübbeke, Krüger, Mayser & Steinle, 2008).

Als zusätzliches Kriterium zu objektiven pass-fail Kriterien ist es nicht möglich, aus den Urteilen auf der Störungsbewertungsskala alleine auf die Beherrschbarkeit von Fahrsituationen zu schließen. Eine frühe Abschätzung der zu erwartenden Urteile auf der SBS kann jedoch bei der Durchführung von Beherrschbarkeitsstudien helfen, da hierdurch vermieden wird, Studien an Szenarien durchzuführen, die aufgrund des subjektiven Störungsempfindens der naiven Probanden ohnehin inakzeptabel sind.

2.12 Signalentdeckungs- und Spieltheorie

Die Bewertung der Beherrschbarkeit von Fahrerassistenzsystemen ist eine mit Unsicherheit behaftete Messaufgabe. Die Messgröße ist die dichotome Eigenschaft der Beherrschbarkeit bzw. Nicht-Beherrschbarkeit und es können verschiedene Methoden, darunter die Expertenbefragung, eingesetzt werden, um diese zu messen. Die mit Unsicherheit behaftete Messung einer dichotomen Eigenschaft wird durch die Signalentdeckungstheorie (SET) beschrieben. Anhand dieser soll im Folgenden dargestellt werden,

unter welchen theoretischen Randbedingungen die Bewertung der Beherrschbarkeit erfolgt und wie die „Qualität“ der Bewertung der Beherrschbarkeit quantifiziert werden kann.

Die SET wurde von Green und Swets (1966) entwickelt und beschreibt einen theoretischen Rahmen, um die Detektion schwierig zu erkennender Signale zu beschreiben.

In der SET wird vorausgesetzt, dass das zu suchende Merkmal dichotomer Natur ist, also entweder vorliegt oder nicht, und dass das Urteil über das Vorhandensein des Merkmals auch nur in binärer Form erfolgen kann. Unter diesen Voraussetzungen ergibt sich in Abhängigkeit des tatsächlichen und gefundenen bzw. gemessenen Zustands die in Tabelle 4 dargestellte Zustandsmatrix (Green & Swets, 1966).

Tabelle 4 Mögliche Zustände der SET (Green & Swets, 1966).

	Eigenschaft vorhanden	Eigenschaft nicht vorhanden
Urteil positiv	True positive	False positive
Urteil negativ	False negative	True negative
	$\sum = 100\%$	$\sum = 100\%$

Bei der Durchführung von Beherrschbarkeitsbewertungen ist es, im Einklang mit aktueller Literatur (RESPONSE3 CoP), nützlich, den Lösungsraum auf die Zustände „beherrschbar“ und „nicht beherrschbar“ bezüglich der zu untersuchenden Beherrschbarkeitsklasse (C0 bis C2) einzuschränken. Dann bezeichnet „nicht beherrschbar“ Systeme, die nicht das gewählte Beherrschbarkeitsniveau erreichen, und „beherrschbar“ Systeme, deren Beherrschbarkeit besser als der Grenzwert für die betrachtete Klasse ist. In dieser Form kann die Signalentdeckungstheorie auf Beherrschbarkeitsfragestellungen angewendet werden. Dann ergibt sich die folgende Tabelle der möglichen Zustände:

Tabelle 5 Mögliche Zustände der Beherrschbarkeitsbewertung..

	System unbeherrschbar	System beherrschbar
Ablehnung	Korrekte Ablehnung	Fälschliche Ablehnung
Freigabe	Fälschliche Freigabe	Korrekte Freigabe
	$\sum = 100\%$	$\sum = 100\%$

Den Feldern der Tabelle 5 können nun qualitativ Konsequenzen zugeordnet werden. Bei einer korrekten Freigabe eines Systems entstehen keine weiteren Aufwände für eine Entwicklung alternativer Lösungen oder weiterer Beherrschbarkeitsbewertungen. Das

beurteilte System verursacht beim Einsatz keine inakzeptablen Risiken. Bei einer korrekten Ablehnung eines Systems ist eine Weiterentwicklung notwendig, um eine Systemausprägung mit einem geringeren Risiko zu finden. Der Einsatz eines unter Sicherheitsaspekten unzulässigen Systems wurde jedoch verhindert.

Bei einer fälschlichen Ablehnung werden Aufwände für eine Weiterentwicklung oder Neubewertung des Systems ausgelöst, obwohl dies aufgrund der tatsächlich bestehenden Risiken nicht notwendig wäre. Es werden jedoch keine unzulässigen Risiken für die Anwender des Systems ausgelöst. Bei einer fälschlichen Freigabe eines Systems entstehen zunächst keine weiteren Aufwände für die Bewertung oder Weiterentwicklung. Jedoch verursacht das System inakzeptable Risiken, die zu Personenschäden führen können oder eine spätere Korrektur des Systems notwendig machen. Diese werden, wenn sie später entdeckt und dem System zugeordnet werden, auch eine Weiterentwicklung des Systems notwendig machen.

Bereits auf Basis dieser qualitativen Bewertung lässt sich erkennen, dass die Utilität der verschiedenen Szenarien äußerst ungleich verteilt ist. Bei den korrekten Bewertungen entstehen aus den Bewertungen selbst keine zusätzlichen Aufwände. Im Rahmen eines Entwicklungsprozesses von der Konzeptidee bis zum fertigen Produkt kann die Utilität also mit 0 aufgefasst werden, da der Gesamtaufwand durch diese Ausgänge nicht verändert wird. Bei einer fälschlichen Ablehnung fällt für die Neubewertung oder Weiterentwicklung des Systems ein moderater zusätzlicher Aufwand an. Die Konsequenzen einer fälschlichen Freigabe eines Systems überschreiten diese jedoch bei weitem. Die schuldhafte Verursachung von Personenschäden ist aus ethischer sowie finanzieller Sicht äußerst kritisch zu sehen. Auch spätere Korrekturen an einem fälschlicherweise in den Markt eingeführten System sind bezüglich der Utilität sehr negativ zu bewerten.

Die Bewertungsmatrix lässt sich auch mittels der Spieltheorie beschreiben, um die Eigenschaften des Bewerters der Beherrschbarkeit zu beschreiben. So kann die Entscheidungstabelle als Normalformdarstellung eines Zwei-Personen-Spiels mit vollständigen Informationen aufgefasst werden, bei dem die Utilitäten die „Gewinne“ des Bewertenden beschreiben. In diesem Fall ist erkennbar, dass es sich für den Bewerter nicht um ein Nullsummenspiel handelt, da alle „Gewinne“ entweder 0 oder negativ sind.

In dieser Form lassen sich spieltheoretische Ansätze für dieses Problem anwenden. Wenn der Bewerter als ideal angenommen wird, er also die Beherrschbarkeit des Systems immer korrekt beurteilt, ist die ideale Strategie trivial. Zunächst beobachtet der

Urteiler, ob das System beherrschbar oder nichtbeherrschbar ist und urteilt dann entsprechend, um die mit jedem Ausgang verbundenen Kosten zu minimieren.

Nun wird angenommen, dass die tatsächliche Beherrschbarkeit des zu betrachtenden Systems in keiner Weise bekannt ist. Im spieltheoretischen Sinn bedeutet dies, dass die Natur als Gegenspieler des Bewertenden eine sogenannte gemischte Strategie mit unbekannter Verteilung spielt. Der zu erwartende Gewinn bei der Auswahl „Ablehnung" ergibt sich dann aus zwei Komponenten. Wenn sich das System in der Natur als tatsächlich unbeherrschbar herausstellt, entstehen keine durch das Urteil verursachten zusätzlichen Kosten. Wenn das System tatsächlich beherrschbar ist entstehen zusätzliche Kosten, die mit der Neubewertung des Systems verbunden sind. Dies lässt sich wie folgt notieren:

$$U_{Ablehnung} = p * 0 - (1 - p) * NB$$

Wobei *NB* die Kosten für eine Neubewertung des Systems bezeichnet und *p* die unbekannte Wahrscheinlichkeit ist, dass das System tatsächlich unbeherrschbar ist.

Für die Auswahl der Freigabe des Systems ergibt sich ein ähnlicher Zusammenhang. Ist das System tatsächlich beherrschbar, gibt es keine zusätzlichen Kosten, andernfalls treten Kosten auf, die mit eventuellen Personenschäden verbunden sind. Dies kann wie folgt notiert werden:

$$U_{Freigabe} = p * PS + 0 * (1 - p)$$

Wobei *PS* die Kosten für eventuell eintretende Personenschäden bezeichnet.

Da bei praxisrelevanten Problemen p > 0 ist und *PS* >> *NB*, folgt damit, dass für einen Bewerter, der über die Wahrscheinlichkeit der Nicht-Beherrschbarkeit nur weiß, dass sie ungleich Null ist, also keinerlei Expertise in der Bewertung der Beherrschbarkeit hat, die Ablehnung die ideale Strategie ist.

Dies ist die Grundlage dafür, dass neue Fahrerassistenzsysteme grundsätzlich als Nicht-Beherrschbar gelten müssen, bis ihre Beherrschbarkeit bewiesen ist. Dieser Grundgedanke ist auch im Response 3 Code of Practice und der ISO 26262 verankert.

Beim Nachweis der Beherrschbarkeit von Fahrerassistenzsystemen kann jedoch nicht davon ausgegangen werden, dass diese zweifelsfrei bestimmt werden kann. Der nächste Abschnitt beschäftigt sich mit dem Fall, dass eine Wahrscheinlichkeit für einen Irrtum bei der Beherrschbarkeitsbewertung berücksichtigt werden muss.

2.12.1 Strategie bei begrenzter Expertise

In diesem Fall wird angenommen, dass das System mit der Wahrscheinlichkeit *p* nicht beherrschbar ist und der Bewerter ein nicht-beherrschbares System mit der Wahrscheinlichkeit *q* als solches identifiziert. Dann ist das System mit der Wahrscheinlichkeit *1-p* beherrschbar. Es wird angenommen, dass der Bewerter diesen Zustand mit der Wahrscheinlichkeit *r* korrekt beurteilt. Ein idealer Bewerter hat dann die Eigenschaft $q = r = 1$ und ein Bewerter ohne jegliche Expertise die Eigenschaft $q + r = 1$. In diesem Fall heißt q die Richtig-Positive-Rate oder Sensitivität und r die Richtig-Negative-Rate oder Spezifität. Die Wahrscheinlichkeiten *q* und *r* müssen nicht identisch sein. Ein Bewerter ohne Expertise, der beispielsweise in 90% aller Fälle das Urteil „nicht beherrschbar“ abgibt, erzielt ein *q* von 90%, jedoch nur ein *r* von 10%.

Für bekannte *q* und *r* lassen sich die sog. Likelihood-Ratios *LR* bestimmen:

$$LR_{+} = \frac{q}{1-r}$$

$$LR_{-} = \frac{1-q}{r}$$

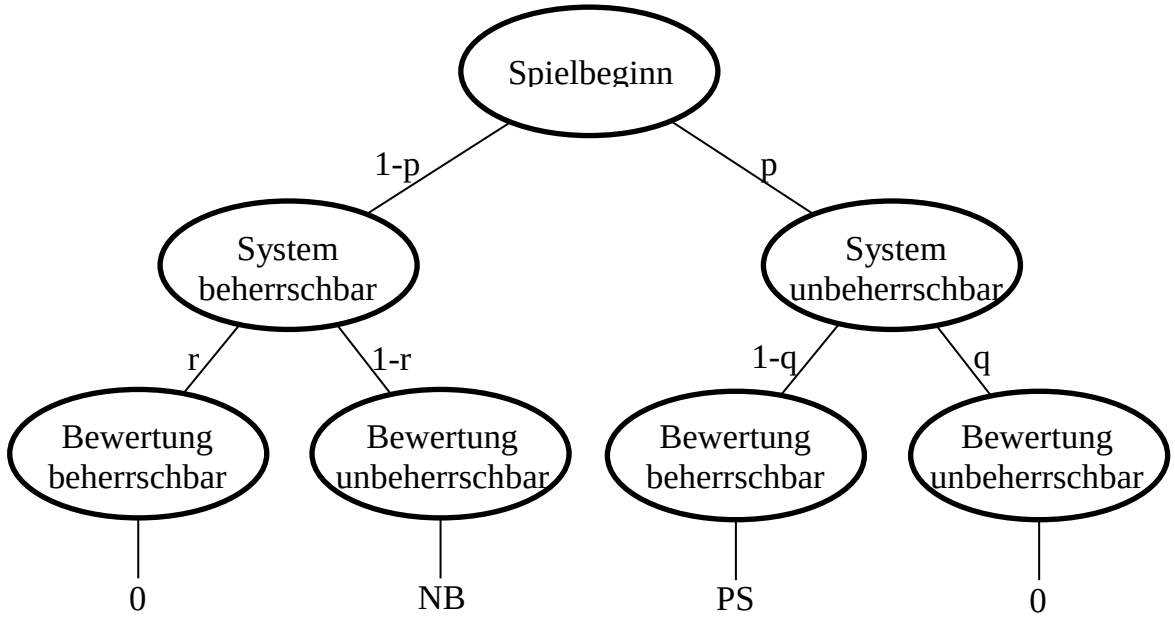

Abbildung 4 Extensivformdarstellung der Beherrschbarkeitsbewertung mit Erwartungswert der Auszahlung für den Bewerter.

Abbildung 4 stellt die Bewertung eines unbekannten Systems durch einen Bewerter mit begrenzter Expertise in Extensivform dar. Daraus ergibt sich für den Bewerter mit begrenzter Expertise folgende Auszahlung für eine beliebige Beherrschbarkeitsbewertung:

$$U = (1-p) * r * 0 + (1-p) * (1-r) * NB + p * (1-q) * PS + p * q * 0$$
$$U = (1-p) * (1-r) * NB + p * (1-q) * PS$$

Dabei ist *p* eine unbekannte Konstante und *r* und *q* beschreiben zusammen das Ausmaß der Expertise und das Urteilsverhalten. Da die Expertise insgesamt begrenzt ist, müssen Randbedingungen für *q* und *r* festgelegt werden. Der Zusammenhang zwischen *q* und *r* ist jedoch eine Eigenschaft des Bewerters und kann nicht a-priori festgelegt werden. An dieser Stelle wird deswegen nur angenommen, dass es einen eindeutigen Zusammenhang zwischen der wahren Ablehnungsrate und der wahren Annahmerate besteht:

$$q = f(r), f\ monoton\ fallend$$

Die Funktion *f* muss monoton fallend sein, da große *q* mit kleinen *r* und kleine *q* mit großen *r* verbunden sein müssen. Die Eigenschaft der Monotonie ergibt sich aus der Annahme, dass der Bewerter das Potential der eigenen Bewertungsmethode zumindest annähernd ausschöpfen soll.

Mit dieser Annahme folgt eine Bedingung für die Ableitung der Ausschüttung des Bewerters mit begrenzter Expertise:

$$f'(r) = -\frac{(1-p) * NB}{p * PS}$$

Dies lässt sich so interpretieren, dass ein Bewerter, der häufig nicht beherrschbare Systeme bewertet (großes *p*), ein Nützlichkeitsoptimum bei einer wenig negativen Ableitung der *q*-*r*-Funktion findet, das heißt weit verschoben zu kleinen *r*, also geringen wahren Annahmerate und einer geringen falschen Ablehnungsrate. Ein Bewerter, der häufig beherrschbare Systeme bewertet, eine Methode mit hohen Neubewertungs-/Weiterentwicklungskosten verwendet oder ein System mit geringen zu erwartenden Personenschäden bewertet, findet ein Optimum bei stark negativen Ableitungen der *q*-*r*-Funktion, das heißt bei vergleichsweise hohen *r*, also einer hohen wahren Annahmerate und einer hohen falschen Ablehnungsrate.

2.12.2 Direkte Messung der Urteilsgüte von Experten

Ein System hat die dichotome Eigenschaft, auf dem geforderten Niveau beherrschbar oder nicht beherrschbar zu sein. Diese kann jedoch vor dem Ende des Produktlebenszyklus des Systems nicht abschließend beobachtet werden. Sie kann vor Ende der Entwicklung des Systems nur indirekt durch den Probandenversuch bestimmt werden. Die Wahrscheinlichkeit, dass ein System, das in 15% der Fälle zu einem Unfall führt, also gerade an der Grenze zwischen C2 und C3 liegt, durch einen Probandenversuch als beherrschbar bestimmt wird, beträgt nach Code of Practice $B(0|20; 0{,}15) \approx 3{,}88\%$. Um für einen Experten nachzuweisen, dass sein Urteil besser oder schlechter ist, muss gezeigt werden, dass seine Fehleinschätzungsrate geringer ist als diese. Die dafür notwendige Stichprobe kann mittels der inversen Betafunktion abgeschätzt werden. Dazu wird die obere Grenze des Intervalls für die Wahrscheinlichkeit der Einschätzung als nichtbeherrschbares System bestimmt (Krengel, 2005).

$$p_o = \beta^{-1}\left(1 - \frac{5\%}{2}; 0; 93\right) \approx 3{,}88\%$$

Diese obere Grenze p_o ist, zum 5% Signifikanzniveau, das erste Mal geringer als beim Probandentest, wenn bei 0 Fehleinschätzungen 93 kritische Situationen richtig eingeschätzt worden sind. Abgesehen von der Schwierigkeit, eine Situation zu identifizieren, deren Beherrschbarkeit bewiesenermaßen gerade 15% beträgt, wird es weiterhin nichttrivial sein, 93 Bewertungen davon durchführen zu lassen, ohne Lerneffekte kompensieren zu müssen. Ein direkter Nachweis, dass Experten oder eine Gruppe von Experten eine geringere Wahrscheinlichkeit von Fehleinschätzungen bei der Beherrschbarkeitsbewertung von kritischen Fahrsituationen haben, erscheint deswegen schwierig.

3 Konkretisierung der Forschungsfragen

In dem vorhergehenden Kapitel wurde der relevante Stand der Technik für die Themengebiete Fahrerassistenzsysteme, Gebrauchssicherheitsbewertung hiervon und Expertiseforschung dargestellt. In diesem Kapitel werden die wissenschaftlichen Fragestellungen aus dem erläuterten Stand der Technik abgeleitet und die dahinterstehende Motivation erklärt.

Es wurde festgestellt, dass durch die wachsende Anzahl und Komplexität an Fahrerassistenzsystemen in modernen PKW ein großer Bedarf an Methoden besteht, um den Aufwand der Beherrschbarkeitsbewertung bei der Gebrauchs- und Funktionssicherheitsprüfung von Fahrerassistenzsystemen zu reduzieren.

Zu diesem Zweck wurden bereits probabilistische Methoden entwickelt, um die objektive Beherrschbarkeit von kritischen Fahrsituationen bereits während der Systemdesignphase abschätzen zu können. Diese Methoden benötigen jedoch eine erhebliche Menge an Wahrscheinlichkeitswerte, für die verschiedenen möglichen Ereignisse während einer kritischen Fahrsituation. Weiterhin müssen die Ergebnisse für jede neu zu betrachtende Situation zunächst validiert werden, um Fehlabschätzungen zu vermeiden. Das Verhältnis von Nutzen zu Aufwand ist bei diesen Methoden derzeit dann vertretbar, wenn die Anforderungen an die Genauigkeit der Methode gering sind, der Aufwand einer herkömmlichen empirischen Beherrschbarkeitsbestimmung nicht vertretbar ist, beispielsweise wenn sehr niedrige Unfallraten nachgewiesen werden müssen, oder wenn die zu betrachtende Situation so weit eingeschränkt werden kann, dass die Bildung eines validen Modells mit geringem Aufwand möglich ist.

Es besteht also Bedarf an einer Methode, die es erlaubt, bereits früh im Entwicklungsprozess von Fahrerassistenzsystemen eine Abschätzung der Beherrschbarkeit treffen zu können, um besonders kritische Szenarien für detailliertere Betrachtungen priorisieren zu können. Weiterhin besteht ein Bedarf an Methoden, die Ergebnisse von empirischen Beherrschbarkeitsstudien frühzeitig abschätzen zu können, um misslungene Nachweise zu vermeiden, die unnötige Anpassungen des Funktionskonzepts und erneute Beherrschbarkeitsstudien nach sich ziehen würden.

Probabilistische Methoden können eine Abschätzung der objektiven Beherrschbarkeit liefern, erfordern jedoch einen hohen a-priori Aufwand und müssen bei jeder Anpassung neu validiert werden. Das subjektive Empfinden der Störung in der kritischen Fahrsituation kann auch hiermit nicht erfasst werden. Da diese Größe jedoch ebenfalls zu einem Scheitern von Beherrschbarkeitsstudien führen kann, ist eine Vorhersage der Störung wünschenswert.

Expertenbewertungen können prinzipiell dafür eingesetzt werden, eine Abschätzung sowohl der objektiven Beherrschbarkeit als auch der subjektiven Störung zu bestimmen. Andererseits ist die Methode für diesen Einsatzzweck noch nicht validiert worden. Es gibt möglicherweise verschiedene nichtrationale Einflussfaktoren auf die Urteile von Experten, die eine unüberlegte Anwendung der Methode nicht sinnvoll erscheinen lassen. In dieser Dissertation soll also untersucht werden, in welchem Rahmen Expertenbewertungen bei der Beherrschbarkeitsbewertung von Fahrerassistenzsystemen angewendet werden können. Hierbei sei explizit die Schätzung von Grundwahrscheinlichkeiten für probabilistische Methoden durch Experten ausgeschlossen, da für diese Aufgabe bereits ein ausreichender Stand der Technik vorliegt.

Aus dieser Hauptaufgabe ergeben sich direkt weitere Fragestellungen. Sofern Expertenbewertungen einen Beitrag zur Beherrschbarkeitsuntersuchung von FAS beitragen sollen, stellt sich unmittelbar die Frage, in welcher Versuchsumgebung die Bewertung zu erfolgen hat. In Frage kommende Varianten sind etwa Urteile auf Basis von Unterlagen in Textform, Versuche im statischen oder dynamischen Fahrsimulator oder in Versuchsfahrten im Realfahrzeug. Eine Beantwortung der Hauptaufgabe muss auch klären, welche Versuchsumgebung für die Bewertung geeignet ist.

Weiterhin ergibt sich die Fragestellung, anhand welcher Kriterien Experten ausgewählt werden können und wie der Nachweis über ihren Status zu führen ist. Dabei ist besonderes Augenmerk auf die verschiedenen Definitionen des Begriffs „Experte“ zu legen.

Urteile von Experten unterliegen nachweislich einer Vielzahl von Faktoren, die nicht zum Untersuchungsgegenstand gehören. Dies hat, wenn nicht alle Randbedingungen sorgfältig kontrolliert werden, eine Varianz der Urteile zur Folge, die die Reliabilität der Urteile vermindert. Die Reliabilität der Urteile der Experten muss also bestimmt werden.

Untersuchungen von Experten sind immer auch Untersuchungen an Individuen. Bevor mehrere Experten zu einem Kollektiv zusammengefasst werden können, muss zunächst

untersucht werden, welche Eigenschaften die individuellen Experten haben. Das Urteilsverhalten der Experten muss also quantifiziert werden. Aus dem quantifizierten Urteilsverhalten können dann die Unterschiede zwischen den Experten hergeleitet werden. Hängen die Urteile der Experten bei gleichem Untersuchungsgegenstand nicht miteinander zusammen, hängt das Ergebnis der Studie nicht vom Untersuchungsgegenstand, sondern von der befragten Person ab. Dies hätte eine Verringerung der Objektivität der Untersuchung zur Folge. Es wird also notwendig sein, die Unterschiede im Urteilsverhalten zwischen den Experten zu bestimmen.

Wenn eine Abschätzung der subjektiv empfundenen Störung der Probanden in kritischen Fahrsituation durch Experten erfolgen soll, muss dafür auch geprüft werden, in welchem Maß die Urteile der Experten die Probandenurteile vorhersagen. In diesem Zusammenhang muss die Validität der Expertenurteile untersucht werden.

Für einzelne Untersuchungen wird in der Praxis nur eine begrenzte Anzahl von Experten zur Verfügung stehen. Deswegen muss untersucht werden, welchen Einfluss die Anzahl der Experten in einer Studie auf die Qualität der Urteile hat.

Zusammenfassend soll anhand der Ergebnisse dieser Arbeit erkennbar sein, ob und in welchem Maße Expertenbewertungen die Beherrschbarkeitsbewertung von Fahrerassistenzsystemen unterstützen können, welche Voraussetzungen dafür erfüllt sein müssen, welche Methoden sich dafür eignen und worin die Einschränkungen der Methode liegen. Eine abschließende Validierung der Methode Expertenbewertung für die Beherrschbarkeit von Fahrerassistenzsystemen liegt außerhalb des Fokus dieser Arbeit.

4 Empirische Erhebung

Dieses Kapitel beschäftigt sich mit der Beschreibung der Methode und den Ergebnissen der empirischen Untersuchungen, die zur Beantwortung der Forschungsfragen durchgeführt wurden. Zunächst wird ein Vorversuch beschrieben, aus dessen Ergebnissen die weitere Methode abgeleitet wird. Anschließend werden der Aufbau und die Resultate von vier verschiedenen Fahrsimulatorstudien beschrieben.

4.1 Vorversuch

Bei der Bewertung von kritischen Fahrsituationen sind verschiedene Versuchungsumgebungen möglich. So können Situationen in einfacher Textform, im statischen oder dynamischen Fahrsimulator, im Realfahrzeug auf der Teststrecke oder im tatsächlichen Verkehr erlebt und bewertet werden. Für die Untersuchung der Forschungsfragen kommen Versuche im Fahrsimulator oder im Realfahrzeug auf der Versuchsstrecke in Frage, da diese ein geeignetes Verhältnis aus externer Validität, Reproduzierbarkeit und Aufwand bieten. Um eine Auswahl zu treffen, wurde eine Vorstudie durchgeführt, auf deren Grundlage die weitere Methode entschieden wird. Das Vorgehen und die Ergebnisse dieser Studie werden in den folgenden Unterkapiteln dargestellt.

4.1.1 Methode

Das Ziel der Vorstudie ist es, einen Methodenvergleich zwischen den Versuchsumgebungen Realfahrzeug, dynamischer Fahrsimulator und statischer Fahrsimulator für die Expertenbewertung von Fahrerassistenzsystemen durchzuführen. Daraus ergibt sich direkt die erste unabhängige Variable: Das Versuchssetting mit den genannten drei Ausprägungen. Um die Eignung der einzelnen Settings für Expertenbewertungen beurteilen zu können, soll als Referenz der Effekt des Versuchssettings auf naive Teilnehmer aufgenommen werden. Sofern es keine Unterschiede des Effekts der Versuchsumgebung auf Teilnehmer mit hoher und geringer Expertise gibt, können die Settings als austauschbar behandelt werden. Daraus ergibt sich die zweite unabhängige Variable: Das Expertiseniveau.

Die abhängige Variable ist gemäß der Problemstellung die subjektive Bewertung der Kritikalität von Fahrsituationen. In diesem Versuch sowie in den folgenden Studien wird

die Kritikalitätsskala aus (Neukum et al., 2008) verwendet, die in Abbildung 5 dargestellt wird. Die Instruktion der Skala entspricht der aus Abschnitt 2.11.

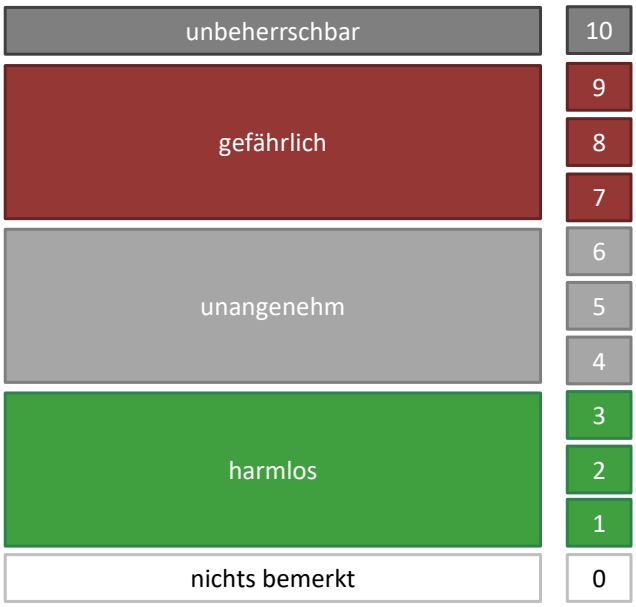

Abbildung 5 Skala zur Bewertung der Kritikalität von Fahr- und Verkehrssituationen (Neukum et al., 2008).

Das Expertiseniveau wird in den Stufen „gering“ und „hoch“ durch verschiedene Rekrutierungsstrategien variiert, da kein geeignetes Vorgehen zur Manipulation der Expertise im hier verwendeten Sinne existiert. Deswegen stellt diese Untersuchung ein Quasi-Experiment dar. Die Definition von „geringer“ und „hoher“ Expertise ist hier im Verhältnis zur jeweils anderen Gruppe zu verstehen. Die Teilnehmer der Gruppe mit geringer Expertise werden aus einem Probandenpool ohne erweiterte Erfahrung mit Fahrerassistenzsystemen ausgewählt. Insbesondere darf keiner der Teilnehmer dieser Gruppe im Rahmen seiner beruflichen Arbeit mit Fahrerassistenzsystemen beschäftigt sein. Außerdem müssen Teilnehmer dieser Gruppe laut Selbstauskunft nur geringes Vorwissen zu FAS vorweisen. Damit ist die Definition eines Normalfahrers nach Ullmann (2006) erfüllt. Die Teilnehmer der Gruppe mit höherer Expertise werden aus dem Umfeld der Beherrschbarkeitsbewertung für Fahrerassistenzsysteme ausgewählt und müssen alle in ihrer beruflichen Praxis bereits an Beherrschbarkeitsbewertungen teilgenommen haben. Mit diesen Einschränkungen konnten 7 Teilnehmer mit höherer Expertise und 6 Teilnehmer mit geringerer Expertise rekrutiert werden, die für Versuche in allen drei Versuchssettings zur Verfügung standen. Die Verfügbarkeit eines geeigneten Realfahrzeugs stellte für diese Vorstudie eine obere Grenze für die Gesamtanzahl an Teilnehmern dar.

Der Bewertungsgegenstand, die kritische Fahrsituation, wurde aus versuchspraktischen Überlegungen heraus konstant gehalten. Da die Bewertung in drei verschiedenen Versuchsumgebungen repliziert werden sollte, wurde hier auf die Darstellung verschiedener Fahrsituationen verzichtet und lediglich eine Situation mit verschiedenen Parametriersierungen verwendet. Insbesondere wegen der Darstellbarkeit im Realfahrzeug wurde als kritisches Szenario ein fehlerhaftes überlagertes Lenkmoment bei manueller Fahrt auf gerader Strecke gewählt. Das überlagerte Lenkmoment konnte in der Amplitude und dem Gradienten verändert werden, sodass 10 verschiedene Eingriffe dargestellt wurden. Jeder Eingriff wurde für insgesamt 3 Sekunden gehalten und war nicht übersteuerbar. Die Fehleingriffe wurden jeweils drei Mal dargestellt, sodass jeder Teilnehmer in jedem Versuchssetting 30 Bewertungen abgegeben hat. Tabelle 6 stellt dar, wie die verschiedenen Lenkmomente durch Variation des Gradienten und der Amplitude erzeugt wurden.

Tabelle 6 Fehllenkmomente im Vorversuch.

Nummer	Gradient [Stufe 1-3]	Amplitude [Stufe 1-4]
1	1	1
2	2	1
3	3	1
4	2	2
5	3	2
6	2	3
7	3	3
8	1	4
9	2	4
10	3	4

Das Versuchsdesign wird in Tabelle 7 dargestellt. Dabei liegen in jeder Versuchsbedingung 30 Bewertungen je Teilnehmer vor, sodass insgesamt 1170 Urteile für die Analyse zur Verfügung stehen (*30 Urteile/Bedingung/Teilnehmer * (6 Novizen + 7 Experten) * 3 Versuchsbedingungen = 1170 Urteile*).

Tabelle 7 Darstellung des Versuchsdesigns des Vorversuchs.

		Versuchssetting		
		Statisch	Dynamisch	Real
Expertise	Gering	n = 6	n = 6	n = 6
	Hoch	n = 7	n = 7	n = 7

4.1.2 Ergebnisse

Der zuvor diskutierte Versuchsaufbau führte zu den Mittelwerten der Störungsbewertungen für die jeweiligen Lenkmomente, die in Abbildung 6 dargestellt werden.

Abbildung 6 Balkendiagramm der Mittelwerte je Gruppe und Lenkmoment im Vorversuch.

Jeder Balken der Abbildung stellt den Mittelwert über alle Urteile jeder Gruppe dar. Für Experten sind dies 7 Teilnehmer bei drei Wiederholungen, also der Mittelwert aus 21 Urteilen je Balken. Bei Novizen sind dies 6 bei ebenfalls drei Wiederholungen, also der Mittelwert aus 18 Urteilen je Balken.

Eine Varianzanalyse zeigt einen hochsignifikanten Effekt der Lenkmomentenamplitude auf die Urteile auf der SBS ($F = 220{,}7$; $p < 0{,}001$). Auch für den Gradienten wurde ein hochsignifikanter Effekt gefunden ($F = 22{,}3$; $p < 0{,}001$). Für das Expertiseniveau wurde jedoch kein signifikanter Effekt gemessen ($F = 3{,}2$; $p = 0{,}074$).

Qualitativ lässt sich hier bereits erkennen, dass die Urteile der Experten und Novizen nicht systematisch auseinander zu liegen scheinen. So ergibt sich eine Pearson-Korrelation von $\rho = 0{,}90$ ($p < 0{,}01$). Eine Varianzanalyse mit Messwiederholung zeigt auch keinen Effekt des Versuchssettings auf die Urteile ($F = 0{,}42$; $p = 0{,}658$). Die Pearson-Korrelation zeigt jedoch einen hochsignifikanten Zusammenhang zwischen den Urteilen in den einzelnen Settings. Die gefundenen Korrelationen können Tabelle 8 entnommen werden.

Tabelle 8 Korrelationen zwischen den Versuchssettings im Vorversuch.

Zusammenhang	p-Wert	ρ
Dynamischer Simulator – Statischer Simulator	< 0.01	0.680
Dynamischer Simulator – Realfahrzeug	< 0.01	0.551
Statischer Simulator – Realfahrzeug	< 0.01	0.627

4.1.3 Diskussion

Der Vorversuch hatte den Zweck, eine Entscheidungsgrundlage für die Auswahl des Versuchssettings für weitere Studien zu bilden. Das Versuchsdesign hat erfolgreich hochsignifikante Effekte für die Amplitude- und Gradientenstufe erzeugt, jedoch keinen signifikanten Unterschied zwischen den Urteilen der Experten und Novizen gefunden. Tatsächlich wurde eine starke Korrelation zwischen den Urteilen der beiden Probandengruppen gefunden.

Mit dem gewählten Versuchsdesign wurde kein signifikanter Effekt für das Versuchssetting identifiziert. Es konnte festgestellt werden, dass die Urteile in den verschiedenen Settings miteinander korrelieren und dieser Zusammenhang wurde quantifiziert. Der Korrelationskoeffizient liegt bei allen drei Vergleichen in einer ähnlichen Größenordnung. Dies könnte eine Eigenschaft der betrachteten Simulatoren sein, da bei anderen Simulatorsystemen ebenfalls für überlagerte Lenkeingriffe ein signifikanter Unterschied zwischen Realfahrzeug und Simulator gefunden wurde, der jedoch von der Lenkwinkelübersetzung des verwendeten Simulators abhängig war (Neukum, 2009).

4.2 Versuchskonzept

Die Forschungsfrage fordert einen Vergleich der subjektiven Störungsbewertung zwischen Teilnehmern mit geringer und hoher Expertise in der Beherrschbarkeit von Fahrerassistenzsystemen. Zu diesem Zweck müssen diese Versuchsgruppen möglichst identische Fahrsituationen unterschiedlicher Kritikalität erleben und bewerten können. Dabei hat die Kontrolle von Störvariablen eine entscheidende Rolle, da alle nicht kontrollierten Störvariablen eine Varianz der Urteile erzeugen, die den zu untersuchenden Zusammenhang überdeckt. Bei der Auswahl des Versuchssettings ist also auf eine hohe Reproduzierbarkeit zu achten. Dies spricht prinzipiell für eine Durchführung der weiteren Studien im Fahrsimulator. Da weiterhin Fahrsituationen mit variierender Kritikalität erlebt werden sollen, um das gesamte Spektrum der Störungsbewertung abzudecken, ist

auch unter Sicherheitsaspekten eine Durchführung der Studien im Fahrsimulator zu bevorzugen. Schließlich ist es wünschenswert, eine große Bandbreite an Situationen für die folgenden Studien zu implementieren, damit die externe Validität der Ergebnisse nicht zu stark eingeschränkt ist. Da der Aufwand bei der Implementierung vieler Fahrsituationen – speziell solcher mit Interaktion mit dem Fremdverkehr – im Realfahrzeug mit einem erheblich größeren Aufwand als im Fahrsimulator verbunden ist, spricht auch dieser Aspekt für eine Durchführung der weiteren Studien im Fahrsimulator.

Weiterhin fordert die Forschungsfrage, dass eventuell existierende Unterschiede im Urteilsverhalten zwischen den einzelnen Experten zu identifizieren sind. Diese Forderung macht ein within-subject Design notwendig und führt dazu, dass die notwendige Teststärke nur durch eine ausreichend hohe Anzahl an geprüften Fahrsituationen zu Stande kommen kann. Aussagen, die nur auf einer geringen Anzahl an Urteilen basieren, leiden unter hoher Varianz bzw. geringer Teststärke. Dies ist gerade bei subjektiven Urteilen zu beachten, da hier eine Reihe von unkontrollierbaren Faktoren die Ergebnisse beeinflussen können. Die in weiteren Versuchen zu erwartende Varianz lässt sich mittels der Daten des Vorversuchs abschätzen. Hier wird eine Standardabweichung von 1,94 Skaleneinheiten angenommen, die die Standardabweichung der Urteile über alle Probanden im Vorversuch darstellt. Es wird festgesetzt, dass Effekte mit einem Cohens *d* von 0,5 identifizierbar sein sollen. Daraus folgt bei einer geschätzten Standardabweichung der Urteile von $s = 1{,}94$ auf der SBS eine zu erwartende Differenz der Mittelwerte von $\mu_D = d * s = 0{,}5 * 1{,}94 = 0{,}97$.

Für eine gewünschte Teststärke von 0,5 ergibt sich bei einem Signifikanzniveau von 5% eine Stichprobengröße von 32 (Kenny, 1987). Diese Anzahl an Urteilen ist für jeden zu betrachtenden Experten notwendig, um eventuell vorhandene Unterschiede zwischen den Experten auch identifizieren zu können.

Aus dieser Berechnung folgt direkt, dass diese Anzahl an Urteilen nicht im Rahmen einer einzelnen Studie durchgeführt werden kann, ohne erhebliche Beeinflussung der Ergebnisse durch Ermüdung der Probanden und nachlassende Motivation in Kauf zu nehmen. Deswegen wurden diese 32 Urteile auf vier Teilstudien aufgeteilt. Um jedoch weiterhin die Anforderung nach hoher Reliabilität zu erfüllen, werden alle vier Studien im gleichen Fahrsimulator durchgeführt und darauf geachtet, dass technische Veränderungen am Fahrsimulator auf ein aus Wartungsgründen notwendiges Minimum reduziert werden. Die folgenden Unterkapitel beschreiben diese Teilstudien im Detail.

4.3 Rekrutierung der Experten

Bei der Untersuchung des Effekts von Expertise gibt es zwei mögliche Herangehensweisen. Entweder wird die Expertise der Teilnehmer durch eine Manipulation beeinflusst oder es werden Probanden mit einem bestimmten a-priori Level an Expertise rekrutiert. Um unabhängige Stichproben zu bekommen ist es sinnvoll, die Manipulation der Expertise einer Zufallsstichprobe zu bevorzugen. Da jedoch keine valide Methode zur Quantifizierung der Beherrschbarkeitsexpertise existiert, gibt es folglich auch keine Methode zur validen Manipulation dieser Expertise. Es wäre zwar möglich solche Methoden zu definieren, aber nicht möglich, ihren Effekt zu validieren.

In Abwesenheit einer Methode zur Manipulation der Beherrschbarkeitsexpertise bleibt nur die Möglichkeit, Studienteilnehmer mit vorhandener a-priori Expertise zu rekrutieren. Hierzu sind Kriterien für die Zugehörigkeit zur Expertengruppe zu definieren. Da wieder keine Methode zur Quantifizierung der Expertise existiert, lassen sich auch diese Kriterien nicht validieren. Dies spricht dafür, wenig strikte Kriterien für die Zugehörigkeit zur Expertengruppe zu formulieren. Dadurch stehen mehr Personen für die Expertengruppe zur Verfügung, sodass auch die Größe der Stichprobe wächst. Damit sinkt das Risiko, Teilnehmer mit tatsächlich hoher Expertise aufgrund fehlerhafter Kriterien auszuschließen. Da am Ende der Versuche eine Quantifizierung der Expertise vorliegen soll, führt die Inklusion von Teilnehmern mit geringer Expertise auch nicht zu einem Nachteil. Im Gegenteil könnten dadurch eher Unterschiede innerhalb der Expertengruppe festgestellt werden, die bei zu strikten Kriterien eventuell nicht existieren.

In Anbetracht praktischer Überlegungen über die zur Verfügung stehenden Personen wurden folgende Kriterien für die Zugehörigkeit zur Expertengruppe definiert:

1. Die Teilnehmer müssen zur Zeit der Untersuchung in der Entwicklung von Fahrerassistenzsystemen beschäftigt sein.

2. Die Teilnehmer müssen vor Beginn des ersten Versuchs bereits mindestens einmal an einer Beherrschbarkeitsbewertung eines Fahrerassistenzsystems teilgenommen haben.

3. Die Teilnehmer müssen über eine erweiterte Fahrsicherheitsausbildung verfügen.

Die ersten zwei Kriterien dienen lediglich zur Absicherung eines Grundverständnisses der Zusammenhänge. Kriterium eins stellt sicher, dass alle Teilnehmer der Expertengruppe ein Grundverständnis von Fahrerassistenzsystemen verfügen. Kriterium zwei

dient zur Sicherstellung, dass die zu rekrutierenden Experten bereits die Begrifflichkeiten der Beherrschbarkeit von Fahrerassistenzsystemen in ihrer Arbeit verwendet haben. Kriterium drei schließlich stellt sicher, dass die Teilnehmer der Expertengruppe über ein vergleichbares Niveau an Fahrfertigkeit verfügen.

Die Rekrutierung von Teilnehmern, die diese Kriterien erfüllen erfolgte über ein Schneeballverfahren. Dazu wurden, beginnend bei der Abteilung Konzepte integrale Sicherheit der BMW AG, Mitarbeiter mündlich gefragt, welche Kollegen ihnen bekannt sind, die die oben genannten Kriterien erfüllen. Die genannten Personen wurden anschließend kontaktiert und gebeten, Selbstauskunft über die Erfüllung der Kriterien abzugeben. Sofern die Selbstauskunft positiv war, wurden die Kandidaten gefragt, ob sie freiwillig an der Versuchsreihe teilnehmen möchten. Die rekrutierten Personen wurden erneut gefragt, ob ihnen weitere Personen bekannt sind, die die Kriterien erfüllen. Der Prozess wurde abgebrochen, als keine zusätzlichen Personen mehr genannt wurden.

Die Teilnehmer wurden mit dieser Methode auf Grundlage von Freiwilligkeit rekrutiert und mussten keinen formellen Test zur Erfüllung der Kriterien bestehen. Dies sind Randbedingungen, die aufgrund des arbeitsrechtlichen Rahmens gewählt worden sind. Aus dem gleichen Grund wurde auf eine Erhebung von Persönlichkeitsmerkmalen verzichtet.

Das gewählte Rekrutierungsverfahren schränkt die rekrutierten Experten auf Mitarbeiter der BMW AG ein. Dies ist eine praktische Einschränkung, die aufgrund des Untersuchungsorts, das Forschungs- und Innovationzentrum (FIZ) der BMW AG in München, ohnehin gegeben ist. Damit erfolgt keine weitere Einschränkung des zulässigen Personenkreises durch die Rekrutierungsmethode.

Da kein validierter Test auf tatsächlich vorhandene Expertise im Bereich Beherrschbarkeitsbewertung existiert, dürfen die so rekrutierten Probanden nur als Nenn-Experten gelten. Die Rekrutierungsmethode stellt kein bekanntes Maß an minimaler Expertise sicher. Ob das Resultat der gewählten Kriterien zu einem Teilnehmerkollektiv führt, dessen Bewertungen der Kritikalität von Fahrsituationen besser oder schlechter für eine Expertenbewertung der Kritikalität geeignet ist, wird im Rahmen der nachfolgenden Versuche untersucht. Den Teilnehmern der Expertengruppe wird im weiteren Verlauf dieser Arbeit aufgrund ihrer Zugehörigkeit zur Expertengruppe eine erhöhte Expertise zugewiesen. Ob dies gerechtfertigt ist, wird erst anhand der Ergebnisse der Auswertung erkennbar sein.

Insgesamt wurden auf diese Weise 21 Teilnehmer für die Gruppe mit höherer Expertise – die Expertengruppe – rekrutiert. Davon waren 20 Teilnehmer männlich und das mittlere Alter lag bei 34,3 Jahren. Die durchschnittliche Fahrleistung betrug 15250 km im Jahr 2013.

4.4 Versuch 1

Die erste Teilstudie wurde im April 2013 durchgeführt. Neben der Sammlung von Daten für den Vergleich des Urteilsverhaltens von Teilnehmern mit geringer und hoher Expertise lag hier auch ein Fokus darauf, erste Erkenntnisse über Unterschiede im Fahrverhalten zwischen den Gruppen zu sammeln. Teile dieses Unterkapitels wurden in Galaske, Farid und Bengler (2015) veröffentlicht.

4.4.1 Theorie

Wenn Experten zur Bewertung der Beherrschbarkeit von Fahrerassistenzsystemen verwendet werden, wird typischerweise davon ausgegangen, dass Experten bessere, also genauere, Urteile abgeben, als es naive Testsubjekte tun würden. Andererseits wäre der Status als Experte falsch zugeordnet worden. Da bei der Bewertung der Beherrschbarkeit eines Fahrerassistenzsystems nur ein Urteil als wahre Wahrheit zugelassen wird, sollte sich hohe Expertise in einer geringen Differenz zu diesem Referenzpunkt zeigen. Folglich ist zu erwarten, dass eine Gruppe von Experten homogenere Urteile über die Kritikalität einer Fahrsituation als eine Vergleichsgruppe mit relativ geringer Expertise im Bereich Beherrschbarkeit von Fahrerassistenzsystemen trifft. Weiterhin ist zu erwarten, dass Experten nur wenig durch Gewöhnungseffekte innerhalb einer Studie beeinflusst werden. Ein idealer Experte der Beherrschbarkeit von Fahrerassistenzsystemen sollte ähnliche Szenarien oder Effekte bereits mehrfach erlebt haben und deswegen wenig durch Überraschung oder Gewöhnung von dem wahren Urteil abgelenkt werden. Bei einer wiederholten Bewertung eines Szenarios ist also davon auszugehen, dass Experten weniger als naive Probanden in ihrem Urteilen schwanken. Schließlich ist zu erwarten, dass ein Experte der Beherrschbarkeit von FAS, also der Sicherheit solcher Systeme, sich der Konsequenzen seines Handelns bewusst ist und deswegen eher als ein naiver Versuchsteilnehmer bestrebt ist, eine Unterbewertung der Kritikalität einer Fahrsituation zu vermeiden.

Die folgende Studie wurde entwickelt, um die im vorangegangenen Absatz beschriebenen Erwartungen zu überprüfen.

4.4.2 Methode

Die erste Studie hatte das Ziel, das Urteilsverhalten zweier Gruppen in einem Versuchssetup zu vergleichen, das an eine Probandenstudie mit naiven Testsubjekten gemäß RESPONSE 3 Code of Practice angelehnt ist. Eine Gruppe Teilnehmer bestand aus 33 Personen aus dem Umfeld des BMW Forschungs- und Innovationszentrums mit geringer oder keiner Erfahrung mit Fahrerassistenzsystemen. Das Durchschnittsalter dieser Gruppe betrug ca. 33 Jahre bei einer Standardabweichung von 12,7 Jahren. 10 % der Gruppenmitglieder waren weiblich. Alle Teilnehmer dieser Gruppe mit geringer Expertise gaben an, in ihrer täglichen Arbeit nicht an der Entwicklung von Fahrerassistenzsystemen beteiligt zu sein. Im Kontext des RESPONSE 3 CoP stellt dies eine geeignete Stichprobe für eine Probandenstudie dar und kann deswegen für einen Vergleich mit Experten herangezogen werden. Die Gruppe mit hoher Expertise bestand aus 19 Mitgliedern der zuvor rekrutierten Expertengruppe. Die Teilnehmer dieser Gruppe waren aus verschiedenen Abteilungen und Disziplinen und gaben jeweils an, in ihrer täglichen Arbeit an der Entwicklung von Fahrerassistenzsystemen beteiligt zu sein. Damit erfüllt die Gruppe mit hoher Expertise die Anforderungen des RESPONSE 3 CoP für eine Expertenbewertung.

Die Studie wurde gemäß Vorstudie im statischen Fahrsimulator durchgeführt. Dabei wird davon ausgegangen, dass die Simulatorumgebung das Urteilsverhalten von Probanden und Experten in gleicher Art und Weise beeinflusst. Der verwendete Fahrsimulator war der statische Fahrsimulator 2 im BMW Forschungs- und Innovationszentrum (FIZ) in München. Die Fahrer erlebten die Fahrsituationen in einem Mockup, dass aus der Vorderhälfte eines BMW 5ers bestand. Die Darstellung der Fahrszene erfolgte über eine 5 Kanal-Projektion auf einer gekrümmten Leinwand, sodass sich ein 210° Blickwinkel ergab. Zusätzlich dienten 3 weitere Kanäle zur Darstellung der Rückspiegel, die über 3 Plasmafernsehr hinter dem Mockup dargestellt wurden, sodass in jedem der drei Spiegel die richtigen Perspektive sichtbar war. Das simulierte Fahrzeug verfügte über eine automatische Gangschaltung. Abbildung 7 stellt das verwendete Mockup und die Projektionsfläche dar. Die Bildschirme für die Darstellung der Rückspiegelinhalte werden der Übersichtlichkeit halber nicht gezeigt.

Abbildung 7 Der statische Fahrsimulator 2 im BMW FIZ.

Alle Versuchsteilnehmer, unabhängig von vorheriger Erfahrung oder Sachexpertise, nahmen an der gleichen Versuchseinführung teil. Jedem Probanden wurden der Versuchszweck, die Störungsbewertungsskala und der Simulator selbst erklärt. Nach der verbalen Instruktion absolvierte jeder Teilnehmer eine 15-minütige Einführungsfahrt, um sich an die Bedienung des simulierten Fahrzeugs zu gewöhnen. Während dem Betrieb des Fahrsimulators waren der Teilnehmer und der Versuchsleiter über eine Gegensprechanlage verbunden.

Der Hauptteil der Studie wurde als nicht-permutiertes gemischtes 2-Faktor-Design dargestellt. Als unabhängige Variablen wurden Expertise und Wiederholung gewählt. Die Expertise wurde durch die Rekrutierung der Teilnehmer mit geringer und hoher Expertise manipuliert, wodurch diese Studie, wie die Vorstudie, ein Quasi-Experiment darstellt. Die geforderte Wiederholung wurde implementiert, indem die ersten vier kritischen Szenarien der Studie nahtlos wiederholt wurden, sodass jeder Proband insgesamt 8 Szenarien erlebt hat. Die vier einzigartigen Szenarien wurden in gleichem Abstand auf einem 20-minütigen Rundkurs verteilt, sodass die Szenarien nach dem Abschluss der ersten Runde ohne Unterbrechung wiederholt werden konnten, indem die Teilnehmer dem Kurs schlicht ein zweites Mal folgten.

Es wurde darauf verzichtet, die Reihenfolge der Szenarien zu permutieren, um die Streuung der Ergebnisse zu minimieren. Eine Permutation wurde als nicht notwendig angesehen, da die Bewertung der wahren Kritikalität nicht das Ziel der Studie war, sondern lediglich ein Vergleich der Urteile zwischen den Gruppen.

Die vier dargestellten Szenarien bestanden aus zwei Szenarien für jeweils zwei verschiedene Arten von Fahrerassistenzsystemen. System A assistierte dem Fahrer bei der longitudinalen und transversalen Steuerung des Fahrzeugs, erforderte jedoch, dass der Fahrer dauerhaft die Hände am Lenkrad beließ. In den kritischen Fahrsituationen für dieses Szenario wurde auf einem geraden Stück Landstraße die Lateralsteuerung durch ein Fehllenkmoment zum rechten Fahrbahnrand hin ersetzt. In dieser kritischen Fahrsituation gab es keinen umgebenden Fremdverkehr. Der Fahrer musste, um die Situation zu beherrschen, das Fehllenkmoment am Lenkrad abstützen und das Fahrzeug zurück in die Spur führen, ohne dabei den rechten oder linken Fahrbahnrand zu überschreiten. System B bestand in einer Funktion, die in etwa hochautomatisiertem Fahren entsprach. Dem Fahrer war es also gestattet, den Verkehr und die Straße vor dem Auto nicht zu beobachten. In den kritischen Szenarien für dieses System wurde ein stehendes Fremdfahrzeug an einer schlecht einsehbaren Stelle platziert. Der Fahrer erhielt etwa sechs Sekunden vor der zu vermeidenden Kollision eine audiovisuelle Warnung. Diese Fahrerübernahmeaufforderung war zuvor in der Einführungsfahrt ohne bevorstehendes Hindernis geübt worden. Während dieser Situation verhinderte dichter Gegenverkehr einen Fahrstreifenwechsel, sodass der Fahrer, um eine Kollision zu vermeiden, nur noch die Möglichkeit hatte, das Fahrzeug durch Betätigung der Bremse rechtzeitig zum Stillstand zu bringen. Im Gegensatz zur strikten Interpretation von hochautomatisierten Fahren war hier also zur Kollisionsvermeidung eine Fahrerreaktion notwendig. Um die Streuung, die durch unterschiedliches Blickverhalten der Probanden während der Phase mit System B verursacht werden könnte, zu minimieren, wurden die Probanden gebeten, eine Nebenaufgabe im zentralen Infotainmentsystem des Mockups zu bedienen. Die verwendete Nebenaufgabe war der Surrogate Reference Task (SuRT). Diese Nebenaufgabe erfordert vom Fahrer, dass er einen Zielkreis aus einer Menge an Ablenkungskreisen identifiziert, und mittels eines Dreh-Drück-Stellers im Mockup eingibt, ob der Zielkreis in der rechten oder linken Hälfte des Bildschirms lokalisiert ist. Die Größe der Distraktoren betrug aus der Fahrerperspektive etwa 26 arcmin und die Größe des Zielkreises betrug ca. 32 arcmin. Die Anzahl der Distraktoren war auf 50 gesetzt. Um die Motivation zur Betätigung der Nebenaufgabe zu standardisieren, wurde für die beste Leistung in der Nebenaufgabe ein kleinerer Preis in Form eines Modellautos ausgeschrieben. Abbildung 8 zeigt einen Screenshot aus der Nebenaufgabe in dem verwendeten Setup. Der Zielkreis befindet sich im rechten oberen Quadranten.

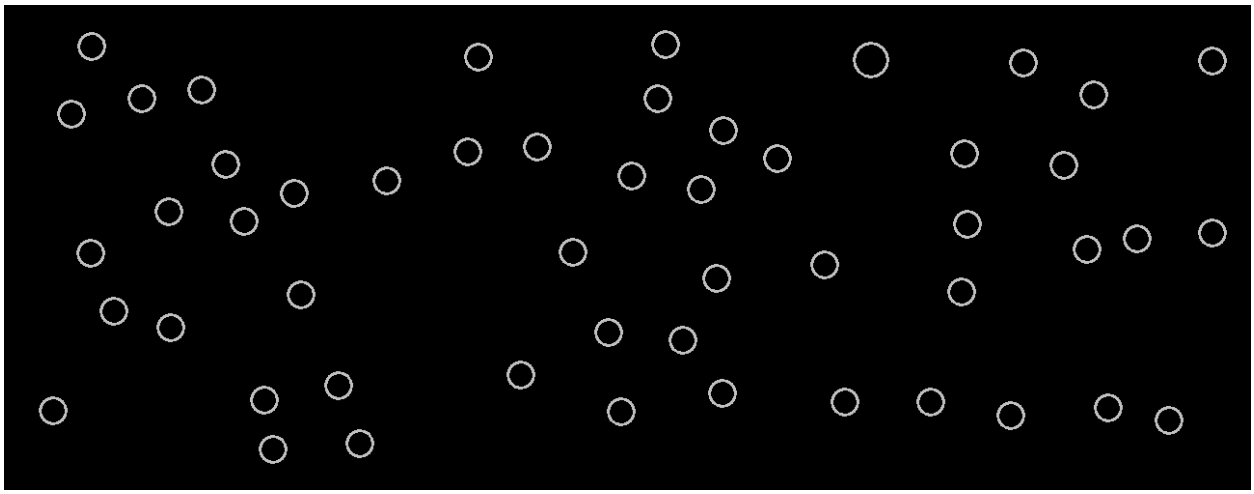

Abbildung 8 Screenshot aus dem verwendeten Surrogate Reference Task in Versuch 1.

Für jedes der zwei Systeme wurden zwei Szenarien mit verschiedener Kritikalität dargestellt. Diese vier Situationen wurden zwei Mal durchfahren, sodass jeder Proband insgesamt 8 Szenarien erlebte. Nach jeder Situation wurde jeder Teilnehmer gebeten, das Fahrzeug anzuhalten und die erlebte Situation auf der Störungsbewertungsskala zu bewerten. Nach der Befragung des Teilnehmers wurde dieser instruiert, dem Streckenverlauf bis zur nächsten Situation weiter zu folgen, bis alle acht Situationen bewertet worden sind. Tabelle 9 zeigt die Reihenfolge, in der die Situationen präsentiert wurden. Darin ist erkennbar, dass die Reihenfolge der ersten vier Situationen identisch mit der zweiten Hälfte der Versuchsfahrt ist.

Tabelle 9 Reihenfolge der dargestellten kritischen Fahrsituationen in Versuch 1.

Nummer	1	2	3	4	5	6	7	8
System	B	A	A	B	B	A	A	B
Kritikalität	Hoch	Niedrig	Hoch	Niedrig	Hoch	Niedrig	Hoch	Niedrig

Mit diesem Versuchsaufbau ist es nun möglich, aus den in 4.4.1 formulierten Erwartungen die zu testenden Hypothesen für den Versuch zu formulieren:

H_1: Die Varianz der Urteile der Gruppe mit hoher Expertise ist je Szenario geringer als bei der Gruppe mit geringer Expertise.

H_2: Der Mittelwert der Urteile der Gruppe mit hoher Expertise ist je Szenario höher als bei der Gruppe mit geringer Expertise.

H_3: Die Hedges-g Effektstärke der Wiederholung ist in der Gruppe mit hoher Expertise je Szenario geringer als in der Gruppe mit geringer Expertise.

4.4.3 Ergebnisse

Mittels des Lillieforstest auf Normalverteilung wurde festgestellt, dass bei der Gruppe mit geringer Expertise für vier von acht dargestellten Szenarien die Hypothese der Normalverteilung abgelehnt werden muss. Bei der Gruppe mit höherer Expertise war dies in zwei von acht Szenarien der Fall. Im Sinne konservativer Annahmen wird deswegen im Folgenden davon ausgegangen, dass die Daten nicht normalverteilt sind. Tabelle 10 enthält die Ergebnisse der Tests auf Normalverteilung. Nicht signifikante Ergebnisse werden als „n.s." gekennzeichnet.

Tabelle 10 Ergebnisse des Lillieforstests auf Normalverteilung in Versuch 1.

Szenario	1	2	3	4	5	6	7	8
Geringe Expertise	$p < 0,05$	n.s.	n.s.	$p < 0,05$	n.s.	$p < 0,05$	n.s.	$p < 0,05$
Hohe Expertise	$p < 0,05$	n.s.	$p < 0,05$	n.s.	n.s.	n.s.	n.s.	n.s.

Anschließend wurde der Levenetest auf Gleichheit der Varianz verwendet, um zu untersuchen, ob die Gruppe mit höherer Expertise tatsächlich Urteile mit geringerer Streuung als die Gruppe mit geringerer Expertise gefällt hat. Auf den Gesamtdatensatz angewendet wird die Hypothese H_1 abgelehnt ($F = 0,357$; $df = 414$; $p > 0,05$). Auf die einzelnen Szenarien angewendet, erreichte der Test in nur einem von acht Szenarien Signifikanz. Es wird also Gleichheit der Varianz der Urteile der beiden Probandengruppen angenommen. Die Ergebnisse des Levenetests für die einzelnen Szenarien werden in Tabelle 11 dargestellt.

Tabelle 11 Ergebnisse des Levenetests auf Gleichheit der Varianz der Urteile zwischen den Gruppen mit geringer und hoher Expertise in Versuch 1.

Szenario	1	2	3	4	5	6	7	8
Ergebnis	$p \sim 0,03$	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.

Aufgrund der Gleichheit der Varianz der Urteile bei beiden Teilnehmergruppen wurde der einseitige Mann-Whitney-Wilcoxon-Rangsummentest auf Gleichheit der Mediane gewählt, um Hypothese H_2 zu testen. Der Test ergab einen signifikanten Unterschied zwischen Urteilen der beiden Gruppen ($Z = 4,27$; $p < 0,001$). Um zu untersuchen, wie sich dieser Unterschied auf die Daten aufteilt, wurde der Test anschließend auf die einzelnen Szenarien angewendet. Tabelle 12 stellt die Ergebnisse dieses Tests sowie die

Hedges-g Effektstärken mit dem dazugehörigen 95% Konfidenzintervall, die numerisch mittels Bootstrapping ermittelt wurden, dar. Eine Bonferroni-Korrektur der p-Werte wird nicht durchgeführt, da es sich nicht um die Prüfung einer alternativen Hypothese auf den gleichen Daten handelt.

Tabelle 12 Vergleich der Urteile der Gruppen mit geringer und hoher Expertise in Versuch 1 für die einzelnen Szenarien.

Szenario	1	2	3	4	5	6	7	8
MWW-Rangsummentest	n.s.	n.s.	p ~ 0,03	n.s.	n.s.	n.s.	p ~ 0,03	p ~ 0,03
Hedges-g Obergrenze	0,88	0,85	1,17	1,32	1,05	0,93	1,25	1,27
Hedges-g Median	0,31	0,26	0,52	0,67	0,41	0,31	0,57	0,61
Hedges-g Untergrenze	-0,18	-0,31	-0,03	0,11	-0,17	-0,25	0,04	0,08

Diese Ergebnisse zeigen, dass die Grupppe der Teilnehmer höherer Expertise Urteile mit einem höheren Median als die Gruppe mit geringerer Expertise abgegeben hat. Die Mediane und Konfidenzintervalle der Hedges-g Effektstärken zeigen jedoch, dass dieser Effekt nicht in allen Szenarien gleich hoch ist. So wird in Szenario 2 eine Effektstärke von 0,26 gemessen und in Szenario 4 eine Effektstärke von 0,67. Die Größe des Konfidenzintervalls der Effektstärke verdeutlicht den Bedarf an großen Stichproben, um diesen Effekt zu untersuchen.

Unter der Annahme normalverteilter Mittelwerte, die aufgrund der Teilnehmeranzahl getroffen werden kann, wird im Folgenden die Wahrscheinlichkeit berechnet, dass der Mittelwert der Urteile der Gruppe höherer Expertise unterhalb des Mittelwerts der Urteile der Gruppe mit geringerer Expertise liegt. Diese Ergebnisse werden in Tabelle 13 für jedes der acht Szenarien dargestellt.

Tabelle 13 Wahrscheinlichkeit, dass der Mittelwert der Urteile der Gruppe höherer Expertise geringer liegt als der Mittelwert der Urteile der Gruppe mit geringerer Expertise.

Szenario	1	2	3	4	5	6	7	8
Ergebnis	11%	18%	3%	1%	8%	14%	17%	14%

Um die dritte Hypothese über den Effekt der Wiederholung auf die beiden Gruppen zu prüfen, wurden die Hedges-g Effektstärken und ihre 95% Konfidenzintervalle für die vier wiederholten Szenarien berechnet. Diese werden in Abbildung 9 dargestellt.

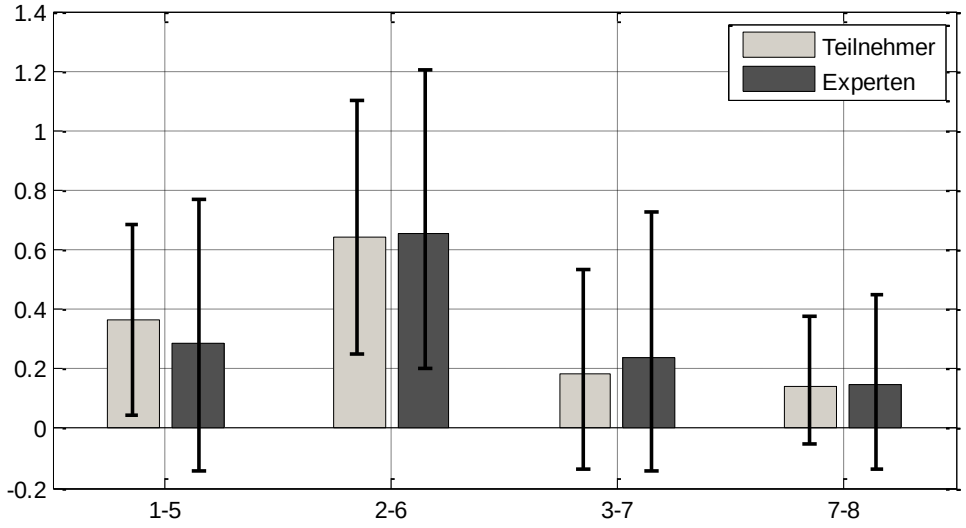

Abbildung 9 Hedges-g Effektstärke der Wiederholung für die vier wiederholten Szenarien für die Gruppe mit hoher und geringer Expertise in Versuch 1.

Das Ergebnis in Abbildung 9 legt nahe, dass es bezüglich der Stärke des Effekts der Wiederholung auf die Höhe der Urteile auf der SBS keinen relevanten Unterschied zwischen den Gruppen mit hoher und geringer Expertise gibt.

Neben dem Vergleich der subjektiven Urteile der beiden Probandengruppen besteht ein weiteres Ziel des ersten Versuchs im Vergleich des objektiven Fahrverhaltens. Dazu wurden die aufgezeichneten Fahrdaten aus der Simulation ausgewertet. Zunächst werden die Übernahmesituationen für System B betrachtet. Abbildung 10 stellt den Mittelwert der Abstände zwischen dem Egofahrzeug und dem Hindernis beim Stillstand des Egofahrzeugs dar. Situation 1 und 5 waren die Übernahmeszenarien mit geringerem Abstand zum Hindernis bei der Auslösung der Übernahme und Szenarien 4 und 8 waren die Szenarien, in denen die Fahrer eine größere Zeitreserve hatten, um eine Kollision mit dem Hindernis zu verhindern. Da die Daten anhand eines Lillieforstests keine signifikante Abweichung von der Normalverteilung aufwiesen, wurde ein t-Test zum Vergleich zwischen den Gruppen angewendet. Dabei zeigte sich, dass nur in den Situationen 4 und 8 ein signifikanter Unterschied zwischen den Probandengruppen vorlag, der jeweils einen geringeren mittleren Abstand zum Hindernis für die erfahreneren Teilnehmer anzeigte (Situation 4: $df = 50$; $T = 2{,}05$; $p < 0{,}05$; Situation 8: $df = 50$; $T = 3{,}20$; $p < 0{,}01$).

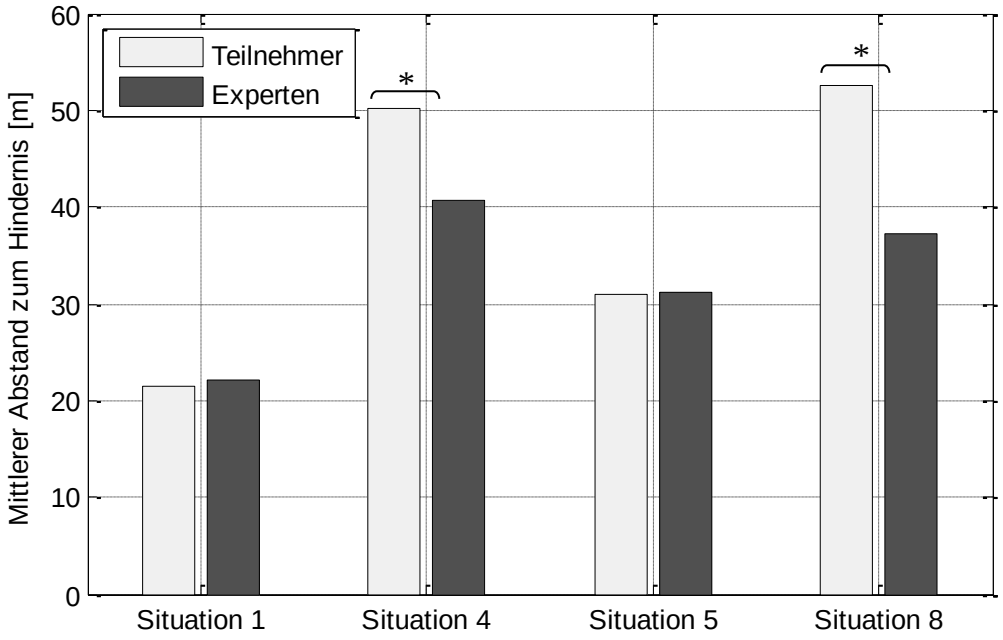

Abbildung 10 Mittlere minimale Abstände zum Hindernis in den Übernahmeszenarien in Versuch 1.

Bei der Betrachtung der in dem Szenario realisierten Längsverzögerung ergibt sich ein anderes Bild. Hierfür wurde die durchschnittliche Verzögerung zwischen Beginn und Ende des Verzögerungsvorgangs betrachtet. Abbildung 11 stellt einen Vergleich zwischen den Gruppen bezüglich der mittleren Längsverzögerung während der Übernahmeszenarien dar. Auch hier zeigt sich ein signifikanter Unterschied zwischen den Gruppen in den Szenarien 4 und 8, wobei dieser jedoch einen geringeren Bremseingriff durch die Gruppe mit höherer Expertise zeigt (Situation 4: $df = 53$; $T = 2{,}13$; $p < 0{,}05$; Situation 8: $df = 53$; $T = 3{,}05$; $p < 0{,}01$). Diese beiden Beobachtungen legen gemeinsam betrachtet nahe, dass die Experten nicht die volle zur Verfügung stehende Verzögerung genutzt haben und deswegen näher zum Hindernis zum Stehen kamen. Der geringere Abstand ist also kein Ausdruck höherer, sondern geringerer Kritikalität der Fahrsituation.

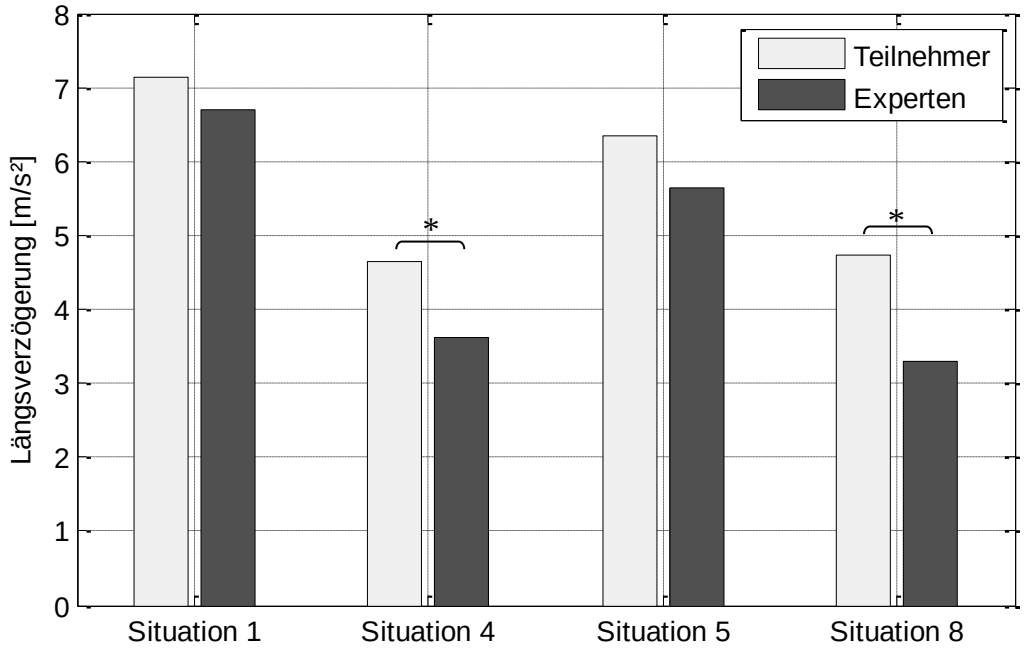

Abbildung 11 Vergleich der Gruppen bezüglich der mittleren Verzögerung in den Übernahmeszenarien in Versuch 1.

Die restlichen vier Szenarien waren die Lenkmomentszenarien. Bei diesen wurde der maximal erreichte Querversatz relativ zur ursprünglichen Position in dem eigenen Fahrstreifen ausgewertet. Auch hier bestätigt sich das Bild, dass die objektiven Maße der Kritikalität der Fahrsituation für die Experten geringer als bei den naiven Teilnehmern liegen. Signifikant ist jedoch nur der Unterschied zwischen den Gruppen in Szenario 2 ($df = 50$; $T = 3{,}32$; $p < 0{,}01$).

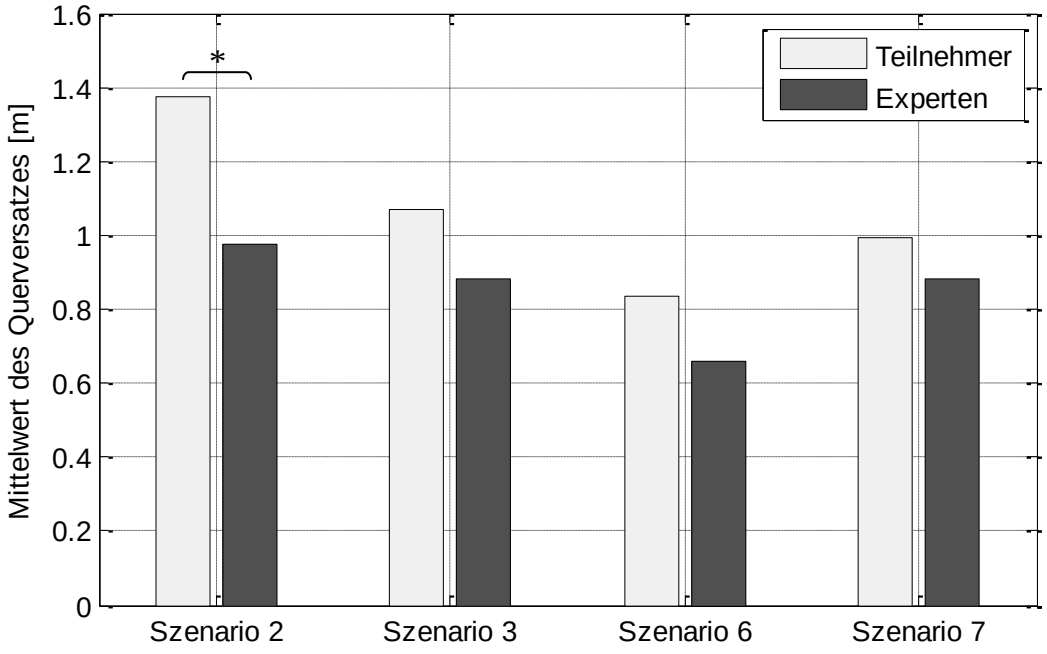

Abbildung 12 Mittelwerte des maximalen Querversatzes in den Lenkmomentszenarien in Versuch 1.

4.4.4 Diskussion

Die durchgeführte Studie legt keinen signifikanten Effekt der Expertise auf die Varianz der Urteile auf der SBS nahe, obwohl eine im Kontext von Expertenbewertungen große Anzahl an Probanden gewählt worden ist. Das bedeutet, dass bei praxisrelevanten Stichprobengrößen und den gewählten Rekrutierungskriterien nicht davon ausgegangen werden kann, dass Probenden mit höherer Expertise im Bereich Beherrschbarkeitsbewertung Urteile mit geringerer Streuung als naive Probanden abgeben. Damit können Experten auch nicht ohne weitere Änderung des Versuchsdesigns dafür genutzt werden, die Anzahl der notwendigen Probanden für eine subjektive Beherrschbarkeitsbewertung zu verringern. Verbesserte Fragemethoden sind notwendig, wenn Studien mit Teilnehmern höherer Expertise eine geringere Streuung der Urteile als Studien mit naiven Probanden aufweisen sollen.

In der Studie wurde beobachtet, dass höhere Expertise bei der Bewertung der Beherrschbarkeit von Fahrerassistenzsystemen mit höheren – also kritischeren – Urteilen auf der Störungsbewertungsskala einhergeht. Falls dieser Effekt bestätigt wird, kann er dafür genutzt werden, mittels Probanden höherer Expertise unbeherrschbare Szenarien zuver-

lässig zu erkennen, wodurch ungeeignete Systemkonzepte frühzeitig im Entwicklungsprozess erkannt werden können, sodass die Wahrscheinlichkeit, ein sicheres Design zu erhalten, steigt.

Die Ergebnisse der Studie legen nahe, dass die Probanden mit hoher und geringer Expertise sehr ähnlich durch die Wiederholung von Szenarien beeinflusst werden. Das heißt, dass höhere Expertise bei der Bewertung der Beherrschbarkeit von FAS keine praxisrelevante Verteidigung gegen Gewöhnungseffekte darstellt. Diese Erkenntnis sollte beim Design von Expertenstudien berücksichtigt werden. Die Reihenfolge und Anzahl der präsentierten Szenarien scheint für Teilnehmer mit hoher Expertise ebenso relevant wie für naive Teilnehmer.

Bei der Auswertung der objektiven Fahrdaten konnte gezeigt werden, dass die objektiven Fahrdaten der Experten entweder denen der naiven Teilnehmer ähneln oder eine geringere Kritikalität der erlebten Fahrsituation andeuten, während die subjektiven Urteile der Experten eine Tendenz zu einer höheren Bewertung der Kritikalität haben. Dies zeigt, dass die Experten nicht nach dem gleichen Bewertungsschema wie die naiven Probanden vorgehen. Die Ergebnisse können so interpretiert werden, dass die der Expertengruppe zugordnete Probanden tatsächlich nicht den erlebten Verlauf der Fahrsituation bewerten, sondern in der Lage sind, über ihr eigenes Erleben hinaus zu extrapolieren. An dieser Stelle kann jedoch keine Aussage über die Qualität der Extrapolation erfolgen. Dies kann erst auf Basis einer größeren Datenmenge erfolgen, was in Abschnitt 4.8 dargestellt wird.

Insgesamt zeigen die Ergebnisse der ersten Teilstudie, dass die Durchführung von Expertenstudien über die Kritikalität von Fahrsituationen nicht trivial ist. Wenn keine Maßnahmen zur Vermeidung von Fehleinschätzungen ergriffen werden, ist eine Unterschätzung der Kritikalität auch dann möglich, wenn eine große Anzahl an Urteilern zur Verfügung steht. Die dargestellten Ergebnisse zeigen, dass es notwendig ist, Expertenbewertungen der Beherrschbarkeit von Fahrerassistenzsystemen genauer zu analysieren.

4.5 Versuch 2

Dieser Abschnitt beschäftigt sich mit der Beschreibung des zweiten Teilversuchs in der Versuchsreihe, der im Oktober 2013 durchgeführt wurde. Neben der Erweiterung der Datenbasis für den Vergleich des Urteilsverhaltens von Experten und naiven Teilnehmern soll dieser Versuch zusätzlich Erkenntnisse über die Konfidenz der Teilnehmer in

ihre Urteile liefern. Teile dieses Unterkapitels wurden in Galaske, Weinbeer, Farid und Bengler (2015) veröffentlicht.

4.5.1 Theorie

Die Eignung von Expertenbewertungen wurde für die Sektoren Sicherheit und Medizin spätestens seit der Beschreibung von kognitiven Verzerrungen (Tversky & Kahneman, 1974) kritisch hinterfragt. Es konnte gezeigt werden, dass hohe Expertise nicht gegen die Effekte von kognitiven Verzerrungen schützt. Versuche, diesem Effekt durch ein Debiasing entgegenzuwirken waren nur eingeschränkt erfolgreich (Fischoff, 1981). Es wurde beobachtet, dass intuitive Urteile von sogenannter Overconfidence, also überschätzter Sicherheit in die Genauigkeit eigener Urteile, betroffen sein können (Lin & Bier, 2008; Moore, Tenney & Haran, 2016; Soll & Klayman, 2004). In anderen Feldern wurde jedoch beobachtet, dass Expertenurteile einen großen Nutzen haben können (Klein, 2008). Zusammengenommen ergibt sich, dass es notwendig ist, die tatsächliche Leistung von Expertenbewertungen in jedem Feld einzeln zu untersuchen, um die Vor- und Nachteile vor dem Einsatz sorgsam gegeneinander abzuwägen (Kahneman & Klein, 2009).

Der scheinbare Konflikt um die Eignung von Expertenbewertungen kann durch die individuellen Eigenschaften des Feldes, in dem die Untersuchung stattfinden soll, aufgehoben werden. Obwohl die genaue Wirkungsweise von Expertise bei Beurteilungsaufgaben noch nicht abschließend beschrieben wurde, ist es naheliegend, dass Feedback über die vorherige Schätzleistung von großer Bedeutung ist (Klein, 2008). Dieser Faktor erklärt zumindest teilweise, warum bei manchen Aufgaben Expertise einen erheblichen Effekt hat, während in anderen Bereichen selbst einfache statistische Verfahren menschlicher Expertise überlegen sind (Grove et al., 2000). Im Bereich Beherrschbarkeit von Fahrerassistenzsystemen ist solches Feedback sehr schwierig zu beschaffen. Die Entwicklung von Fahrerassistenzsystemen dauert Jahre und die Effekte solcher Systeme in realen Verkehrssystemen werden selbst unter optimalen Umständen erst Jahre nach Markteinführung objektiv messbar, wenn überhaupt. Den Grad an Beherrschbarkeit bestimmter realer Fahrsituationen aufgrund von objektiven Daten aus dem tatsächlichen öffentlichen Verkehrsgeschehen zu bestimmen, um dies dann als Feedback für Entscheider zu verwenden, die bis dahin möglicherweise bereits anderen Tätigkeiten nachgehen, scheint also wenig zielführend. Deswegen kann nicht davon ausgegangen werden, dass Experten der Beherrschbarkeit durch natürliches Feedback gut kalibriert sind. Eine genaue Untersuchung der möglichen Schätzleistung ist also notwendig.

In der vorangegangenen Studie (siehe Abschnitt 4.4) wurde eine große Streuung in den Urteilen der Kritikalität von Fahrsituationen bei Experten im Fahrsimulator beobachtet. Dafür gibt es mehrere mögliche Erklärungen. Offensichtlich ist es möglich, dass die Expertise in der Expertengruppe schlicht nicht ausreichend war, um einen messbaren Effekt auf die Streuung der Urteile zu haben. Andererseits wurde ein Effekt der Expertise auf die Höhe der Urteile der Kritikalität gefunden. Es gibt keinen offensichtlichen Grund, warum es einen Effekt auf die Höhe der Urteile gibt, jedoch nicht auf die Streuung der Urteile. Eine alternative Erklärung besteht darin, dass es für die Experten schwierig ist, eine gemeinsame Interpretation der Skala beizubehalten, da die Teilnehmer individuell und ohne Möglichkeit zur gegenseitigen Beratung befragt wurden. Dies ist der Realität einer Expertenbefragung in der Praxis tendenziell eher fern, wurde jedoch so gewählt, um soziale Interaktionseffekte auszuschließen. Dieser Mangel an zweiten Meinungen könnte dazu geführt haben, dass die Urteile unreflektiert mit einer hohen eingeschätzten Sicherheit auch dann abgegeben werden, wenn sie weit von dem wahren Wert entfernt liegen. Dieser Effekt entspricht der Verfügbarkeitsverzerrung (availability bias). Da der Teilnehmer beim Absuchen des eigenen Gedächtnisses nach Informationen, die der präferierten Lösung widersprechen, nicht fündig wird, geht er davon aus, dass keine Kontraindikatoren für die Lösung existieren und geht deswegen von einer hohen Vertrauenswürdigkeit der eigenen Lösung aus (Kahneman, 2012). Urteiler, die sich jedoch die Vielzahl möglicher Antworten bewusst sind, werden deswegen eine geringere selbsteingeschätzte Sicherheit berichten. Nach dieser Argumentation müssten Urteile, die mit geringer selbsteingeschätzter Konfidenz belegt worden sind, näher am Mittelwert der Urteile der Gruppe liegen. Weiterhin sollte die Wiederholung der gleichen Aufgabe zu einer Erhöhung der selbsteingeschätzten Urteilssicherheit einhergehen. Diese Mechanismen sollten für Teilnehmer unabhängig vom Grad der Expertise gelten.

Expertenbewertungen werden verwendet, um Fragestellungen zu beantworten, die mit anderen Methoden nicht oder nur schwierig zu beantworten sind. Für den Fall, dass dies auch mit Expertenbewertungen nicht möglich ist, ist es sinnvoll, diese Unkenntnis diagnostizieren zu können. Bei anderen Methoden kann beispielsweise die Streuung der Messergebnisse ein Indikator für eine fehlgeschlagene Messung sein. Aufgrund der geringen Anzahl an Urteilen in einer Expertenbewertung ist dies bei dieser Methode jedoch nur sehr eingeschränkt möglich. Dies liegt daran, dass die Unfähigkeit zur Übereinstimmung nicht immer eine Unfähigkeit zur korrekten Schätzung darstellt. Es wäre deswegen besser, die Konfidenz der abgegebenen Urteile direkt von den urteilenden

Experten bewerten zu lassen. Frühere Untersuchungen haben ergeben, dass die selbsteingeschätzte Konfidenz oft einer erheblichen Selbstüberschätzung unterliegt (Soll & Klayman, 2004). Für diesen Effekt gibt es mehrere Gründe. Ein Grund ist, dass bei der individuellen Bewertung der Fragestellung selten alternative Schätzungen präsentiert werden, oder die Möglichkeit besteht, nach widersprüchlichen Informationen zu suchen. Dieser Mangel an alternativen Meinungen führt zu einer Überschätzung der Konfidenz in die eigene Lösung. In diesem Fall ist die Überschätzung der Konfidenz mit dem sogenannten Availability Bias verwandt (Lin & Bier, 2008). Um die Eignung von Expertenbewertungen für die Bewertung der Beherrschbarkeit von Fahrerassistenzsystemen beurteilen zu können, wäre es wünschenswert zu untersuchen, ob dieser Effekt der Verfügbarkeit von alternativen Informationen auf die Überschätzung der Konfidenz auch bei Experten auftritt, die nach den Kriterien des RESPONSE3 Code of Practice rekrutiert worden sind.

Abgesehen von einer geringen Varianz und der richtigen Kalibrierung ist die Reliabilität ein weiterer wichtiger Aspekt einer Expertenbewertung. Eine Art diese zu messen besteht in der Test-Retest-Reliabilität. Es ist anzunehmen, dass fähige Experten unter ähnlichen Umständen die gleiche Situation ähnlich bewerten.

4.5.2 Methode

Das Ziel dieser Studie ist es, die Eigenschaften von Expertenurteilen mit den Urteilen von naiven Studienteilnehmern zu vergleichen. Zu diesem Zweck wurde, wie in der ersten Studie, als Versuchssetting ein Fahrsimulator gewählt. Verglichen mit rein textbasierten Befragungsmethoden ist hier ein geringeres Abstraktionsniveau erforderlich. Gegenüber einer Realfahrzeugstudie besteht hier der Vorteil der Kontrolle äußerer Einflüsse sowie größerer Freiheit bei der Auswahl der Versuchsteilnehmer, da keine besonderen Fahrausbildungen etwa für das Fahren auf einer Versuchsstrecke notwendig sind. Weiterhin besteht ein Vorteil in dem allgemein höheren Sicherheitsniveau im Fahrsimulator, das insbesondere beim Erleben von beherrschbarkeitsrelevanten Fahrsituationen von Bedeutung ist. Wie im Rahmen des Konzepts der Versuchsreihe beschrieben, wird hier angenommen, dass die Simulatorumgebung das Urteilsverhalten von Experten und naiven Versuchsteilnehmern auf vergleichbare Art und Weise beeinflusst.

Die Studie wurde wieder im statischen Fahrsimulator 2 des Forschungs- und Innovationszentrums FIZ in München durchgeführt. Dadurch konnte die Versuchsumgebung

gegenüber den anderen Versuchen konstant gehalten werden und somit Störeinflüsse vermieden werden.

Vor Versuchsbeginn wurden alle Versuchsteilnehmer unabhängig von der zugeordneten Gruppe der gleichen Einführungsprozedur unterworfen. Die Teilnehmer wurden über den Versuchszweck aufgeklärt und eine schriftliche Einverständniserklärung unterzeichnet. Anschließend wurden ihnen die verwendeten Beurteilungsskalen und der Simulator selbst erklärt. Alle Versuchsteilnehmer haben eine 15-minütige Einführungsfahrt absolviert. Auch diese Einführungsfahrt wurde unabhängig von der Gruppenzuordnung und damit von der vorherigen Erfahrung mit dem Simulator erlebt.

Das erste Ziel dieses Experiments war es, den Effekt von Feedback auf das Urteilsverhalten zu untersuchen. Als Form des Feedbacks wurde das zuvor gemessene Ergebnis eines von den Experten zu schätzenden Wertes auf der Störungsbewertungsskala gewählt. Diese Form des Feedbacks wurde gewählt, da sie Effekte der Formulierung des Feedbacks und der Interpretation durch den Urteiler minimiert. Es wurden zwei Szenarien geschaffen, in denen der Fahrer eine einfache manuelle Fahrt ohne relevanten Fremdverkehr auf der Autobahn erlebt hat. Als kritisches Element wurde ein kurzes überlagertes Lenkmoment zum rechten Fahrbahnrand hin aufgespielt. In einer Situation war dies ein geringes Lenkmoment, in der anderen Situation ein hohes Lenkmoment. Dieses Szenario ist in der Literatur ausreichend dokumentiert und wurde in früheren Studien auch in diesem Simulator unter Verwendung der gleichen Skala bereits validiert. Diese zwei Szenarien werden im Folgenden als Kalibrierungsszenarien bezeichnet.

Den Teilnehmern, die diese Kalibrierungsszenarien erlebt haben, wurde zunächst das Szenario mit dem geringeren Lenkmoment gezeigt. Anschließend wurden sie um eine Abschätzung der Kritikalität auf der SBS gebeten. Danach wurde ihnen der Referenzwert für dieses Szenario aus früheren Studien mit naiven Teilnehmern genannt und darum gebeten, diesen bei zukünftigen Bewertungen zu berücksichtigen. Dieses Vorgehen wurde anschließend für das Kalibrierungsszenario mit dem größeren Lenkmoment wiederholt. Dieses Vorgehen stellt eine Kalibrierung dar, bei der die verwendete Skala anhand von zwei Beispielen erneut verankert wird. Gleichzeitig erhalten die Versuchsteilnehmer Feedback über ihr eigenes Urteilsverhalten.

Um den Effekt dieser Kalibrierungsprozedur auf Versuchsteilnehmer mit hoher und geringer Expertise zu untersuchen, wurde ein unvollständiges zwei-Faktor Versuchsdesign mit den Faktoren Expertise und Kalibrierung gewählt. Da weiterhin keine Methode zur

Manipulation von Expertise bei der Bewertung der Beherrschbarkeit von Fahrerassistenzsystemen besteht, wurde dieser Faktor nicht manipuliert, sondern, wie in der ersten Studie, durch die Rekrutierungsstrategie kontrolliert. Die Teilnehmer mit hoher Expertise wurden dem bekannten Expertenpool entnommen, die Teilnehmer mit geringer Expertise mit der gleichen Methode wie in den vorangegangenen Studien rekrutiert. Die Teilnehmergruppe mit geringer Expertise hatte ein durchschnittliches Alter von 32,4 Jahren zwischen 18 und 45 Jahren. 84% der Teilnehmer waren männlich und die mittlere Jahresfahrleistung betrug 15 000 km. Für die Expertengruppe standen 15 Personen zur Verfügung.

Der zweite Faktor im Versuchsdesign war die Kalibrierung. Um den Effekt der Kalibrierungsprozedur auf die beiden Teilnehmergruppen zu untersuchen, wurden sieben Fahrszenarien entwickelt, die zunächst von allen Probanden ohne weitere Vorbereitung erlebt und bewertet wurden. Anschließend hat eine Hälfte der Teilnehmer mit geringer Expertise die gleichen Szenarien erneut bewertet (Gruppe G_1). Die andere Hälfte der Teilnehmer mit geringer Expertise erlebte nach den ersten sieben Szenarien zunächst die Kalibrierungsmethode und anschließend die Wiederholung der sieben Fahrszenarien (Gruppe G_2). Die gesamte Gruppe mit höherer Expertise (Gruppe G_3) erlebte die Kalibrierung und anschließend die Wiederholung. Dieses Design wird in Tabelle 14 dargestellt.

Tabelle 14 Zusammenfassung des Versuchsdesigns des zweiten Versuchs.

Expertise	Vorher	Kalibrierung	Nachher
Niedrig	G_{1B}	Nein	G_{1A}
Niedrig	G_{2B}	Ja	G_{2A}
Hoch	G_{3B}	Ja	G_{3A}

Wegen der begrenzten Verfügbarkeit von Teilnehmern mit höherer Expertise wurde entschieden, diese nicht aufzuteilen, um eine Bedingung mit hoher Expertise und ohne Kalibrierung durchzuführen. Da der Effekt der Wiederholung in Abhängigkeit des Expertiseniveaus bereits in der ersten Studie untersucht worden ist, wäre die hier ausgelassene Versuchsbedingung nur von sehr begrenztem Nutzen gewesen. Auf eine Validierung der in der früheren Studie erzeugten Ergebnisse wird also zugunsten einer höheren statistischen Teststärke für den Vergleich zwischen den Gruppen G_2 und G_3 verzichtet.

Zunächst wird erwartet, dass die Kalibrierungsprozedur durch die interaktive Demonstration der Skala eine Verringerung der Streuung (Abk. Var) der Urteile über die Kritikalität (Abk. Cr) der Fahrsituationen zur Folge hat. Dieser Effekt sollte sowohl bei Teilnehmern mit hoher sowie mit geringer Expertise auftreten. Die sich ergebenden Hypothesen werden in H_1 (1) und H_2 (2) formalisiert.

$$H_1: \mathrm{Var}\big(\mathrm{Cr}(G_{2A})\big) < \mathrm{Var}\big(\mathrm{Cr}(G_{2B})\big) \quad (1)$$

$$H_2: \mathrm{Var}\big(\mathrm{Cr}(G_{3A})\big) < \mathrm{Var}\big(\mathrm{Cr}(G_{3B})\big) \quad (2)$$

Um den Effekt der Kalibrierung auf die selbsteingeschätzte Konfidenz der Teilnehmer in ihre jeweiligen Urteile zu analysieren, wurden alle Teilnehmer nach jeder Bewertung gebeten, die Konfidenz in das jeweilige Urteil auf einer 5-Punkte Skala zu bewerten. Es wird nun erwartet, dass die selbsteingeschätzte Konfidenz (Abk. Co) in der Gruppe der Teilnehmer mit geringer Expertise, die die Kalibrierung erfahren hat, (G_2) nach der Kalibrierung geringer als bei der Gruppe mit geringer Expertise, die keine Kalibrierung erfahren hat, ist. Diese Hypothese wird mit (3) beschrieben.

$$H_3: \mathrm{Co}(G_{1A}) > \mathrm{Co}(G_{2A}) \quad (3)$$

Weiterhin wird erwartet, dass die selbsteingeschätzte Konfidenz vor dem Erleben der Kalibrierung größer ist, als nach dem Erleben der Kalibrierungsprozedur. Es wird erwartet, dass dieser Effekt sowohl für Teilnehmer mit geringer als auch mit hoher Expertise gilt. Die damit verbundenen Hypothesen sind in (4) und (5) formuliert.

$$H_4: \mathrm{Co}(G_{2B}) > \mathrm{Co}(G_{2A}) \quad (4)$$

$$H_5: \mathrm{Co}(G_{3B}) > \mathrm{Co}(G_{3A}) \quad (5)$$

Wenn tatsächlich überhöhte Konfidenz (Overconfidence) auftritt und die Konfidenz in die abgegebenen Urteile in der Bedingung vor der Kalibrierungsprozedur erhöht, dann sollte eine geringere Konfidenz in der Bedingung nach der Kalibrierungsprozedur mit weniger extremen Urteilen der Kritikalität einhergehen. Es wird deswegen erwartet, dass eine positive Korrelation zwischen der selbsteingeschätzten Konfidenz in die Urteile nach der Kalibrierung und der Differenz zwischen dem verbundenen Urteil und dem Mittelwert der Urteile der jeweiligen Gruppe existiert. Anders ausgedrückt wird

erwartet, dass Urteiler, die durch die Kalibrierung ihre Konfidenz verringern auch weniger extrem urteilen. Diese Hypothesen werden in (6) und (7) für beide Gruppen, die die Kalibrierung erleben, ausgedrückt.

$$H_6: \rho(Co(G_{2A}), |Cr(G_{2A})-\overline{Cr}(G_{2A})|) > 0 \quad (6)$$

$$H_7: \rho(Co(G_{3A}), |Cr(G_{3A})-\overline{Cr}(G_{3A})|) > 0 \quad (7)$$

Schließlich soll die Test-Retest-Reliabilität untersucht werden. Zu diesem Zweck wurden zwei der Szenarien in dieser Studie als Wiederholungen von Szenarien aus der ersten Studie gewählt. Diese wurden durch die gleiche Expertengruppe bereits auf der gleichen Skala unter – in der Versuchsbedingung vor der Kalibrierungsprozedur – sehr ähnlichen Umständen bewertet. Auch die Einführungsprozedur wurde nahezu identisch gehalten. Zwischen den beiden Studien lag eine Zeitspanne von 6 Monaten, sodass angenommen wird, dass keine Übertragungseffekte vorliegen. Da die Reihenfolge der präsentierten Szenarien zwischen den beiden zu vergleichenden Studien nicht identisch ist, kann jedoch keine Gleichheit der Urteile erwartet werden. Stattdessen soll eine Korrelation der Urteile der Teilnehmer mit hoher Expertise untersucht werden.

Jede Kondition bestand aus der Präsentation von sieben verschiedenen Fahrsituationen und der Bewertung der Kritikalität der Situation sowie der selbsteingeschätzten Konfidenz in das Urteil der Kritikalität. Die Anzahl der Szenarien wurde gewählt, um bei gegebener Versuchsfahrtdauer eine möglichst hohe statistische Teststärke zu erreichen. In jedem der sieben Szenarien wurde der Fahrer durch ein Fahrerassistenzsystem unterstützt, das die Längs- und Querführung des Fahrzeugs übernahm, jedoch dauerhaft überwacht werden musste. Nach der Definition der BaST ist dieses System also ein teilautomatisiertes System (Stufe 2). Kurze Beschreibungen der präsentierten kritischen Fahrsituationen können Tabelle 15 entnommen werden. Die Szenarien wurden aufgrund ihrer Relevanz für die Beherrschbarkeit des verwendeten Fahrerassistenzsystems ausgewählt. Prinzipskizzen sowie weitere Informationen zu den Szenarien finden sich im Anhang.

Tabelle 15 Beschreibung der im zweiten Versuch präsentierten Szenarien.

Nummer	Beschreibung
1	Starkes überlagertes Fehllenkmoment nach rechts auf einer Landstraße
2	Fremdfahrzeug schert auf der Autobahn moderat nah von rechts in den Fahrstreifen des Egofahrzeugs ein
3	Moderates überlagertes Fehllenkmoment nach rechts auf einer Landstraße
4	Hindernis auf der Fahrbahn in schlecht einsehbarer Kurve auf Landstraße
5	Fremdfahrzeug schert auf der Autobahn sehr nah von rechts in die Egofahrstreifen ein
6	Hindernis auf der Fahrbahn in eingeschränkt einsehbarer Kurve auf Landstraße
7	Lenkruck auf der Landstraße

4.5.3 Ergebnisse

In diesem Kapitel werden zunächst die Hypothesen aus dem vorangegangenen Abschnitt untersucht und anschließend die Test-Retest-Reliabilität für Probanden mit höherer Expertise bestimmt.

Ein Test der Bewertungen der Kritikalität der Fahrsituationen auf Normalverteilung führte zu einer Ablehnung der Hypothese der Normalverteilung. Deswegen wurden nicht-parametrische Verfahren für die weitere Analyse verwendet. Tabelle 16 stellt die Varianz der Urteile der Gruppe G_2 vor und nach der Kalibrierung in den präsentierten sieben Szenarien dar. Außerdem werden die Ergebnisse eines einseitigen Levenetests auf Gleichheit der Varianz zum Signifikanzniveau 5% dargestellt. Nicht-signifikante Ergebnisse werden als „n.s.“ dargestellt.

Tabelle 16 Gemessene Varianzen für die Gruppe G_2 vor und nach der Kalibrierung in Versuch 2.

Szenario	1	2	3	4	5	6	7
Vorher	6,06	2,24	2,92	3,83	2,71	3,46	2,08
Nachher	4,01	4,04	3,51	2,92	1,49	3,06	4,64
Levene Test	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.

Diese Ergebnisse zeigen, dass in keinem der sieben Szenarien ein signifikanter Effekt der verwendeten Kalibrierungsmethode auf die Varianz der Urteile der Kritikalität der Fahrsituationen aufgetreten ist. Auch in dem ersten Szenario direkt nach Erleben der

Kalibrierung wurde kein signifikanter Effekt gemessen. Hypothese H_1 wird deswegen abgelehnt.

Tabelle 17 stellt die Varianzen der Urteile der Gruppe mit höherer Expertise vor und nach der Kalibrierung in den 7 Fahrsituationen dar. Wie zuvor werden die Ergebnisse eines einseitigen Levenetests gezeigt. Auch hier wurden keine signifikanten Ergebnisse erzielt, sodass auch H_2 verworfen wird.

Tabelle 17 Gemessene Varianz der Urteile der Gruppe mit höherer Expertise (G_3) vor und nach Kalibrierung in Versuch 2.

Szenario	1	2	3	4	5	6	7
Vorher	2,83	1,27	2,83	5,52	1,55	3,35	4,35
Nachher	2,70	2,83	2,10	3,14	1,07	3,21	3,98
Levene test	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.

Die durchschnittliche, selbsteingeschätzte Konfidenz in den Versuchssettings wird in Tabelle 18 dargestellt. In der ersten Gruppe, die keine Kalibrierung erfahren hat, steigt die berichtete Konfidenz bei der Wiederholung im Einklang mit der Theorie. In der zweiten Gruppe liegen die Werte in den Versuchsbedingungen nach der Kalibrierung wie erwartet niedriger als vor der Kalibrierung. Für Teilnehmer mit höherer Expertise wurden in beiden Bedingungen sehr ähnliche Mittelwerte der Konfidenz ermittelt.

Tabelle 18 Durchschnittliche selbsteingeschätzte Konfidenz in den Versuchsbedingungen des zweiten Versuchs.

Gruppe	Durchschnitt Konfidenz	
	Vorher	Nachher
G_1	4,09	4,28
G_2	4,16	3,97
G_3	3,78	3,78

Da die Hypothese der Normalverteilung verworfen worden ist, wurde der Wilcoxon-Vorzeichenrangtest verwendet, um die Hypothesen 3 bis 5 zu überprüfen. Die Ergebnisse dieser Tests werden in Tabelle 19 dargestellt und unterstützen die Annahme, dass die selbsteingeschätzte Konfidenz durch die Präsentation abweichender Einschätzungen für Teilnehmer mit geringer Expertise verringert wird. Die Konfidenz der Teilnehmer mit geringer Expertise ohne Kalibrierung (G_1) war in der Nachher-Bedingung geringer als bei den Teilnehmern mit Kalibrierung (H_3) und durch die Kalibrierung ist die Konfidenz in der Gruppe G_2 kleiner geworden (H_4). Dieser Effekt ist bei den Teilnehmern

der Expertengruppe jedoch nicht aufgetreten (H_5). Entgegen der Erwartung hatten Teilnehmer mit höherer Expertise in der Vorher-Bedingung eine signifikant geringere Konfidenz in ihre Urteile als Teilnehmer mit geringerer Expertise unter den gleichen Umständen ($Co(G_{3B}) < Co(G_{1B+2B})$; $p < 0.001$).

Tabelle 19 Ergebnisse für die Hypothesen 3-5 im zweiten Versuch.

Hypothese		p-Wert
H_3	$Co(G_{1A}) > Co(G_{2A})$	< 0,001
H_4	$Co(G_{2B}) > Co(G_{2A})$	< 0,05
H_5	$Co(G_{3B}) > Co(G_{3A})$	n.s.

Die Hypothesen 6 und 7 wurden mittels Spearmans Rangkorrelation getestet und ergaben die Resultate, die in Tabelle 20 dargestellt werden. Sie stützen die Annahme, dass höhere selbsteingeschätzte Konfidenz mit größeren Abweichungen vom Gruppenmittelwert einhergehen, sodass die Erfahrung alternativer Einschätzungen diese Abweichung verringern kann.

Tabelle 20 Ergebnisse für die Hypothesen 6 und 7 des zweiten Versuchs.

Hypothese	ρ	p-Wert
$H_6: \rho(Co(G_{2A}), \lvert Cr(G_{2A}) - \overline{Cr}(G_{2A})\rvert) > 0$	0,11	0,04
$H_7: \rho(Co(G_{3A}), \lvert Cr(G_{3A}) - \overline{Cr}(G_{3A})\rvert) > 0$	0,19	0,04

Dreizehn der Teilnehmer mit höherer Expertise hatten zuvor an der ersten Simulatorstudie teilgenommen. In dieser Studie wurden zwei identische Fahrszenarien mit überlagerten Fehllenkmomenten während teilautomatisierter Fahrt auf der Landstraße erlebt und mittels der gleichen Methode bewertet. In Tabelle 15 wurden diese Szenarien als Szenarien 1 und 3 bezeichnet. Im ersten Versuch waren dies Szenarien 2 und 3. Zwischen den beiden Simulatorversuchen lagen etwa 6 Monate, sodass davon ausgegangen werden kann, dass keine Übertragungseffekte vorlagen. Abbildung 13 stellt die Mittelwerte der Urteile der Experten in den relevanten Szenarien und die Standardabweichung dar. Tatsächlich findet sich bei den schwachen Lenkmomentszenarien ein signifikanter Unterschied. Da ein Lillieforstest keine signifikante Abweichung von der Normalverteilung in diesen Szenarien gefunden hat, wurde ein zweiseitiger t-Test verwendet, der für die Urteile über die Szenarien mit dem schwachen Fehllenkmoment einen p-Wert von 0,03 ergab (n = 13, df = 24, T = 2,29). Eine Erklärung hierfür liegt in der Reihenfolge, in der die Szenarien präsentiert worden sind. Im ersten Versuch war das schwache

Lenkmoment die erste von den Probanden erlebte Lenkstörung während im zweiten Versuch bereits vorher das stärkere Lenkmoment erlebt worden ist. Dies führt durch Reihenfolgeeffekte zu verschiedenen Bewertungen und zeigt, wie schon im ersten Versuch, dass auch Experten durch Reihenfolgeeffekte beeinflusst werden.

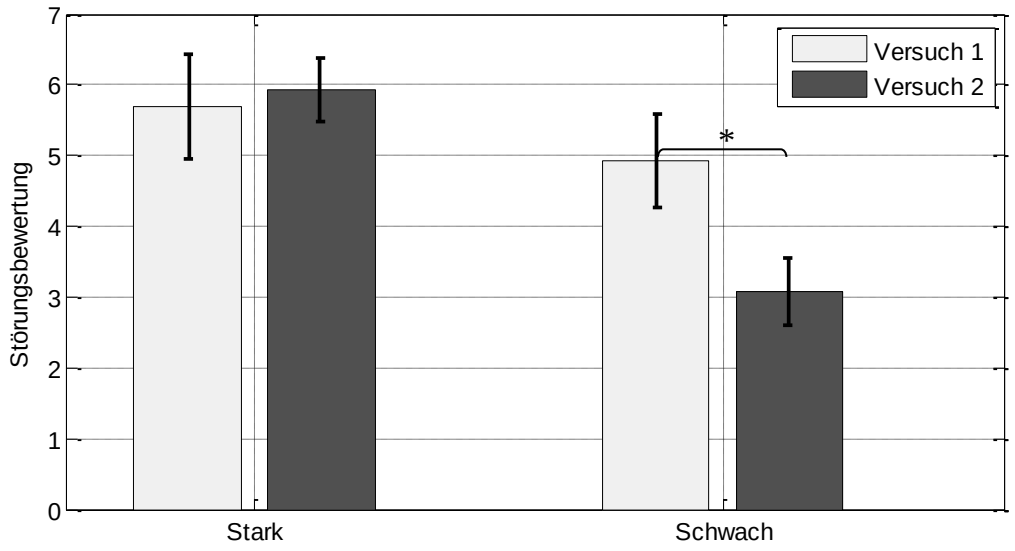

Abbildung 13 Mittlere Störungsbewertung der Lenkmomentszenarien in Versuch 1 und 2 durch Experten.

Das Vorhandensein eines signifikanten Unterschieds zwischen den Ergebnissen der beiden Versuche steht jedoch nicht im Widerspruch zur Reliabilität der Experten. Um dies zu prüfen, wird die Korrelation der Urteile der individuellen Experten zwischen den Versuchen betrachtet. Abbildung 14 stellt hierfür die Urteile der Experten für beide Szenarien als Streudiagramm dar. Zusätzlich wurde eine Regressionsgerade durch die Daten gelegt.

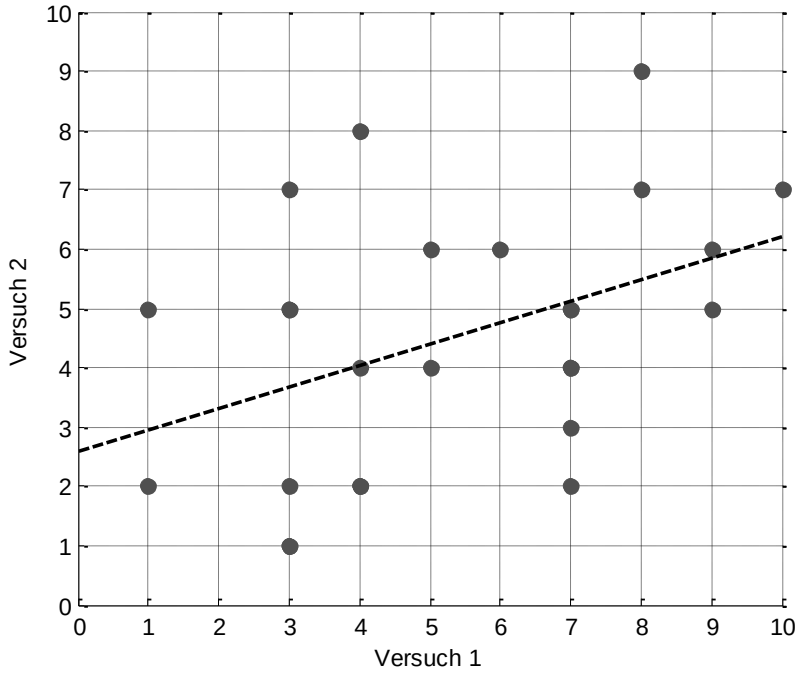

Abbildung 14 Streudiagramm der Urteile der Experten auf der SBS über die Lenkmomentszenarien in Versuch 1 und 2.

Eine Spearman-Korrelation der Urteile der vorangegangenen Studie mit den Ergebnissen der Gruppe G_{3B} ergibt ein ρ von 0,42 und einen p-Wert von 0,03.

4.5.4 Diskussion

Diese Studie untersuchte einige der Eigenschaften von Expertenbewertungen der Beherrschbarkeit von Fahrerassistenzsystemen in einer Fahrsimulatorumgebung. Eine Manipulation der verfügbaren Informationen während des Urteilsprozesses mittels einer interaktiven Demonstration alternativer Urteile verursachte einen negativen Effekt auf die selbsteingeschätzte Konfidenz der Teilnehmer geringer Expertise. Es wurde gezeigt, dass dieser Effekt mit geringeren Abweichungen vom Gruppenmittelwert nach der Behandlung korrelierte. Dies ist ein Hinweis darauf, dass Expertise in der Form von Wissen über die Fehlbarkeit des eigenen Urteilsvermögens gegen die Effekte von überhöhter Konfidenz bei der Bewertung der Beherrschbarkeit von Fahrerassistenzsystemen helfen kann.

Ergebnisse früherer Studien wurden bestätigt, da kein signifikanter Effekt der Expertise auf die Streuung der Urteile identifiziert werden konnte. Es wird daraus geschlossen, dass die aus dem RESPONSE3 Code of Practice adaptierte Rekrutierungsstrategie nicht

automatisch zu homogeneren Einschätzungen der Kritikalität von Fahrsituationen als bei naiven Teilnehmern führt, solange die Urteile isoliert entnommen werden.

Weiterhin wurde die Test-Retest Reliabilität untersucht und für das untersuchte Kollektiv an Experten auf einem niedrigen bis mittleren Niveau identifiziert. Obwohl keine der Auswahlkriterien für die Expertengruppe Teilnehmer mit einer minimalen Reliabilität in ihren Urteilen selektiert, zeigte sich in eine signifikante Reliabilität der Urteile. Es ist zu erwarten, dass die gezielte Wahl geeigneter Kriterien diese Ergebnisse drastisch verbessern kann. Das niedrige Niveau der gefundenen Korrelation legt weiterhin nahe, dass die Rekrutierungsstrategie für Expertenbewertungen aus dem RESPONSE3 Code of Practice nicht in allen Fällen zu reliablen Ergebnissen führt.

4.6 Versuch 3

Dieses Unterkapitel beschäftigt sich mit dem dritten Fahrsimulatorversuch, der im Mai 2014 durchgeführt wurde. Das erste Ziel des Versuchs ist, weitere subjektive Urteile über die Kritikalität von Fahrsituationen zu erhalten, um den angestrebten Vergleich der Probandengruppen mit einer größeren Stichprobe durchführen zu können. Zusätzlich dazu motiviert die intensive Arbeit mit der Störungsbewertungsskala in den vorherigen Versuchen als Nebenziel die Untersuchung eines alternativen Maßes der subjektiven Bewertung der Kritikalität von Fahrsituationen, das besser geeignet ist, um die Ursache der Nicht-Beherrschbarkeit zu identifizieren. Teile dieses Unterkapitels wurden in Galaske, Reisenauer, Farid und Bengler (2015) veröffentlicht.

4.6.1 Theorie

Die Feststellung der Beherrschbarkeit von Fahrerassistenzsystemen (FAS) ist gemäß ISO 26262 ein notwendiger Bestandteil der Entwicklung solcher Systeme (ISO 26262-3). Der Response 3 Code of Practice definiert die Beherrschbarkeit von FAS als ein Zusammenspiel der Fähigkeiten des Fahrers, die Kritikalität einer Fahrsituation wahrzunehmen, sich zu einer geeigneten Gegenmaßnahme zu entscheiden und diese Maßnahme erfolgreich durchzuführen (RESPONSE 3, 2009).

Diese Definition des Konstrukts Beherrschbarkeit ist bisher für eine Bewertung von FAS nur begrenzt geeignet, da keine etablierte Operationalisierung der Unterkonstrukte existiert. Die empirische Bestimmung der Beherrschbarkeit von FAS fokussierte sich

deswegen bislang auf objektiv prüfbare Kriterien der Nicht-Beherrschbarkeit (Simmermacher & Winner, 2011) sowie zusätzlich eindimensionale Bewertungen der subjektiven Störung (Neukum et al., 2008). Diese Ansätze beschränken sich jedoch darauf, die Nicht-Beherrschbarkeit festzustellen und sind nicht dafür geeignet, die Ursache davon zu ermitteln.

Dieser Abschnitt befasst sich mit den Ergebnissen einer Simulatorstudie, die die Entwicklung und Erprobung eines Fragebogens zur Operationalisierung der Subkonstrukte der Beherrschbarkeit von FAS zum Ziel hat.

4.6.2 Methode

Die im Response 3 Code of Practice genannten Unterkonstrukte Wahrnehmbarkeit der Kritikalität der Situation, Entscheidung zu einer geeigneten Gegenmaßnahme und erfolgreiche Durchführung der Maßnahme besitzen, abhängig von der betrachteten Situation, zahlreiche Facetten, für die sich jeweils eine geeignete Skala entwickeln lässt. Eine Operationalisierung für eine spezielle Situation ist jedoch nicht das Ziel dieser Studie. Es soll vielmehr versucht werden, eine möglichst generische Skala zu entwickeln, die dazu geeignet ist, verschiedene Situationen bezüglich ihrer Kritikalität entlang der verschiedenen Subdimensionen der Beherrschbarkeit von FAS zu bewerten.

Deswegen wurde mit einer Gruppe von Experten, die teilweise aus dem Expertenpool für die Versuchsreihe stammten, für jedes Teilkonstrukt ein Itemkatalog mit ca. je 30 Items erstellt, ohne einen Bezug zu einer bestimmten Situation zu erstellen. Diese wurden in anschließenden Fokusgruppen priorisiert und schließlich zu drei 6-Punkte-Likert-Skalen mit insgesamt 20 Items für alle drei Unterkonstrukte zusammengefasst. Die 6-Punkte-Skala wurde gewählt, um mittlere Urteile zu vermeiden. Die Skala wurde auf Basis der Empfehlungen in (Rohrmann, 1978) verankert. Dieses Entwicklungsverfahren wurde in Anlehnung an (Bortz & Döring, 2006) gewählt.

Diese situationsunabhängig entworfenen Skalen wurden in einem Simulatorversuch im statischen Fahrsimulator 2 des BMW Forschungs- und Innovationszentrums (FIZ) in einem gemischten between-within Design insgesamt 44 Probanden in vier Situationen präsentiert. Dabei wurden die Wahrnehmbarkeit der Kritikalität sowie die Durchführbarkeit der Gegenmaßnahme manipuliert. Auf eine Manipulation der Schwierigkeit der Entscheidung zu einer Gegenmaßnahme wurde verzichtet, da dies selbst unter Simula-

torbedingungen nur begrenzt möglich ist. Durch dieses Experiment sind 176 Bewertungen der beiden verbleibenden Konstrukte entstanden. Weiterhin wurden die subjektiv empfundene Störung der Fahrer mittels der Störungsbewertungsskala (SBS) erhoben.

Zu Beginn aller 4 Situationen befand sich das Ego-Fahrzeug auf dem rechten Fahrstreifen einer Autobahn in Folgefahrt bei einer Geschwindigkeit von 80 km/h, die mittels Tempomat eingestellt wurde. Auf dem linken Fahrstreifen befand sich dichter Verkehr, rechts vom Egofahrstreifen war ein Standstreifen. Bei Beginn der kritischen Situation wechselte das Vorderfahrzeug auf den linken Fahrsteifen und gab damit die Sicht auf ein stehendes Fahrzeug in dem Egofahrstreifen frei. Gleichzeitig erschien auf dem Standstreifen eine Person, sodass ein Ausweichen des Ego-Fahrzeugs in beide Richtungen unmöglich war.

Die Wahrnehmbarkeit der Kritikalität der Situation wurde manipuliert, indem 50 Meter vor Beginn der kritischen Situation ein Warndreieck auf dem Standstreifen platziert wurde. Damit wird speziell die Prädizierbarkeit der Situation manipuliert, die laut Response3 Code of Practice eine Komponente der Wahrnehmbarkeit der Situation darstellt. Die Manipulation der Durchführbarkeit der Gegenmaßnahme bestand in einer Veränderung des Abstands zum Hindernis, bei dem das Vorderfahrzeug den Fahrstreifen gewechselt hat. Bei einem späteren Fahrstreifenwechselvorgang des Vorderfahrzeugs verblieb dem Fahrer damit weniger Zeit, das Ego-Fahrzeug zum Stehen zu bringen und damit eine Kollision zu vermeiden, sodass der Anspruch an die rechtzeitige Durchführung des Bremsmanövers stieg. Tabelle 21 stellt die resultierenden vier Varianten dieses Szenarios zusammenfassend dar.

Tabelle 21 Varianten von Szenario 3 und 5 im dritten Versuch.

Variante	Warndreieck	Ausscherzeitpunkt Vorderfahrzeug
A	Vorhanden	Früh
B	Vorhanden	Spät
C	Fehlt	Früh
D	Fehlt	Spät

Insgesamt wurden jedem Teilnehmer fünf Szenarien präsentiert. Die ersten zwei Szenarien fanden auf der Landstraße statt, anschließend wurde die erste Hindernissituation präsentiert. Danach erlebten die Probanden ein kritisches Szenario, bei dem ein Einscherer von rechts in den Fahrstreifen des Egofahrzeugs eindringt. Schließlich wurde eine zweite Variante der Hindernissituation präsentiert, womit der Versuch beendet wurde.

Tabelle 22 Dargestellte Szenarien im dritten Versuch.

#	Beschreibung
1	Fehllenkmoment bei Fahrt auf der Landstraße bei Fahrt mit Tempomat
2	Hindernis hinter schlecht einsehbarer Kurve auf Landstraße bei Fahrt mit Tempomat
3	Stehendes Fahrzeug auf dem eigenen Fahrstreifen auf der Autobahn #1
4	Kritischer Einscherer bei der Fahrt auf dem linken Fahrstreifen der Autobahn mit Tempomat
5	Stehendes Fahrzeug auf dem eigenen Fahrstreifen auf der Autobahn #2

Für die Studie wurden 44 freiwillige Studienteilnehmer aus dem Umfeld des BMW FIZ rekrutiert. Das durchschnittliche Alter der Teilnehmer lag bei 30 Jahren und die mittlere Fahrleistung lag bei ca. 10 000 km pro Jahr. Alle Teilnehmer wurden danach selektiert, einen gültigen Führerschein zu besitzen, jedoch keine erhöhte Fahrausbildung erfahren zu haben und nicht im Umfeld Fahrerassistenzsysteme zu arbeiten. Weiterhin haben als zweite Teilnehmergruppe 18 Experten aus dem Expertenpool teilgenommen.

Die Teilnehmer wurden zufällig einer von vier Versuchsfahrten zugeordnet, die jeweils zwei der vier möglichen Hindernissituationen enthalten haben. Dabei wurde die Reihenfolge der Szenarien permutiert, um Reihenfolgeeffekte auszugleichen. Tabelle 23 stellt die Versuchsfahrten dar.

Tabelle 23 Permutation der Szenarien in den Versuchsfahrten des dritten Versuchs.

Fahrt	Situation 3	Situation 5
1	A	B
2	B	A
3	C	D
4	D	C

Die Teilnehmer haben nach einer Aufklärung über den Versuchszweck und einer kurzen Einführung in die genutzten Skalen eine Einführungsfahrt von 15 Minuten absolviert. Anschließend wurde die erste Prüfsituation gestartet. Nach dem ersten kritischen Szenario wurde zunächst der entworfene Fragebogen vom Probanden ausgefüllt und anschließend eine Bewertung auf der Störungsbewertungsskala abgefragt. Danach wurde die nächste Situation gestartet und dabei wie bei der ersten Situation vorgegangen.

4.6.3 Ergebnisse

Tabelle 24 stellt die Mittelwerte der Scores der beiden Subskalen für die vier Varianten der Szenarien drei und fünf dar. Die Ergebnisse auf der Skala zur Wahrnehmbarkeit der Kritikalität liegen in den Szenarien A und B erwartungsgemäß unter den Scores in den Szenarien C und D. Auch die Scores auf der Skala zur Durchführbarkeit der Gegenmaßnahme verhalten sich erwartungsgemäß (B > A und D > C). Auf der Störungsbewertungsskala wurde die Verschärfung des Szenarios ebenfalls wiedergegeben, da auch hier der mittlere Wert in A geringer liegt als in allen anderen Szenarien und in D am höchsten liegt.

Tabelle 24 Mittelwerte der Subskalen in den kritischen Situationen drei und fünf für Probanden im dritten Versuch.

Szenario	Skala Wahrnehmung [1-6]	Skala Durchführung [1-6]	SBS [0-10]
A	2,80	2,73	4,89
B	3,08	3,21	5,59
C	3,17	2,67	5,47
D	3,32	2,94	6,27

Anhand dieser Ergebnisse werden nun vier Hypothesen überprüft. In den Szenarien mit Warndreieck (A & B) sei der Score auf der Skala zur Wahrnehmbarkeit der Kritikalität geringer als in den Szenarien ohne Warndreieck (C & D) (Hypothese 1). Dies gelte ebenso für die Scores auf der Störungsbewertungsskala (Hypothese 2). In den Szenarien, in denen das Vorderfahrzeug spät ausschert (B & D) sei der Score auf der Skala zur Durchführbarkeit der Gegenmaßnahme höher als in den Szenarien, in denen das Vorderfahrzeug früh ausschert (A & C) (Hypothese 3). Dies gelte analog für die Scores auf der SBS (Hypothese 4).

Ein Lilliefors-Test ergab keine signifikante Abweichung der Scores von der Normalverteilung. Deswegen werden die Effekte der Manipulationen mittels einseitiger t-Tests überprüft. Tabelle 25 gibt einen Überblick über die Ergebnisse der Hypothesentests sowie die ermittelten Hedges-g Effektstärken.

Tabelle 25 Ergebnisse der Hypothesentests und Effektstärken im dritten Versuch.

Hypothese	p-Wert	Hedges-g Effektstärke
1	0,029*	0,42
2	0,059	(0,34)
3	0,021*	0,45
4	0,033*	0,41

Diese Ergebnisse legen nahe, dass die jeweiligen Manipulationen auf den Skalen der Wahrnehmbarkeit der Kritikalität sowie Durchführbarkeit der Gegenmaßnahme erfolgreich wiedergegeben wurden. Auf der Störungsbewertungsskala verursachte lediglich die Manipulation der Durchführbarkeit einen signifikanten Unterschied in den Scores. Die ermittelten Effektstärken liegen für alle signifikanten Hypothesen in der Nähe von 0,4. Sie können also als klein bis mittelgroß eingeordnet werden (Cohen, 1988).

4.6.4 Diskussion

Das Hauptziel dieser Studie bestand darin, weitere Szenarien für einen Vergleich der Urteile von Experten und Probanden zu testen. Dieses Ziel wurde durch die Präsentation der Szenarien an Teilnehmer mit geringerer und höherer Expertise erfüllt. Die Auswertung dieser Daten wird in der Gesamtauswertung über alle Teilversuche in Abschnitt 4.8 diskutiert.

Als Nebenziel sollte untersucht werden, in wie weit es möglich ist, Unterschiede in der Wahrnehmbarkeit der Kritikalität einer Fahrsituation und Durchführbarkeit der Gegenmaßnahme mittels situationsunabhängiger Likert-Skalen zu ermitteln. Dadurch könnte die mittels der SBS festgestellte subjektive Kritikalität auf eine der Grundkonstrukte der Beherrschbarkeit zurückverfolgt werden und damit ein Werkzeug für die Analyse der Ursache der subjektiven Kritikalität entstehen. Die in Abschnitt 4.6.3 berichteten Ergebnisse zeigen exemplarisch, dass dies möglich ist. Als wichtige Grundkonstrukte der Beherrschbarkeit von Fahrerassistenzsystemen können derartige Skalen einen wichtigen Beitrag zur Diagnose der Nicht-Beherrschbarkeit von Fahrerassistenzsystemen darstellen.

Der Vergleich der Effektstärken zwischen den erstellten Skalen sowie der SBS zeigt, dass die Werkzeuge die Manipulationen ähnlich scharf abbilden, wobei die Ursache des Unterschieds zwischen den Situationen anhand der SBS-Scores nicht ermittelt werden kann. Im Gegenzug ist der Einsatz der Störungsbewertungsskala mit deutlich weniger Aufwand verbunden.

Die hier dargestellten Ergebnisse wurden lediglich anhand eines Szenarios mit zwei Manipulationen erzeugt. Dadurch kann eine externe Validität nicht empirisch begründet werden. Da die beiden Skalen jedoch ohne Kenntnis der Prüfsituationen und unter dem Gesichtspunkt der Generalisierbarkeit erstellt worden sind, spricht aus theoretischer Sicht wenig dagegen. Gewissheit indes können nur zusätzliche Studien verschaffen. Die in diesem Versuch verwendete Implementierung darf nur als Hinweis auf die Machbarkeit solcher Skalen verstanden werden, um weitere Untersuchungen in dieser Richtung zu motivieren. Der Hauptzweck dieses Versuchs liegt in der Erzeugung weiterer Daten für die Gesamtauswertung über alle Versuche hinweg, die ab Abschnitt 4.8 dargestellt wird.

4.7 Versuch 4

In diesem Unterkapitel wird die vierte und letzte Teilstudie dargestellt. Diese Studie wurde im Oktober 2014 durchgeführt. Wie bei den anderen Versuchen besteht das erste Ziel darin, weitere Urteile der Probandengruppen zu entnehmen. Darüber hinaus soll die Entnahme von Konfidenzintervallen untersucht werden, wie sie in anderen Fachbereichen angewendet wird (Cooke & Goossens, 2000). Als drittes Ziel soll ein Vergleich der bisherigen Bewertungsform im Fahrsimulator und auf Textbasis erfolgen, da der Response 3 Code of Practice diese Form ebenfalls erlaubt.

4.7.1 Theorie

In den vorangegangenen Versuchen wurde der Zusammenhang zwischen dem Expertiseniveau und dem Urteil auf verschiedenen Skalen untersucht. Der dabei gefundene Zusammenhang ist jedoch nicht für alle Bewertungen der Beherrschbarkeit von Fahrerassistenzsystemen ideal. In Situationen, in denen eine Unterschätzung der Kritikalität der Fahrsituation mit besonderer Sorgfalt verhindert werden muss – etwa, weil die auftretenden Effekte neu sind – wäre es wünschenswert, das Urteilsverhalten manipulieren zu können.

Im zweiten Versuch wurde gezeigt, dass die selbsteingeschätzte Konfidenz kein geeignetes Maß für die objektive Unsicherheit der Urteile darstellt. Dennoch wäre ein solches Maß für eine Bewertung der Kritikalität von großer Bedeutung.

Die bisherigen drei Versuche haben sich darauf konzentriert, einen Vergleich zwischen zwei Gruppen im statischen Fahrsimulator durchzuführen. Diese Auswahl wurde durch

den Vorversuch begründet, in dem der statische Fahrsimulator mit einem dynamischen Fahrsimulator und einem Versuch im Realfahrzeug verglichen wurde. Es gibt aber noch keine Hinweise darauf, in wieweit eine Bewertung der Kritikalität einer Fahrsituation basierend auf einer reinen Textbeschreibung der Fahrsituation möglich ist. Hinweise hierzu würden einen wichtigen Beitrag zur Beurteilung der Eignung von Experten für die Bewertung der Beherrschbarkeit von Fahrerassistenzsystemen leisten.

4.7.2 Methode

Um die drei formulierten Versuchsziele zu erreichen, wurde ein zweistufiges Versuchsdesign entwickelt. Der erste Versuchsteil bestand darin, den Teilnehmern mit höherer Expertise eine textbasierte, illustrierte Beschreibung von zwei Fahrsituationen in Papierform vorzulegen und diese auf der bereits zuvor verwendeten SBS bewerten zu lassen. In der zweiten Versuchsbedingung des ersten Versuchsteils wurden die beiden in Textform beschriebenen Fahrsituationen nach einer Einführungsfahrt im Fahrsimulator erlebt. Es handelt sich also um ein 1-Faktor Experiment mit zwei Ausprägungen der unabhängigen Variable Versuchssetting. In beiden Versuchsumgebungen wurden die Experten gebeten, die 10er, 50er und 90er Perzentile der Urteile von Teilnehmern mit geringer Expertise im Fahrsimulator zu schätzen.

Tabelle 26 Darstellung des Versuchsdesigns für den ersten Versuchsteil des vierten Simulatorversuchs.

Fragebogen			Simulator		
10er Perzentil n = 17	50er Perzentil n = 17	90er Perzentil n = 17	10er Perzentil n = 17	50er Perzentil n = 17	90er Perzentil n = 17

Der zweite Versuchsteil bestand daraus, dass zwei Teilnehmergruppen mit geringer und höherer Expertise im statischen Fahrsimulator nach einer Einführungsfahrt fünf unterschiedliche Fahrsituationen gezeigt wurden. Beide Teilnehmergruppen wurden nach jeder Fahrsituation gebeten, die Fahrsituation auf der SBS zu bewerten. Sie wurden zunächst danach gefragt, wie sie die Situation persönlich einschätzen. Die Teilnehmer der Gruppe mit höherer Expertise wurde zusätzlich gebeten zu schätzen, wie hoch das 10er, 50er und 90er Perzentil in der Gruppe mit geringerer Expertise liegt. Jeder Teilnehmer der Gruppe mit höherer Expertise hat also je Situation 4 verschiedene Fragen bewertet. Es handelt sich damit um ein unvollständiges 2-Faktor Design mit zwei Abstufungen der Variable Expertiseniveau und vier Abstufungen der Variable „Fragestellung“. Die

Reihenfolge der Fragen wurde nicht permutiert. Da keine Hypothesen über die Unterschiede zwischen den Situationen formuliert werden, ist die Fahrsituation lediglich eine Kontrollvariable, die aus praktischen Gründen nicht konstant gehalten werden kann, da sonst starke Übertragungseffekte und Monotonie bei den Teilnehmern eintreten. Das Versuchsdesign wird in Tabelle 27 dargestellt.

Tabelle 27 Darstellung des Versuchsdesigns für den zweiten Teil des vierten Versuchs.

		Frage				
		Text-bewertung	Selbst-einschätzung	10er Perzentil	50er Perzentil	90er Perzentil
Expertise	Gering	-	n = 34	-	-	-
	Hoch	n = 17	n = 17	n = 17	n = 17	n = 17

Jede der Messungen wurde, um die statistische Aussagekraft zu erhöhen, anhand verschiedener Situationen wiederholt. Die Bewertungen auf dem Fragebogen wurden anhand von zwei Situationen durchgeführt, die auch im Fahrsimulatorteil des Versuchs dargestellt wurden. Insgesamt gab es fünf verschiedene Situationen, die die Grundlage des zweiten Versuchsteils bilden. Tabelle 28 beschreibt die Situationen kurz und gibt eine Übersicht, wie die Situationen verwendet wurden. Die Fahrsituationen 1 und 2 wurden so gewählt, dass sie auch in Textform einfach verständlich und dennoch präzise beschrieben werden können, um den Fokus der Bewertung auf die Beherrschbarkeitsexpertise und weniger auf das Situationsverständnis zu legen. Die weiteren drei Situationen wurden nicht in Textform bewertet und unterlagen nicht der gleichen Beschränkung einfacher Beschreibbarkeit in Textform.

Tabelle 28 Beschreibung der dargestellten Fahrsituationen im vierten Simulatorversuch.

#	Beschreibung	Text-basiert	Fahrsimulator
1	Vorderfahrzeug bremst im Stadtverkehr unvermittelt mit Vollverzögerung ab.	Ja	Ja
2	Fehlerhaftes starkes Lenkmoment zum rechten Fahrbahnrand auf der Landstraße.	Ja	Ja
3	Geringes fehlerhaftes Lenkmoment zur rechten Seite bei Überholen von Fahrradfahrer.	Nein	Ja
4	Aktivierte Querführung reagiert nicht auf Fußgänger am rechten Rand des Fahrstreifens.	Nein	Ja
5	Fehlerhaftes Lenkmoment während einer Autobahnbaustelle.	Nein	Ja

Das erste Versuchsziel bestand darin, eine Manipulation der Höhe der erzeugten Urteile der Experten vornehmen zu können. In vorherigen Versuchen wurde üblicherweise nur nach dem Mittelwert oder Median der Urteile der Probanden gefragt. Diese beiden Lageparameter liegen in diesem Anwendungsfall nah beieinander, da die Skala der Bewertung auf 11 Stufen beschränkt ist und die Streubreite der Urteile relativ zur Breite der Skala groß ist. Außerdem werden genug Probanden verwendet, um näherungsweise normalverteilte Urteile zu erhalten. Deswegen können nur in geringem Maße Ausreißer entstehen, die eine Verzerrung des Mittelwerts verursachen können, sodass der Median und der Mittelwert hier als näherungsweise identisch angenommen werden können.

Um nun eine Manipulation der Höhe des Lageparameters der Verteilung der Probanden durchzuführen, wird die Formulierung des Medians als Perzentil ausgenutzt. Damit kann die Frage entlang einer Intervallskala von 0 bis 100 verschoben werden. Der Effekt dieser Manipulation kann durch die drei Fragen nach dem 10er, 50er und 90er Perzentil überprüft werden. Es wird als Hypothese formuliert, dass größere Perzentilwerte in der Frage mit größeren Urteilen der Experten einhergehen. Die verbundenen Hypothesen sind $H_{1\text{-}3}$.

H_1: Die Urteile über das 50er Perzentil sind größer als die Urteile über das 10er Perzentil.

H_2: Die Urteile über das 90er Perzentil sind größer als die Urteile über das 50er Perzentil.

H_3: Die Urteile über das 90er Perzentil sind größer als die Urteile über das 10er Perzentil.

Das zweite Versuchsziel bestand darin, die Streuung der Urteile der Probanden durch Teilnehmer mit höherer Expertise abzuschätzen. Dies ist mittels der 10er und 90er Perzentile, die die Teilnehmer mit höherer Expertise schätzen, möglich. Als Maß für die Güte der Streuungsschätzung wird der Anteil der Urteile der Teilnehmer geringer Expertise gewählt, der innerhalb des von den Teilnehmern höherer Expertise geschätzten Bereichs liegt. Dieses Vorgehen wurde in Anlehnung an Soll & Klayman (2004) gewählt. Dort erreichte eine Expertengruppe, die einen Zielparameter mittels eines 80%-Konfidenzintervalls eingrenzen sollte, eine Treffsicherheit von 48%. In diesem Fall liegt der höchste zu erwartende Wert, bei perfekter Kalibrierung der Teilnehmer höherer Expertise, bei 80%. Als Mindesterwartung wird hier, auch im Vergleich mit der Literatur, eine Treffsicherheit von 40% vorausgesetzt. Dies wird in Hypothese 4 beschrieben.

H_4: Je Szenario liegen mindestens 40% der Urteile der Teilnehmer geringerer Expertise in dem Intervall zwischen dem Median der von den Teilnehmern höherer Expertise geschätzten 10er und 90er Perzentile.

Das dritte Versuchsziel schließlich zielte auf den Vergleich von Szenarienbeschreibungen in Schriftform sowie im Fahrsimulator ab. Dieser wird durch den ersten Versuchsteil ermöglicht. Da zunächst keine Hypothese über die Richtung des Zusammenhangs zwischen den Urteilen in Schriftform sowie im Simulator existiert, wird lediglich ein Unterschied zwischen den Urteilen in den beiden Versuchssettings angenommen. Diese Hypothese wird in H_5 und H_6 formalisiert.

$H_{5/6}$: Die Urteile der Teilnehmer höherer Expertise in Schriftform über die Urteile der Teilnehmer mit geringer Expertise unterscheiden sich von den Urteilen der gleichen Gruppe im Fahrsimulator in Szenario 1/2.

4.7.3 Ergebnisse

Tabelle 29 gibt eine Übersicht über die Mediane der Urteile der einzelnen Gruppen in den verschiedenen Versuchsbedingungen.

Tabelle 29 Mediane der Urteile im Simulator im vierten Versuch.

Expertise	Niedrig	Hoch			
Urteilsart	Subjektiv	Subjektiv	Perzentile der naiven Teilnehmer		
			10er	50er	90er
Szenario 1 Simulator	3	4	3	5	6
Szenario 1 Schriftlich	-	-	3	6	8
Szenario 2 Simulator	4	5	3	5	6
Szenario 2 Schriftlich	-	-	3	6	8
Szenario 3 Simulator	4	4	4	5	7
Szenario 4 Simulator	6	5	3	5	7
Szenario 5 Simulator	7	5	4	6	8

Die Hypothesen eins bis drei erzielten jeweils hochsignifikante Ergebnisse mit p-Werten kleiner als 0,001. Tabelle 30 stellt die statistischen Kennwerte und Hedges-g Effektstärken für diese Hypothesen dar. Diese Ergebnisse legen nahe, dass die Manipulation der Höhe der Urteile durch Veränderung des gefragten Perzentilwerts erfolgreich war.

Tabelle 30 Statistische Kennwerte der Hypothesen 1-3 des vierten Simulatorversuchs.

Hypothese	p-Wert	Hedges g
1	< 0.001	0,95
2	< 0.001	0,90
3	< 0.001	1,85

Die folgende Tabelle 31 stellt dar, wieviel Prozent der Urteile der Teilnehmer geringer Expertise innerhalb des von den Teilnehmern höherer Expertise im Simulator geschätzten Intervalls zwischen dem 10er und 90er Perzentil der Urteile der Teilnehmer geringerer Expertise gelegen haben.

Tabelle 31 Anteil der Urteile der naiven Teilnehmer, die in dem von Experten im Simulator geschätzten Intervall liegen.

Szenario 1	Szenario 2	Szenario 3	Szenario 4	Szenario 5
47,4%	42,7%	47,4%	48,8%	50,3%

Der Mittelwert dieser Anteile liegt bei 47,3%. Damit wird Hypothese 4 angenommen. Dieser Wert liegt nah an dem Wert von 48%, der in (Soll & Klayman, 2004) ermittelt wurde.

Hypothese fünf wird mittels eines Vorzeichenrangtests auf Gleichheit des Medians für die beiden in Schriftform abgefragten Szenarien geprüft. Ein Levenetest auf Gleichheit der Varianz in den beiden Versuchsbedingungen hat keine signifikante Abweichung von Homoskedastizität ergeben, sodass die Voraussetzungen für den Vorzeichenrangtest gegeben sind. Die Ergebnisse des Tests können Tabelle 32 entnommen werden.

Tabelle 32 Ergebnisse zu Hypothesen 5 und 6 im vierten Versuch.

Hypothese	Szenario	p-Wert	Hedges g
H_5	1	< 0,01	0,88
H_6	2	< 0,01	0,96

Diese Ergebnisse zeigen, dass der Median der Urteile in Schriftform signifikant größer waren als die Urteile im Fahrsimulator. Die große Effektstärke zeigt, dass es sich hierbei um einen relevanten Faktor handelt.

Für die Urteile der Teilnehmer mit höherer Expertise im Fahrsimulator wurde zuvor bereits die Effektstärke für eine Änderung des Perzentilwerts in der Frage nach den Ur-

teilen naiver Probanden ermittelt. Zum Vergleich wurden auch die entsprechenden Effektstärken für die Auswirkung des Perzentilwerts in den Fragen in Schriftform ermittelt. Bei Veränderung des Perzentilwerts von 10 auf 50, analog zu Hypothese 1, ergibt sich eine Effektstärke von 1,30. Bei einer Veränderung von 50 auf 90, analog zu Hypothese 2, folgt eine Effektstärke von 1,37. Die Effektstärke für eine Veränderung des Perzentilwerts von 10 auf 90 ergibt sich zu 2,62.

4.7.4 Diskussion

Im vorangegangenen Abschnitt konnte gezeigt werden, dass die Manipulation der Höhe der Urteile der Teilnehmer mit höherer Expertise durch Veränderung des Perzentilwerts bei der Entnahme der Urteile erfolgreich war. Diese Manipulation könnte dafür genutzt werden, Fehlkalibrierungen der Urteiler auszugleichen. Die hier ermittelten Effektstärken müssen jedoch als Obergrenzen gelten, da bei der Manipulation bereits der Bereich von 10 bis 90 ausgereizt worden ist. Bei noch extremeren Perzentilwerten in den Fragen ist von Randeffekten auszugehen, die hier nicht untersucht worden sind. Es wurde auch lediglich die Existenz des Effekts der Manipulation nachgewiesen. Über die Linearität des Effekts zwischen den Datenpunkten 10, 50 und 90 kann keine Aussage getroffen werden.

Weiterhin wurde gezeigt, dass dieser Effekt auch in Schriftform existiert. Es ergeben sich in dieser Versuchsbedingung sogar noch größere Effektstärken, sodass auch hier eine Manipulation der Höhe der Urteile möglich scheint.

Es wurde gezeigt, dass die Fähigkeit der Teilnehmer höherer Expertise, das Intervall der Urteile der Teilnehmer geringerer Expertise einzuschätzen, auf einem Niveau liegt, dass sich in der Literatur auch bei anderen Anwendungen von Expertenbewertung wiederfindet. Damit können Ansätze zur Kompensation von Fehlkalibrierungen, wie etwa Skalierung, übertragen werden.

Schließlich wurde gezeigt, dass die Urteile der Teilnehmer höherer Expertise in Schriftform höher als im Fahrsimulator liegen. Dies stützt die Annahme, dass die Urteiler mit höherer Expertise größere Unsicherheiten bei der Bewertung tendenziell durch höhere Kritikalitätsurteile kompensieren. Ein Werkzeug zur Kompensation dieser durch die Schriftform erhöhten Kritikalität wurde im vorangegangenen Abschnitt diskutiert. Bei den Urteilen in Schriftform wurde keine erhöhte Streuung der Urteile gefunden.

4.8 Gesamtauswertung

Die zuvor beschriebenen vier Versuche fanden alle unter ähnlichen Randbedingungen im gleichen Simulator statt. Die Gruppe der Teilnehmer mit höherer Expertise wurde dabei soweit möglich konstant gehalten. Die 21 Teilnehmer der Expertengrupppe haben im Durchschnitt jeweils an 3,3 Versuchen teilgenommen. 12 Teilnehmer haben an allen vier Versuchen teilgenommen. 5 weitere haben an 3 Versuchen partizipiert. Dadurch können die Ergebnisse der verschiedenen Studien miteinander verglichen werden und gesamthaft ausgewertet werden. Dies wird in diesem Kapitel durchgeführt. Dazu wird die Gruppe mit höherer Expertise zunächst als einheitliches Kollektiv betrachtet und mit den Teilnehmern geringer Expertise verglichen. Im zweiten Unterkapitel werden die Teilnehmer höherer Expertise, die an mehreren Versuchen teilgenommen haben, individuell mit dem Kollektiv der Teilnehmer geringerer Expertise verglichen und ihr Urteilsverhalten quantifiziert. Schließlich wird diskutiert, wie anhand solcher Daten neue Expertenkollektive konstruiert werden können und welche Güte der Urteile damit anhand des gegebenen Kollektivs erreicht werden kann.

4.8.1 Gruppenbetrachtung

In diesem Abschnitt wird davon ausgegangen, dass die in den vorangegangenen vier Versuchen beobachteten Teilnehmer mit höherer Expertise ein Kollektiv mit homogenen Eigenschaften darstellen. Die Expertenbewertung besteht hier in der Prognose des Mittelwerts der Teilnehmer geringerer Expertise anhand des Mittelwerts der Teilnehmer mit höherer Expertise. Zu diesem Zweck wird eine lineare Regression angewendet.

Die verwendeten Daten sind die Mittelwerte der Urteile der beiden Gruppen in insgesamt 30 Szenarien. Dies sind 8 Szenarien aus der ersten Studie, 14 Szenarien aus Studie zwei, die drei unpermutierten Szenarien aus der dritten Studie sowie fünf Szenarien aus der vierten Studie. Die Mittelwerte der Urteile der Teilnehmer mit geringer Expertise sind im Durchschnitt aus je 39,75 Urteilen ermittelt worden. Für die Mittelwerte der Urteile der Teilnehmer mit höherer Expertise wurden im Schnitt 17,25 Urteile herangezogen. Fehlende Werte in Form von abwesenden Experten wurden für diese Auswertung ignoriert. Für beide Datensätze wurde mittels eines Lillieforstest keine signifikante Abweichung von der Normalverteilung gefunden. Ein Test auf Homoskedastizität führte ebenfalls zu keinem signifikanten Unterschied zwischen den Varianzen der Urteile der Gruppen. Damit ist nach dem Gauß-Markov-Theorem die Kleinste-Quadrate-Methode der beste lineare erwartungstreue Schätzer.

Aus den Daten ergibt sich durch eine Pearson Korrelation ein ρ von 0,8235 bei einem p-Wert kleiner 0,001. Ein Lillieforstest auf Normalverteilung der Residuen ergab keine signifikante Abweichung von der Normalverteilung. Es gibt also einen erheblichen Zusammenhang zwischen den beiden Datensätzen. Dieser wird in Form einer Regressionsanalyse in Abbildung 15 dargestellt. Die dort abgebildete Regressionsgerade hat eine Steigung von ca. 0,81 und eine Abszisse von etwa 1,63. Diese Parameter bestätigen frühere Beobachtungen, dass Teilnehmer höherer Expertise bei wenig kritischen Szenarien zu einer Überbewertung der Kritikalität neigen. Die Abbildung enthält weiterhin das 95%-Konfidenzband der Regressionsanalyse, dass den Raum der für diese Daten möglichen Zusammenhänge darstellt.

In der Abbildung ist weiterhin das 90%-Prädiktionsintervall der Regressionsanalyse angegeben. Dieses wurde gewählt, da die untere Grenze dieses Intervalls dafür verwendet werden kann, den niedrigsten zu akzeptierenden Mittelwert der Urteile der Teilnehmer höherer Expertise bei einem gegebenen nicht zu überschreitenden Mittelwert der Urteile der Teilnehmer geringerer Expertise zu ermitteln. Sollen beispielsweise Mittelwerte der Urteile der Teilnehmer mit geringerer Expertise über 5 zum Signifikanzniveau 5% ausgeschlossen werden, gibt diese Kurve Aufschluss darüber, dass bei diesem Modell und Kollektiv der höchste zu tolerierende Mittelwert der Urteile der Teilnehmer höherer Expertise bei etwa 4,7 liegt.

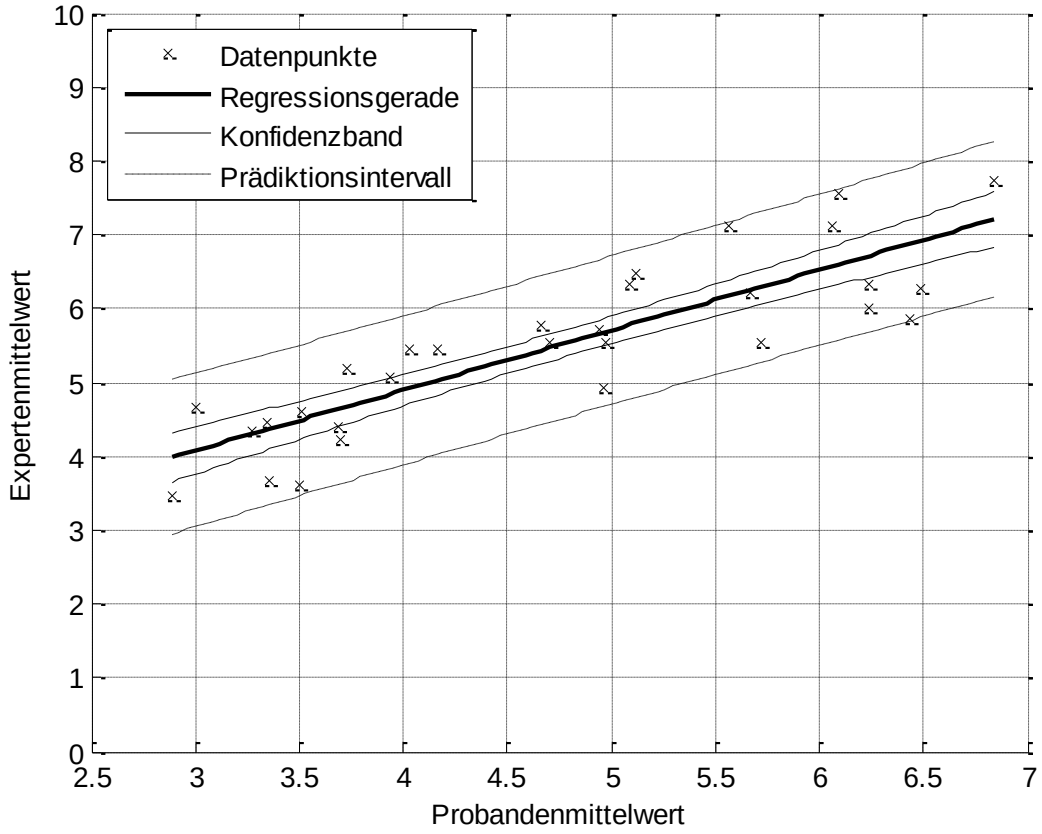

Abbildung 15 Regressionsanalyse der Gruppenbetrachtung.

Für die in Abbildung 15 dargestellte Regression standen durchschnittlich 17,25 Teilnehmer höherer Expertise zur Verfügung, um den Mittelwert der Urteile der Teilnehmer geringerer Expertise abzuschätzen. Im Folgenden soll untersucht werden, wie sich eine Verringerung der Anzahl der Teilnehmer auf die Genauigkeit der Prognose auswirkt.

Dazu wird angenommen, dass der Prognosefehler alleine aufgrund des Standardfehlers des Mittelwerts zustande kommt. In dem Fall ist der Prognosefehler proportional zum Kehrwert der Wurzel der Anzahl der Teilnehmer höherer Expertise (Maybeck, 1979):

$$\varepsilon(n) \sim \frac{\sqrt{17{,}25}}{\sqrt{n}}$$

Damit kann der Prognosefehler von Expertengruppen mit n Teilnehmern geschätzt werden, indem der anhand der durchgeführten Versuche mit durchschnittlich 17,25 Teilnehmern gemessene Prognosefehler mit dem Faktor $\varepsilon(n)$ multipliziert wird.

Dies ist eine konservative Schätzung, da tatsächlich der Standardfehler des Mittelwerts nicht den gesamten Prognosefehler ausmacht. Diese Methode ignoriert, dass die Urteile

der verschiedenen Experten nicht um einen gemeinsamen Mittelwert streuen. Deswegen überschätzt diese Methode die Streuung der Urteile bei kleinen n. Dieser Kompromiss wird zunächst hingenommen. Die Ergebnisse dieser Berechnung stellen also eine konservative Schätzung dar.

Damit lässt sich die minimal notwendige Differenz zwischen dem Mittelwert der Expertenurteile und dem größten zulässigen Probandenmittelwert bestimmen, ab dem für jede Anzahl an Experten n zum Sicherheitsniveau 5% davon ausgegangen werden kann, dass der Mittelwert der Probandenurteile kleiner als der kritische Wert sein wird. Dies wird in Abbildung 16 dargestellt.

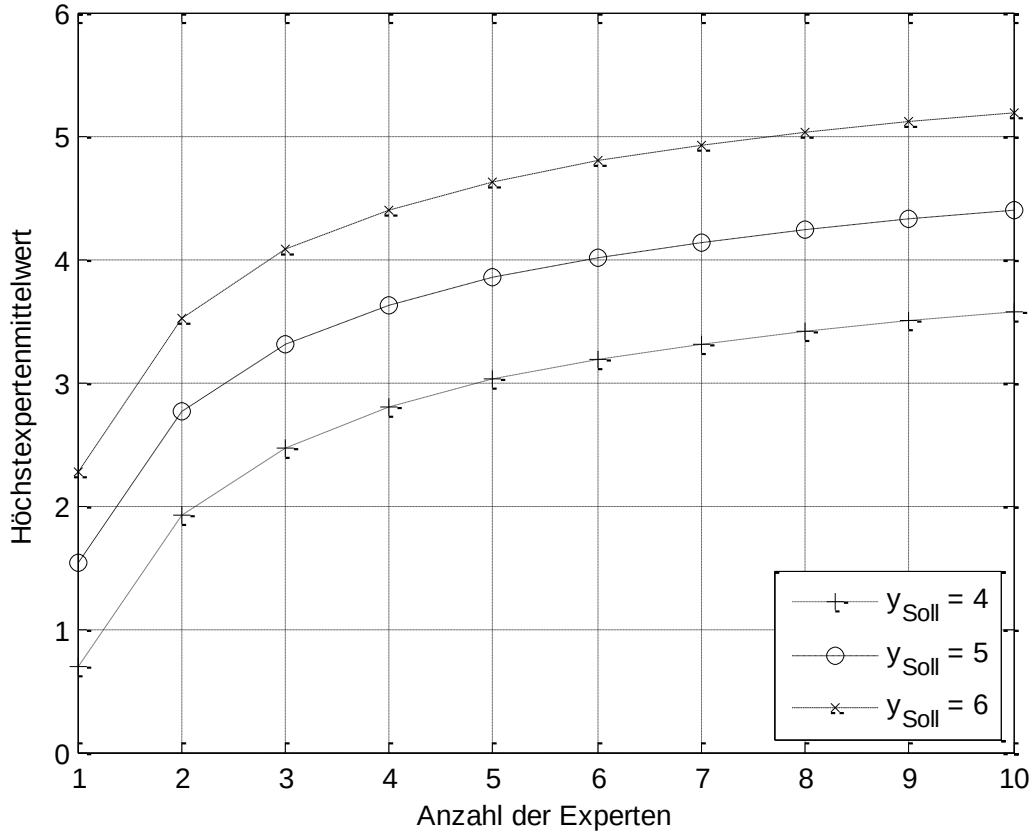

Abbildung 16 Auf Basis des Standardfehlers des Mittelwerts geschätzter höchster zulässiger Mittelwert der Urteile der Teilnehmer höherer Expertise über der Anzahl der Experten.

Unter gleichbleibenden Voraussetzungen lässt sich mit dem Datensatz aus den vier Versuchen auch bestimmen, ab welchen Mittelwerten der Urteile der Experten der Nachweis erbracht ist, dass der Mittelwert der Urteile der naiven Probanden über einem bestimmten Sollwert liegt. Anhand solcher Kurven kann der Nachweis erbracht werden, dass eine Situation mit einer Wahrscheinlichkeit von 95% zu kritisch ist und damit für

weitere Betrachtungen nicht in Betracht kommt. Abbildung 17 stellt dar, wie hoch der Mittelwert der Experten mindestens sein muss, um zu zeigen, dass der Mittelwert der Urteile der naiven Probanden über den eingezeichneten Grenzwerten liegt. Es sei darauf hingewiesen, dass die y-Skala gegenüber Abbildung 16 verschoben ist, um eine höhere Auflösung der Grafik zu erlauben. Zwischen den beiden Kurven gibt es einen Bereich, in dem aufgrund der Daten nicht entschieden werden kann, ob die Kritikalität über oder unter dem Grenzwert liegt. Dieser Bereich wird bei geringeren Anzahlen von Experten breiter.

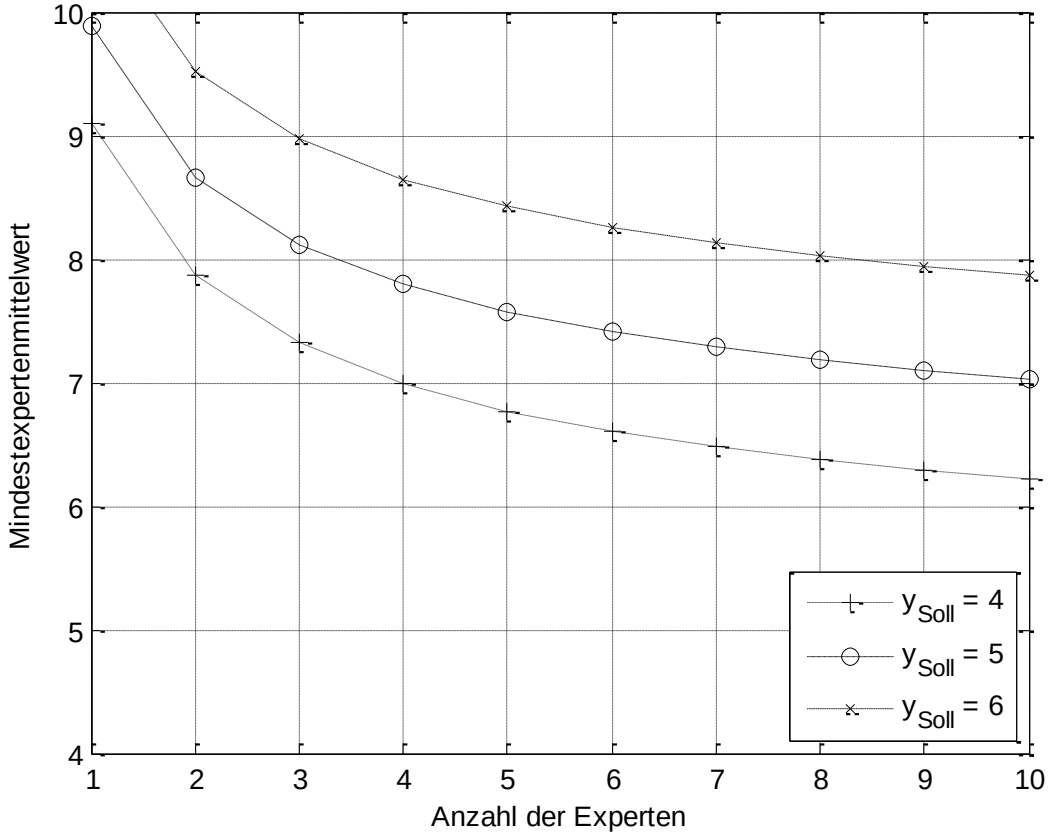

Abbildung 17 Auf Basis des Standardfehlers des Mittelwerts geschätzter geringster notwendiger Mittelwert der Urteile der Experten über der Anzahl der Experten.

Die Kurven in Abbildung 16 und 17 lassen sich durch Bootstrapping numerisch berechnen, um keine konservativen Annahmen über den Prognosefehler machen zu müssen. Dabei werden für jedes der 30 Szenarien k zufällige Expertenurteile mit zurücklegen gezogen und der Prognosefehler für dieses Kollektiv bestimmt. Durch häufige Wiederholung dieser Vorgehensweise und Mittelwertbildung nähert sich das Ergebnis dem wahren Wert an. So wird für jede Anzahl der Experten k der geringste zulässige Mittelwert der Expertengruppe berechnet, bei Unterschreitung dessen zum Signifikanzniveau

von 5% ausgeschlossen werden kann, dass der Mittelwert der Probandengruppe über 4, 5 oder 6 liegt. Das Ergebnis dieser Berechnung wird in Abbildung 18 dargestellt.

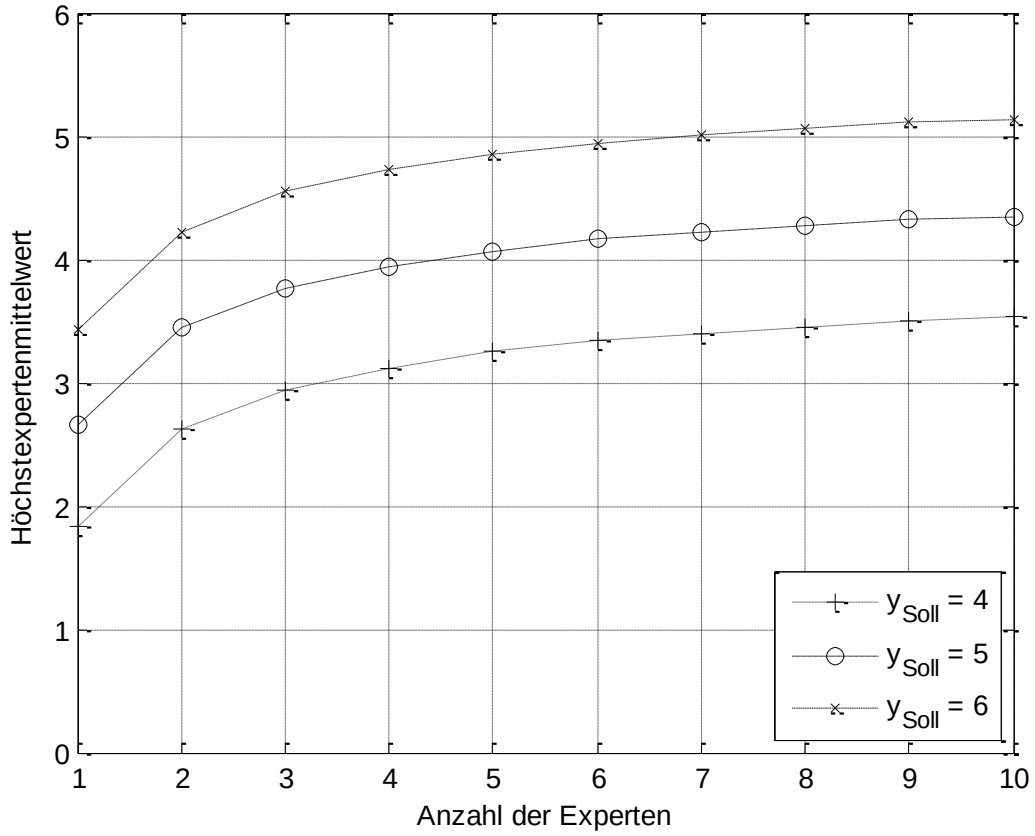

Abbildung 18 Numerisch berechneter höchster zulässiger Mittelwert der Urteile der Teilnehmer höherer Expertise über der Anzahl der Experten.

Beim Vergleich von Abbildung 16 und Abbildung 18 ist erkennbar, dass die Grenzwerte in Abbildung 18 höher liegen. Dies wird dadurch verursacht, dass diese Methode keine konservative Annahme über die Streuung der Urteile der einzelnen Experten machen muss, sondern die Grenzwerte durch wiederholtes Ziehen der Expertenurteile direkt quantifiziert werden. Dies führt insbesondere bei kleinen Anzahlen an Experten zu großen Unterschieden. So wurde durch die konservative Schätzmethode in Abbildung 16 bei zwei Experten ermittelt, dass der Mittelwert der Experten höchstens 2,76 sein darf, um mit 95% Sicherheit ausschließen zu können, dass der Mittelwert der naiven Probanden auf der SBS über 5 liegt. Durch die numerische Methode wurde gezeigt, dass dies bereits bei einem Expertenmittelwert von 3,50 der Fall ist. Bei mehr Experten ist die Differenz der Ergebnisse deutlich kleiner. Dies ist zu erwarten, da der Standardfehler

des Mittelwerts nicht der einzige Beitrag zur Streuung des Mittelwerts der Expertenurteile darstellt, sodass die Schätzmethode den negativen Effekt geringer Expertenanzahlen überschätzt. Auch bei der Betrachtung des geringsten notwendigen Mittelwerts der Urteile der Experten zum Nachweis eines Probandenmittelwerts über den Grenzwerten 4 bis 6 tritt der gleiche Effekt auf. Dies ist beim Vergleich von Abbildung 19 mit Abbildung 17 erkennbar.

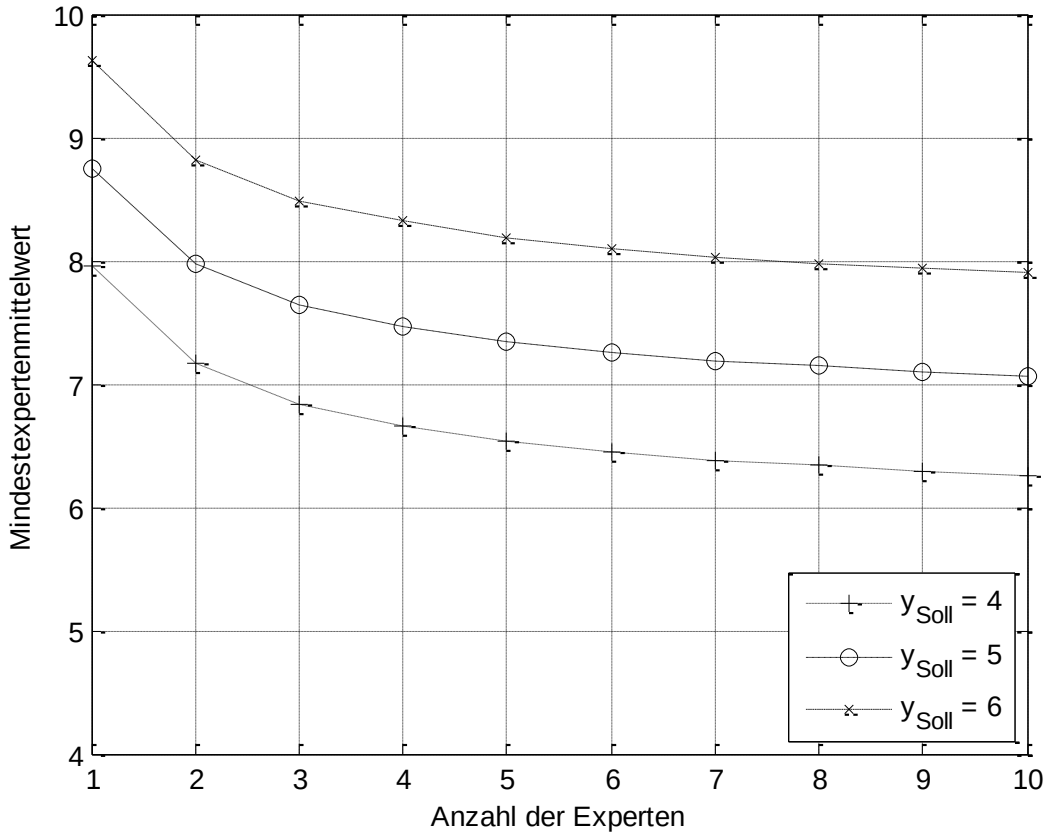

Abbildung 19 Numerisch berechneter geringster notwendiger Mittelwert der Urteile der Experten über der Anzahl der Experten.

4.8.2 Übertragungseffekte, Urteilstendenzen

Im vorherigen Abschnitt wurde davon ausgegangen, dass alle Experten ein gemeinsames Urteilsverhalten haben. In der Modellvorstellung sind sie also eine Stichprobe aus der Gesamtheit der Beherrschbarkeitsexperten und repräsentieren jeweils eine individuelle Instanz aus der Grundgesamtheit. Es wurde davon ausgegangen, dass alle Experten in ihrem Urteil zufällig um den Mittelwert der Expertengruppe streuen.

Jenseits dieser Modellvorstellung ist es möglich, dass die Experten nicht nur die Kritikalität der Fahrsituation bewerten, sondern auch ihre Einstellung zu kritischen Fahrsituationen im Allgemeinen zum Ausdruck bringen. In diesem Fall wären die Urteile einzelner Experten zwischen den Szenarien nicht unabhängig, sondern paarweise Korreliert. Manche Experten würden dazu tendieren, Szenarien überdurchschnittlich hoch zu bewerten und manche die Szenarien konsequent unterdurchschnittlich einstufen. Um dies zu überprüfen, wird eine paarweise Korrelation der Urteile für alle 30 Szenarien vorgenommen. Dadurch ergeben sich $\sum_1^{29} i = 435$ Hypothesen über paarweise Korrelationen der bis zu 21 Expertenurteile. Nach Bonferronikorrektur wird ein Signifikanzniveau von $\frac{0{,}05}{435}$ zu Grunde gelegt. Dadurch ergeben sich 21 signifikante Zusammenhänge zwischen den Szenarien, was 4,8% der betrachteten Hypothesen ausmacht und damit dem Erwartungswert entspricht. Die Daten stützen also nicht die Vermutung, dass die Experten lediglich ein individuelles Risikoakzeptanzniveau auf der Skala zum Ausdruck bringen, sondern tatsächlich die einzelnen Situationen getrennt voneinander betrachten.

Um diesen Effekt im Detail zu betrachten, werden exemplarisch die ersten vier Szenarien des ersten Teilversuchs betrachtet. Szenarien 1 und 4 bestanden aus Übernahmeszenarien aus dem hochautomatisierten Fahren und Szenarien 2 und 3 aus fehlerhaften Lenkmomenten beim teilautomatisierten Fahren. In Tabelle 33 und Tabelle 34 sind die paarweisen Korrelationen der Urteile für Experten und Teilnehmer aufgetragen.

Tabelle 33 Die Korrelationstabelle der ersten vier Szenarien des ersten Versuchs für die Teilnehmer geringer Expertise.

r	Szenario 1	Szenario 2	Szenario 3	Szenario 4
1	1	(n.s.)	(n.s.)	0.56***
2		1	0.55***	(n.s.)
3			1	(n.s.)
4				1

Tabelle 34 Die Korrelationstabelle der ersten 4 Szenarien des ersten Versuchs für die Experten.

r	Szenario 1	Szenario 2	Szenario 3	Szenario 4
1	1	0.57*	0.63**	0.56*
2		1	0.63**	0.46*
3			1	0.54*
4				1

In Tabelle 33 ist erkennbar, dass es einen signifikanten Zusammenhang zwischen den Urteilen in Szenario 2 und 3 sowie in Szenarien 1 und 4 gab. Dies sind die einzigen signifikanten Zusammenhänge und beide sind signifikant zum p = 0,001 Niveau. Szenarien 2 und 3 sowie 1 und 4 sind jeweils artgleich, sodass der signifikante Zusammenhang bedeutet, dass ein hohes Urteil in einem Lenkmomentszenario auch zu einem vergleichsweise hohen Urteil im nächsten Lenkmomentszenario geführt hat und das gleiche auch für Übernahmeszenarien gilt. Es wurde jedoch kein Einfluss der Urteile in den Lenkmomentszenarien auf die Urteile in den Übernahmeszenarien gefunden. Die Teilnehmer scheinen diese also getrennt zu bewerten. In Tabelle 34 ist erkennbar, dass es signifikante Zusammenhänge zwischen allen Szenarien gab. Der Mittelwert der signifikanten Korrelationen (erste 4 Szenarien) beträgt 0,54 für die Experten sowie 0,43 für die Probanden. Das heißt, während bei den Probanden nur etwa 18% der Varianz durch individuelle Urteilstendenzen erklärt werden, ist bei den Experten 29% der Varianz auf individuelle Effekte zurückzuführen. Erst bei Betrachtung der Datenreihe mit 30 Szenarien wird wie zuvor berichtet erkennbar, dass die Experten tatsächlich keine nennenswerten Übertragungseffekte zeigen. Dies verdeutlicht den Wert der Betrachtung über viele Szenarien hinweg.

4.8.3 Individuelle Auswertung

In diesem Abschnitt wird eine ähnliche Auswertung wie in 4.8.1 durchgeführt, jedoch wird hier davon ausgegangen, dass jeder Teilnehmer der Gruppe mit höherer Expertise eine individuelle Urteilsstrategie hat und die Urteile also nicht einfach über der Gruppe gemittelt werden dürfen. Deswegen wird im Folgenden für jeden Experten ein eigenes Modell des Urteilsverhaltens erstellt.

Die Teilnehmer der Gruppe mit höherer Expertise hatten nicht jeweils alle Fahrsituationen bewertet, da diese an vier Terminen über den Zeitraum von zwei Jahren präsentiert wurden. Deswegen kommt es zu einer je nach Teilnehmer unterschiedlichen Anzahl an fehlenden Werten. Die Anzahl der bewerteten Szenarien variiert dabei von mindestens 8 bis zu den vollen 30 von 30 Szenarien. Wie im vorangegangenen Abschnitt werden hier nur die 30 Szenarien betrachtet, die von allen Teilnehmern ohne Permutation erlebt wurden. Durchschnittlich hat jeder Teilnehmer höherer Expertise 23,7 Szenarien bewertet. Die fehlenden Werte werden im Folgenden behandelt, indem die Berechnung der Regressionsgeraden je Teilnehmer lediglich die tatsächlich erlebten Fahrsituationen berücksichtigt. Die nachfolgenden Berechnungen erfolgen also zunächst ohne Betrachtung der unterschiedlichen Anzahl der bewerteten Szenarien.

Abbildung 20 stellt die 21 Regressionsgeraden der Teilnehmer höherer Expertise dar, die den Zusammenhang der Urteile jedes Experten mit dem Mittelwert der Urteile der Teilnehmer geringer Expertise modelliert.

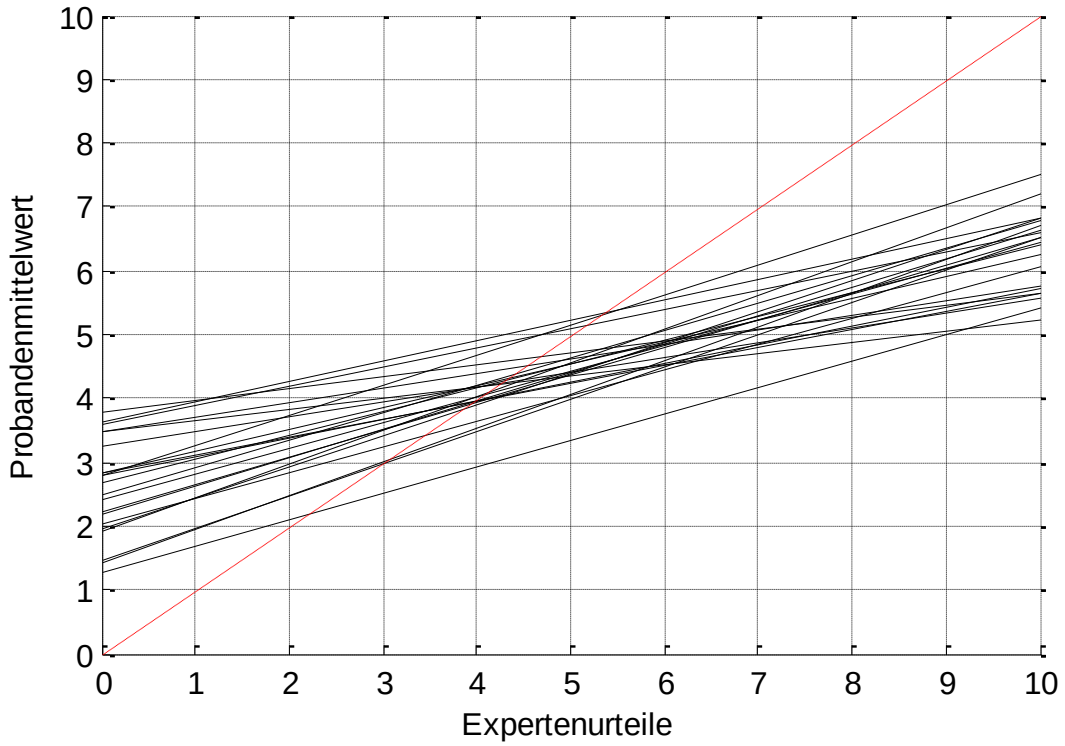

Abbildung 20 Regressionsgeraden der 21 Teilnehmer mit höherer Expertise.

Tabelle 35 enthält für jeden Teilnehmer die Anzahl der bewerteten Fahrsituationen n, den F- und p-Wert des linearen Modells, das zugehörige R^2 sowie die Abszisse und Steigung des resultierenden linearen Modells. P-Werte kleiner als 0,05 werden mit einem Stern gekennzeichnet. Nichtsignifikante Modelle sind grau hinterlegt.

Tabelle 35 Statistische Kennwerte der individuellen linearen Urteilsmodelle.

#	n	F	p	R^2	ε	Abszisse	Steigung
1	30	8,93	0,01*	0,24	1,49	2,27	0,43
2	16	3,03	0,10	0,18	1,10	2,79	0,29
3	8	3,60	0,11	0,37	1,37	2,07	0,50
4	22	27,83	0,00*	0,58	1,04	0,86	0,60
5	30	5,19	0,03*	0,16	1,66	3,39	0,24
6	8	3,59	0,11	0,37	0,55	3,50	0,17
7	30	27,77	0,00*	0,50	0,99	2,32	0,56
8	30	4,13	0,05	0,13	1,71	2,70	0,33
9	30	6,78	0,01*	0,19	1,58	3,26	0,36
10	30	34,62	0,00*	0,55	0,88	1,85	0,48
11	16	8,41	0,01*	0,38	0,83	2,22	0,43
12	30	5,44	0,03*	0,16	1,65	2,44	0,33
13	30	10,32	0,00*	0,27	1,44	1,77	0,43
14	16	1,78	0,20	0,11	1,18	3,26	0,23
15	30	6,46	0,02*	0,19	1,60	2,81	0,33
16	30	16,92	0,00*	0,38	1,22	2,37	0,43
17	8	0,76	0,42	0,11	0,78	1,29	0,41
18	22	1,31	0,27	0,06	1,96	3,44	0,21
19	30	13,76	0,00*	0,33	1,32	1,30	0,59
20	22	18,88	0,00*	0,49	1,27	0,70	0,60
21	30	8,89	0,01*	0,24	1,49	3,51	0,33

Tabelle 35 stellt die statistischen Kennzahlen der linearen Modelle der individuellen Expertenurteile in Abhängigkeit der durchschnittlichen Probandenurteile dar. Sie zeigt einige wichtige Zusammenhänge. So wurde für keinen der Teilnehmer, die lediglich 8 Fahrsituationen bewertet haben, das lineare Modell signifikant. Dies zeigt, dass für eine Modellierung des Urteilsverhaltens der Teilnehmer eine größere Datenmenge notwendig ist. Das erste signifikante lineare Modell wurde für Teilnehmer 11 gebildet, der insgesamt 16 Szenarien bewertet hat. Die Urteile der zwei anderen Teilnehmer mit der gleichen Anzahl der bewerteten Szenarien – Nummer 2 und 14 – erzeugten jedoch kein signifikantes Modell. Für 11 von 12 Teilnehmer, von denen 30 Urteile für die Modellbildung zur Verfügung standen, wurde ein signifikantes lineares Modell gefunden. Lediglich die Urteile von Teilnehmer 8 führten nicht zu einem signifikanten Modell, obwohl dieser 30 Szenarien bewertet hat.

Das Bestimmtheitsmaß R^2 zeigt, dass es erhebliche Unterschiede in der Qualität des Modells zwischen den Teilnehmern gibt. So variiert der Anteil der durch das Modell erklärten Varianz zwischen 58% bei Teilnehmer 4 und 16% bei Teilnehmer 5 und 12. Dies spricht dafür, dass nicht alle Experten im gleichen Maße die Urteile der naiven

Probanden vorhersagen. Diese Feststellung wird später in diesem Unterkapitel für die Optimierung der Ergebnisse genutzt werden. Bemerkenswert ist weiterhin, dass alle gebildeten linearen Modelle Steigungen unterhalb von 1 aufweisen. Tatsächlich liegen die Steigungen zwischen 0,24 und 0,60. Gleichzeitig sind jedoch alle ermittelten Abszissen positiv. Dies ist die numerische Repräsentation der qualitativen Beobachtung, dass die Teilnehmer höherer Expertise im Bereich geringer Kritikalität zwischen drei und fünf nahe an den Urteilen der Teilnehmer geringer Expertise liegen, jedoch bei höherer Kritikalität dazu neigen, die Kritikalität zu hoch einzuschätzen. In anderen Worten bedeutet dies, dass eine Erhöhung des Urteils eines Teilnehmers höherer Expertise von 4 auf 8 nur eine Erhöhung des Mittelwerts der Urteile der Teilnehmer geringerer Expertise um etwa 1 bis 2 Punkte auf der SBS anzeigt.

Da jeder Experte im Durchschnitt 23,7 Szenarien bewertet hat, von denen die subjektive Kritikalitätsbewertung durch naive Probanden bekannt ist, können die untersuchten Experten individuell bezüglich ihrer Sensitivität und Spezifität beschrieben werden. Dazu muss zunächst ein Kriterium für die Probandenurteile festgelegt werden, ab dem ein Fahrszenario als zu kritisch bewertet wird. In (Neukum & Krüger, 2003) wird hierfür die Kategorie „Gefährlich“ auf der SBS verwendet. Wenn ein oder mehrere Probanden das Szenario als gefährlich, also höher als 6 auf der SBS, einstufen, wird das Szenario als nicht tolerabel eingestuft. Dieses Kriterium wird nun auf jeden Experten angewendet, um festzustellen, ob er das Szenario als zu kritisch einstuft. Damit ergeben sich die Sensitivitäten q, Spezifitäten r, positive Likelihood-Ratio LR+ und negative Likelihood-Ratio LR- in Tabelle 36. In der Tabelle ist zunächst auffällig, dass für die meisten Experten die positive Likelihood-Ratio nicht definiert ist. Dies liegt daran, dass sie kein Szenario, dass von den Probanden unkritisch bewertet wurde, selbst als zu kritisch eingestuft haben. Damit ist die Wahr-Negativ-Rate oder Spezifität gleich 1 und damit der Quotient der positiven Likelihood-Ratio gleich 0. Deswegen kann LR+ nicht angegeben werden. Weiterhin kann in Tabelle 36 beobachtet werden, dass für einige Experten die Spezifität r nicht definiert ist. Diese Experten haben keine Szenarien bewertet, die nach dem oben beschriebenen Kriterium durch die Probanden als nicht-kritisch betrachtet wurden. Deswegen kann die Spezifität oder Wahr-Negativ-Rate nicht bestimmt werden. Dies betrifft die Experten 6, 17 und 18. Weiterhin hat Experte 12 nach diesem Kriterium kein Szenario als unkritisch bewertet, das auch von den Probanden als unkritisch eingestuft wurde (r = 0,00). Deswegen ist in diesem Fall die negative Likelihood-Ratio nicht definiert.

Die beobachtete Sensitivität für die subjektive Kritikalität nach dem oben genannten Kriterium schwankt bei den untersuchten Experten stark. Experte 9 hat eine Wahr-Positiv-Rate von nur 0,04 während Experte 13 eine Wahr-Positiv-Rate von 0,77 erreicht. Für Experte 17 wurde q von 1,00 berechnet, da dieser nur 8 Szenarien bewertet hat (siehe Tabelle 35). Die Spezifität r ist bei dem untersuchten Kriterium entweder 0, 1 oder undefiniert. Dies liegt daran, dass mit dem gewählten Kriterium 29 von 30 Szenarien als intolerabel kritisch bewertet worden sind. Deswegen liegen nicht genug Daten vor, um die Wahr-Negativ-Rate zu bestimmen. Die negative Likelihood-Ratio schwankt, in Abhängigkeit von der gemessenen Sensitivität, ebenfalls stark. Wegen der schlecht definierten Spezifität ist diese für das hier untersuchte Kriterium nicht interpretierbar.

Tabelle 36 Sensitivität, Spezifität und Likelihood-Ratios der Experten.

#	q	r	r+q	LR+	LR-
1	0,31	1,00	1,31	-	0,69
2	0,40	1,00	1,40	-	0,60
3	0,14	1,00	1,14	-	0,86
4	0,57	1,00	1,57	-	0,43
5	0,31	1,00	1,31	-	0,69
6	0,50	-	-	-	-
7	0,10	1,00	1,10	-	0,90
8	0,31	1,00	1,31	-	0,69
9	0,04	1,00	1,04	-	0,97
10	0,38	1,00	1,38	-	0,62
11	0,27	1,00	1,27	-	0,73
12	0,59	0,00	0,59	0,59	-
13	0,66	1,00	1,66	-	0,35
14	0,40	1,00	1,40	-	0,60
15	0,38	1,00	1,38	-	0,62
16	0,28	1,00	1,28	-	0,72
17	1,00	-	-	-	-
18	0,27	-	-	-	-
19	0,28	1,00	1,28	-	0,72
20	0,57	1,00	1,57	-	0,42
21	0,07	1,00	1,70	-	0,93

Die Berechnungen für Tabelle 36 stützten sich auf ein Kriterium, dass die Anzahl der naiven Probanden nicht berücksichtigte. Da für die hier betrachteten Szenarien jedoch durchschnittlich 42 Probandenurteile vorliegen, führt dies, zusammen mit der starken Streuung der Urteile auf der SBS, zu einer geringen Anzahl an als unkritisch eingestuften Szenarien, sodass die Auswertung der Sensitivität und Spezifität der Experten keine

interpretierbaren Daten ergibt. Um dies zu korrigieren, wird die gleiche Auswertung mit einem alternativen Kriterium wiederholt, das die betrachteten Szenarien in die kritischsten 15 und unkritischsten 15 Szenarien aufteilt. Dies wird anhand des Mittelwerts der Urteile der Probanden durchgeführt. Liegt der Mittelwert der Urteile der Probanden über 4,7, wird das Szenario als kritisch betrachtet, andernfalls als unkritisch. Es ist zu bemerken, dass dieses alternative Kriterium damit die Interpretation der Daten verändert. Die Urteile der Experten werden nun so verstanden, dass ein Wert größer 6 auf der SBS als Schätzung verstanden wird, dass das bewertete Szenario zu der kritischeren Hälfte der Szenarien in der Versuchsreihe gehört. Die Sensitivität, Spezifität und Likelihood-Ratio unter Berücksichtigung des alternativen Kriteriums für die Festlegung der wahren Kritikalität können Tabelle 37 entnommen werden. Dort sind trotz der Änderung des Kriteriums einige ähnliche Effekte wie in Tabelle 36 zu beobachten. Für Experten 3, 7, 9, 11 und 19 ergibt sich ein r von 1. Dies liegt jetzt jedoch nicht an einer zu geringen Datenbasis, sondern daran, dass sie tatsächlich eine Wahr-Negativ-Rate von 100% erreicht haben. Experte 3 hat 4 von 4 unkritischen Szenarien mit 6 oder weniger auf der SBS bewertet, Experten 7 und 9 15 von 15 Szenarien und Experte 11 8 von 8 Szenarien, also jeweils die Hälfte der von den Experten bewerteten Szenarien. Dies schränkt die Interpretierbarkeit der Daten deswegen nicht ein, wobei gerade bei Experte 3 weiterhin beachtet werden muss, dass er in dieser Versuchsreihe nur 8 Bewertungen abgegeben hat. Die hier betrachteten Kennwerte spiegeln nicht die Größe der Datenbasis wider, anhand der sie berechnet worden sind, und treffen keine Aussage über die statistischen Streubreiten der Werte.

Anhand der Daten in Tabelle 37 können die Unterschiede in den Urteilscharakteristiken der Experten nun erkannt werden. Experte 4 fällt auf, da bei ihm die Summe aus Sensitivität und Spezifität mit 1,73 am höchsten ist. Sein individuelles Urteil ist also am besten dazu geeignet, um festzustellen, ob das betrachtete Szenario zu den kritischeren oder weniger kritischen Szenarien gehört. Die Urteile von Experten 17 und 21 sind hingegen nicht geeignet, diese Schätzung vorzunehmen. Bei ihnen ist die Summe aus der Sensitivität und der Spezifität gerade 1,00. Das bedeutet, dass ihre Urteile keine nutzbaren Informationen über die Kritikalität der Szenarien enthalten. Experte 6 hat gleiche Sensitivität und Spezifität von jeweils 0,75. Seine Wahr-Positiv-Rate ist also genauso hoch wie die Wahr-Negativ-Rate. Er stellt für das alternative Kriterium einen erwartungstreuen Schätzer dar. Bei Experte 2 wurde bei gleicher Summe r+q ein anderes Verhältnis dieser Werte festgestellt. Seine Urteile sind besser dazu geeignet, weniger kritische Szenarien zu entdecken, aber weniger geeignet, besonders kritische Szenarien zu finden. Wenn

Wert darauf gelegt wird, dass ein Urteiler mit großer Wahrscheinlichkeit alle kritischen Szenarien selektiert, würde eine Charakteristik wie bei Experte 20 geeignet sein. Er verfügt über eine höhere Wahrscheinlichkeit, tatsächlich zur kritischeren Hälfte der Szenarien gehörende Situationen als solche zu erkennen (q = 0,82). Dies führt, aufgrund der gleichen Summe aus q und r wie bei den zwei zuvor betrachteten Experten, jedoch zu einer vergleichsweise geringen Wahr-Negativ-Rate r von 0,73.

Wie an den Beispielen diskutiert, können diese Kennwerte genutzt werden, um die Fähigkeit und Urteilscharakteristik der Experten numerisch zu beschreiben. Diese Informationen können verwendet werden, um Experten für bestimmte Aufgaben auszuwählen.

Tabelle 37 Sensitivität, Spezifität und Likelihood-Ratios der Experten mit alternativem Kriterium.

#	q	r	r+q	LR+	LR-
1	0,31	0,87	1,33	3,50	0,62
2	0,63	0,88	1,50	5,00	0,43
3	0,25	1,00	1,25	-	0,75
4	0,90	0,82	1,73	5,00	0,11
5	0,40	0,80	1,20	2,00	0,75
6	0,75	0,75	1,50	3,00	0,33
7	0,20	1,00	1,20	-	0,80
8	0,53	0,93	1,47	8,00	0,50
9	0,07	1,00	1,07	-	0,93
10	0,37	0,93	1,60	10,00	0,36
11	0,50	1,00	1,50	-	0,50
12	0,87	0,67	1,53	2,60	0,20
13	0,93	0,67	1,60	2,80	0,10
14	0,50	0,75	1,25	2,00	0,67
15	0,60	0,87	1,47	4,50	0,46
16	0,47	0,93	1,40	7,00	0,57
17	1,00	0	1,00	1,00	-
18	0,46	0,90	1,36	5,00	0,60
19	0,53	1,00	1,53	-	0,47
20	0,82	0,73	1,55	3,00	0,25
21	0,07	0,93	1,00	1,00	1,00

Um zu überprüfen, ob die Unterschiede zwischen den Experten tatsächlich durch individuelle Bewertungsstrategien verursacht worden sind, werden diese nun einem paarweisen Vergleich der Urteile unterzogen. Für die 21 Experten gibt es 210 mögliche Paare, die verglichen werden können. Davon sind nur 203 Paarvergleiche zulässig, da

es in den übrigen sieben keine Szenarien gibt, die von beiden Experten bewertet wurden. Diese Vergleiche wurden mittels paarweiser t-Tests durchgeführt.

Tabelle 38 stellt die Ergebnisse dieser Vergleiche dar. Signifikante Unterschiede zum Bonferroni-korrigierten Signifikanzniveau von $\frac{0,05}{203}$ sind durch ein dunkles Feld dargestellt. Helle Zellen zeigen nicht signifikante Ergebnisse an. Insgesamt wurden bei 33 Paarvergleichen signifikante Unterschiede zwischen den Urteilen der Experten bzgl. der Situationen festgestellt. Dies sind 16,26% der Hypothesen, also mehr als bei einem Signifikanzniveau von 5% durch Zufall zu erwarten gewesen wären. Der Berechnung lagen im Durchschnitt 19,24 Urteile je Paarvergleich zu Grunde. 59 bzw. 29,06% aller Paarvergleiche stützten sich nur auf 8 gemeinsame Urteile. Wenn nur die 13 Experten betrachtet werden, die jeweils 30 Szenarien bewertet haben, finden sich nach Bonferroni-Korrektur 18 signifikant verschiedene Paare bei insgesamt 66 möglichen Paarungen. Dies macht 27,3% der möglichen Paarungen aus. Für die Probanden wurden, als Methodencheck, nach Bonferroni-Korrektur in keinem der vier Versuche mehr als 5% signifikante Unterschiede im Urteilsverhalten festgestellt. Damit wird bestätigt, dass es mehr Unterschiede zwischen den Urteilen der Experten gibt, als durch Zufall zu erwarten gewesen wäre.

Tabelle 38 Signifikante Unterschiede zwischen individuellen Experten nach Bonferroni-Korrektur.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1	-																				
2	-	-																			
3	-	-	-																		
4	-	-	-	-																	
5	-	-	-	-	-																
6	-	-	-	-	-	-															
7	-	-	-	-	-	-	-														
8	-	-	-	-	-	-	-	-													
9	-	-	-	-	-	-	-	-	-												
10	-	-	-	-	-	-	-	-	-	-											
11	-	-	-	-	-	-	-	-	-	-	-										
12	-	-	-	-	-	-	-	-	-	-	-	-									
13	-	-	-	-	-	-	-	-	-	-	-	-	-								
14	-	-	-	-	-	-	-	-	-	-	-	-	-	-							
15	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-						
16	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-					
17	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-				
18	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-			
19	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
20	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
21	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Die bisherigen Ergebnisse über das Urteilsverhalten von Expertengruppen in Abschnitt 4.8.1 haben sich auf Gruppen bezogen, die durch die Anwesenheit der individuellen Probanden zu den Versuchen zustande gekommen sind. Die Anzahl und Auswahl der Experten wurde also bisher nicht modifiziert. Im Folgenden soll untersucht werden, welche Auswirkung die Rekrutierung einer Teilmenge der Experten auf die Ergebnisse hat.

Zu diesem Zweck werden neue Expertengruppen auf Basis der Daten der individuellen Experten erstellt. Die Eigenschaften dieser neuen Gruppen können dann mit der ursprünglichen Gruppe verglichen werden. Bei der Betrachtung der Experten als homogene Gruppe wurde durch Kreuzvalidierung ein mittlerer quadrierter Fehler (engl. Mean Square Error, MSE) von 0,40 festgestellt. Dieser Wert bezieht sich auf die Abweichung des Mittelwerts aller verfügbaren Experten – im Durchschnitt 17,25 je Szenario – vom Mittelwert der Probandenurteile. Eine geschickte Auswahl der Experten kann nun jedoch zu einer Verringerung dieser Abweichung führen.

Als praktikable Größe werden zunächst Gruppen von je fünf Experten ausgewählt. Bei insgesamt 21 untersuchten Experten ergeben sich somit 20359 mögliche Gruppen mit

je 5 Experten, ohne, dass ein Experte in einer Gruppe zweimal vorkommt, sowie ohne Permutation der Reihenfolge. 5710 Gruppen wurden aus der Auswertung ausgeschlossen, da sie weniger als 8 Szenarien gemeinsam bewertet haben. Abbildung 21 stellt die Verteilung des durch 10-fache Kreuzvalidierung geschätzten mittleren quadrierten Fehlers als Histogramm dar. In den Gruppen, in denen nur 8 Szenarien gemeinsam bewertet wurden, wurde eine 8-fach Kreuzvalidierung bzw. Leave-One-Out-Kreuzvalidierung angewendet. Der mittlere quadratische Fehler bei einfacher Mittelwertbildung über alle 21 Experten ist als durchgezogene vertikale Linie eingezeichnet. Der mittlere Fehler bei Auswahl von 5 zufälligen Experten als gestrichelte vertikale Linie. 22,9% der Fünfergruppen weisen einen Fehler auf, der unterhalb des Fehlers bei Verwendung aller Experten liegt. Von diesen 3121 Gruppen haben jedoch 2975 oder 95,32% nur acht Szenarien bewertet. Da aber die Experten, die nur acht Szenarien bewertet haben, nach Tabelle 35 nicht über erkennbar bessere R^2 oder ε als andere Experten verfügen, ist an diesen Zahlen lediglich der Effekt einer zu geringen Szenarienanzahl erkennbar.

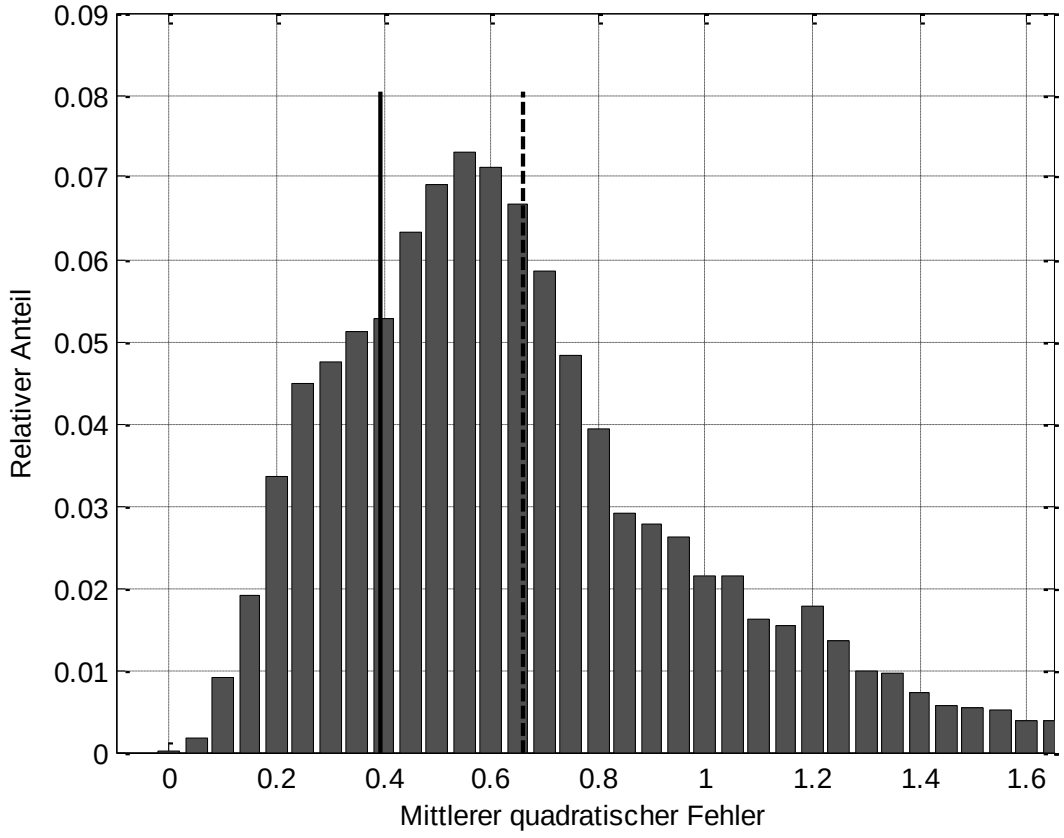

Abbildung 21 Histogramm des mittleren quadrierten Fehlers bei Expertengruppen mit je fünf Mitgliedern.

Das Zehnerperzentil des mittleren quadratischen Fehlers beträgt 0,26. In diesen 2035 Gruppen mit je fünf Experten finden sich nicht alle Experten gleich oft. Bestimmte Experten treten deutlich häufiger auf als andere. Abbildung 22 stellt ein Histogramm der Häufigkeit des Auftretens einzelner Experten in Fünfergruppen mit einem geringeren mittleren quadrierten Fehler als unter Verwendung aller 21 Experten dar. Dort ist erkennbar, dass beispielsweise Experte 6 in über 55% dieser Expertengruppen vorkommt, Experte 20 jedoch nur in 0,7%. Dies lässt sich mit dem Unterschied dieser Experten im Fehlerterm ε der linearen Regresssion in Tabelle 35 erklären. Während der Fehlerterm bei Experte 6 nur 0,55 beträgt, liegt er bei Experte 20 mit 1,27 besonders hoch.

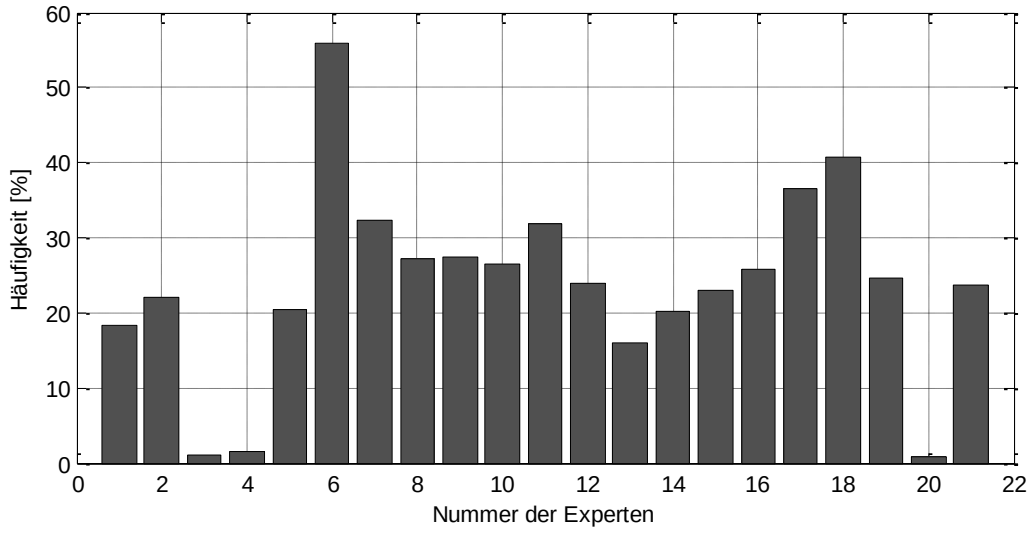

Abbildung 22 Häufigkeit des Auftretens einzelner Experten in Fünfergruppen mit einem geringeren geschätzten MSE als unter Verwendung aller Experten.

Wenn Experten ausgeschlossen werden, die nur 8 Szenarien bewertet haben und eine Mindestanzahl von 14 bewerteten Szenarien angesetzt wird, ergibt sich für das Histogramm der mittleren quadrierten Fehler Abbildung 23. Der mittlere quadrierte Fehler wurde wieder durch 10-fache Kreuzvalidierung geschätzt.

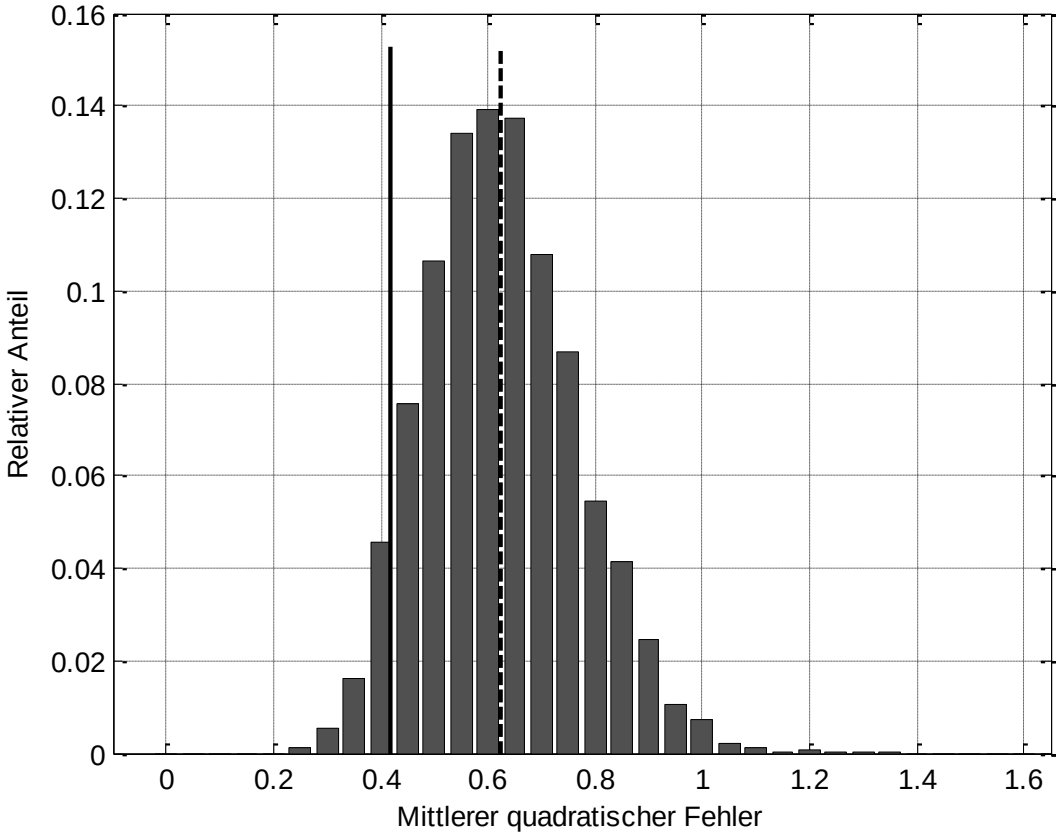

Abbildung 23 Mittlerer quadratischer Fehler bei 5 Experten je Gruppe und mindestens 14 bewerteten Szenarien je Gruppe.

Mittels einer schwarzen vertikalen Linie ist der durch 10-fache Kreuzvalidierung geschätzte Fehler der 21er-Gruppe von Experten eingetragen, der ca. 0,385 beträgt. Mittels der gestrichelten vertikalen schwarzen Linie ist der Mittelwert der MSEs der Fünfergruppen mit mindestens 14 bewerteten Szenarien aufgetragen. Dieser liegt bei 0,624. Lediglich 2,92% der Gruppen mit 5 Experten, 192 Stück, besitzen einen MSE geringer als der Erwartungswert bei Verwendung aller 21 Experten. Die Häufigkeit des Auftretens einzelner Experten in dieser Untergruppe wird in Abbildung 24 als Balkendiagramm dargestellt. Erkennbar ist, dass die Experten, die nur 8 Szenarien bewertet haben, nicht in den betrachteten Szenarien auftauchen. Weiterhin erkennbar ist, dass es zu starken Unterschieden in der Auftretenshäufigkeit einzelner Experten kommt. Insbesondere Experte 10 sticht in dieser Darstellung hervor, da er in über 80% der ausgewählten Fünfergruppen vertreten ist. Dies lässt sich mit Blick auf Tabelle 35 begründen. Zunächst hat Experte 10 alle 30 Szenarien bewertet und tritt alleine deswegen häufiger auf als Experten, die weniger Szenarien bewertet haben. Weiterhin ist bereits bei der Auswertung in Tabelle 35 aufgefallen, dass diese Person ein hohes R^2 mit einem niedrigen ε

verbindet. Dies wirkt sich positiv auf den geschätzten MSE aus und führt somit zu einem häufigen Auftreten in den besten Fünfergruppen.

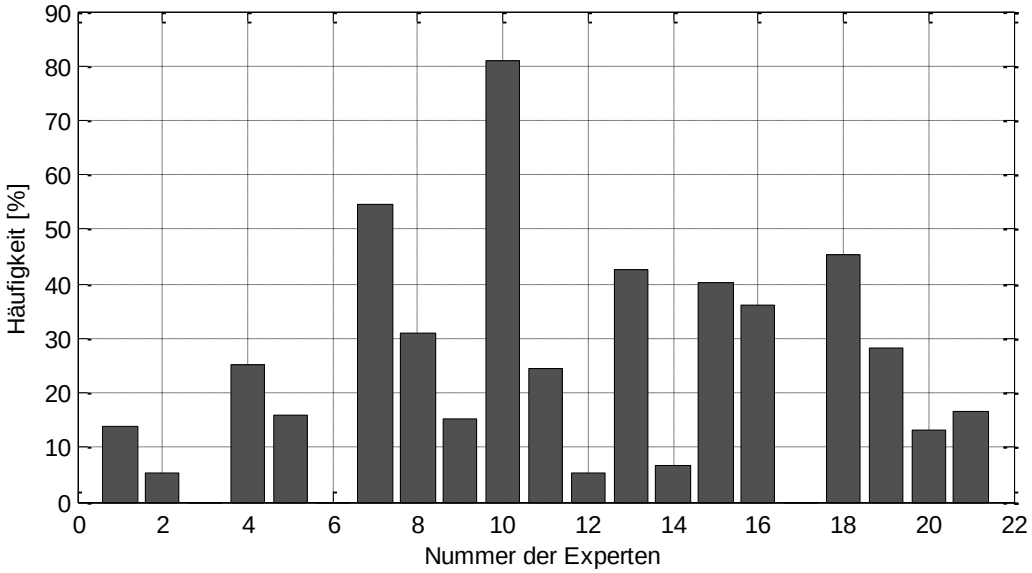

Abbildung 24 Häufigkeit des Auftretens einzelner Experten in Fünfergruppen mit einem geringeren geschätzten MSE als unter Verwendung aller 21 Experten.

Diese Daten wurden durch einfache Mittelwertbildung der Urteile der einzelnen Experten ermittelt. Die Abweichung zwischen dem gemittelten Expertenurteil und dem realen Wert der naiven Probanden hängt in diesem Fall stark von der Auswahl der Experten ab. Dies ist an den großen Unterschieden in der Häufigkeit des Auftretens einzelner Experten in den Gruppen geringerer mittlerer Quadratsummenfehler in Abbildung 24 erkennbar. Bei bekannter Abweichung der einzelnen Experten von den Probandenurteilen kann eine Gewichtung der Experten innerhalb der Gruppen vorgenommen werden. Nach dem zentralen Grenzwertsatz wird die Schätzung mit der geringsten Gesamtstreuung durch die folgende Formel gegeben (Einicke, 2012; Maybeck, 1979).

$$Cr_{Gruppe} = \sigma_{Gesamt}^2 \sum_{i=1}^{k} \sigma_i^{-2} Cr_i$$

Wobei Cr_i das Kritikalitätsurteil des Experten i darstellt, k die Anzahl der Experten in der Gruppe ist, σ_i die geschätzte Standardabweichung des Experten i und σ_{Gesamt}^2 die geschätzte Standardabweichung über alle Expertenurteile ist. Abbildung 25 stellt das Histogramm der mittleren quadrierten Fehler für Fünfergruppen mit mindestens 14 bewerteten Szenarien unter Verwendung von Gewichtung dar.

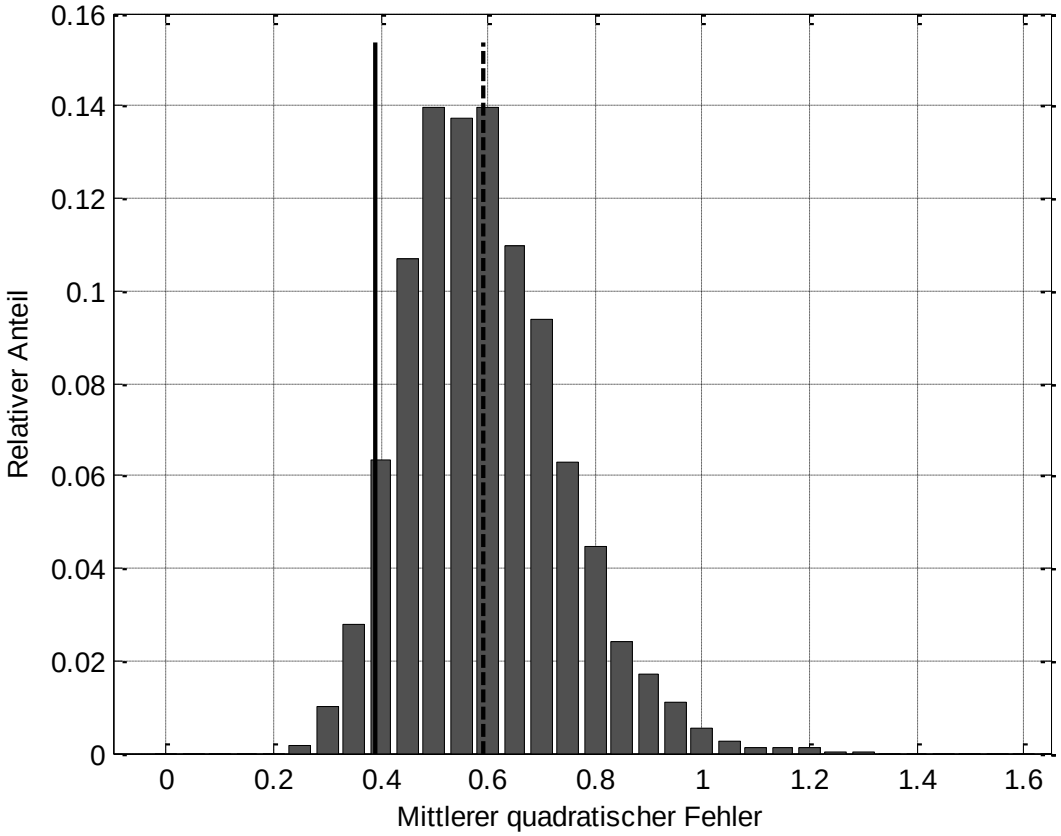

Abbildung 25 Mittlerer quadratischer Fehler bei 5 gewichteten Experten je Gruppe und mindestens 14 bewerteten Szenarien je Gruppe.

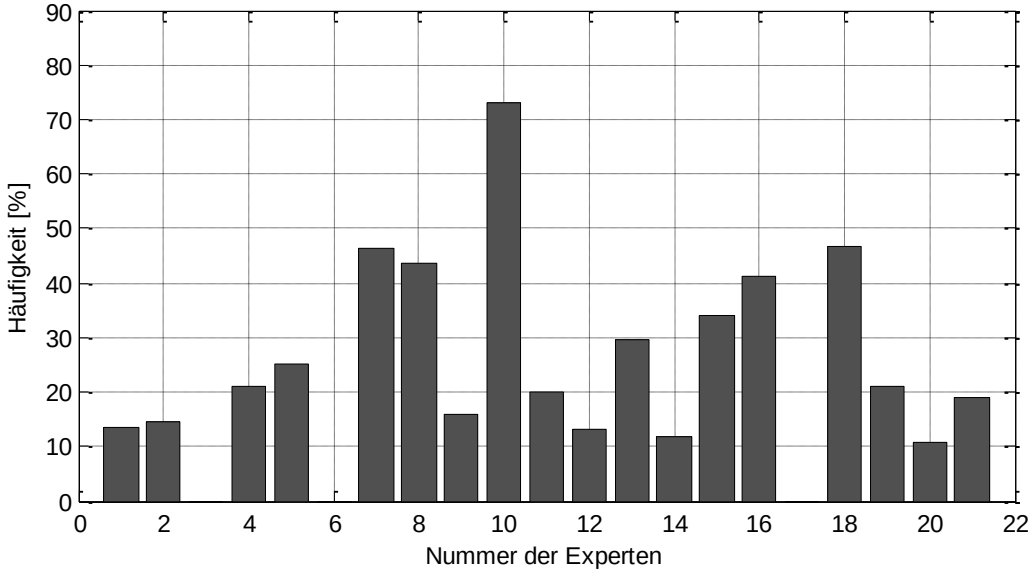

Abbildung 26 Häufigkeit des Auftretens einzelner Experten in gewichteten Fünfergruppen mit einem geringeren geschätzten MSE als unter Verwendung aller Experten.

Nach der Gewichtung der Experten können neue Konfidenzintervalle bestimmt werden, um den praktischen Nutzwert dieser Maßnahme zu beurteilen. Dazu werden für die besten 6,1% der Fünfergruppen, deren MSE geringer als bei Verwendung aller Experten lag, erneut die Konfidenzintervalle wie bereits in Abschnitt 4.8.1 bestimmt. Erneut werden für den Vergleich die Grenzwerte 4, 5 und 6 gewählt. Abbildung 27 stellt die obere Grenze des Mittelwerts der ausgewählten Expertengruppen dar, wenn mit 95% Konfidenz ausgeschlossen werden soll, dass der Mittelwert der Probanden oberhalb des Grenzwerts liegt. Bei allen Grenzwerten ist der geringste zulässige Mittelwert der Experten mit Gewichtung größer als ohne Gewichtung. Die Differenzen an den Grenzwerten sind in Tabelle 39 ablesbar. Die Unterschiede liegen jedoch im einstelligen Prozentbereich. Damit sind sie zwar größer als der Monte-Carlo-Fehler durch die verwendete Kreuzvalidierung, jedoch geringer als etwa der Effekt eines zusätzlichen Experten in der Gruppe.

Tabelle 39 Vergleich des geringsten zulässigen Mittelwerts der Fünfergruppen aus Experten mit und ohne Gewichtung.

Grenzwert	4	5	6
Ohne Gewichtung	3,39	4,20	4,99
Mit Gewichtung	3,44	4,41	5,33
Differenz	0,06	0,20	0,34
Relativ	1,7%	4,6%	6,4%

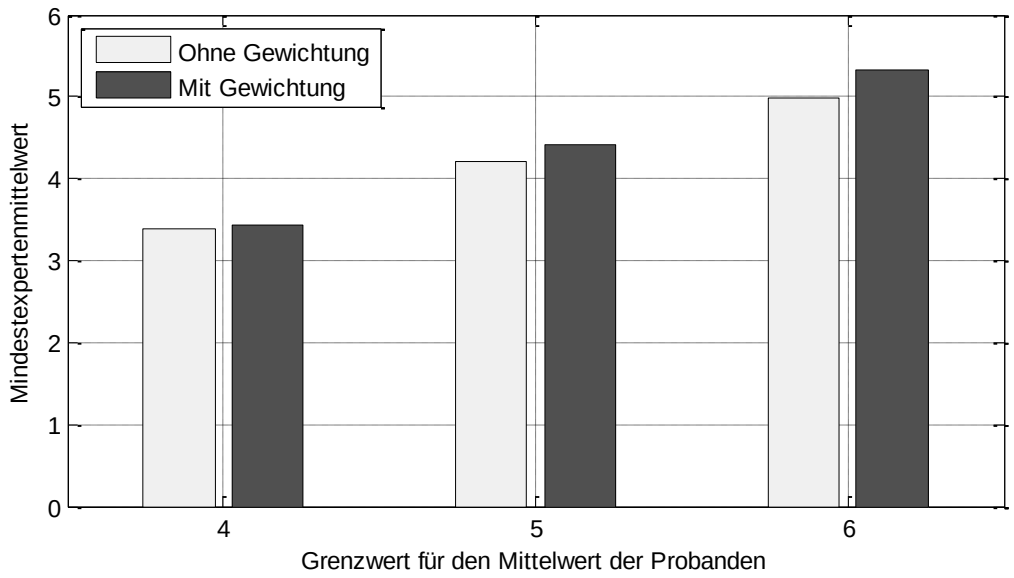

Abbildung 27 Geringster zulässiger Mittelwert der Experten bei Gruppen aus 5 Experten. Vergleich zwischen gewichteter und ungewichteter Mittelwertbildung.

5 Diskussion

Im vorangegangenen Kapitel wurden die Ergebnisse von insgesamt vier Studien beschrieben und eine Auswertung der Einzelversuche sowie der gesamten Versuchsreihe dargestellt. Dieses Kapitel widmet sich der Diskussion und Interpretation dieser Ergebnisse und zielt auf eine Einordung der einzelnen Teilergebnisse in den Kontext der Dissertation.

Bereits anhand der Daten des ersten Versuchs konnte gezeigt werden, dass das Fahrverhalten von naiven Probanden und Teilnehmern mit höherer Expertise unterschiedlich ist. Dieses Ergebnis war zu erwarten gewesen, da in der Expertengruppe mehr Fahrer mit erhöhter Fahrausbildung vertreten waren. Die gleichen objektiven Kritikalitätsmaße, wie sie bei Beherrschbarkeitsstudien mit naiven Probanden verwendet werden, zeigten bei Teilnehmern mit höherer Expertise keinen Zusammenhang mit der Kritikalität der Fahrsituation. Daraus folgt, dass es im Allgemeinen nicht möglich ist, von objektiven Beherrschbarkeitskriterien, die mittels Experten gemessen worden sind, auf die objektiven Kriterien der Beherrschbarkeit von naiven Probanden zu schließen. Bei einzelnen Experten kann dies möglich sein, jedoch nicht ohne eine genaue Analyse des jeweiligen Situationshergangs.

Ebenso konnte gezeigt werden, dass Teilnehmer mit höherer Expertise die subjektive Störung in den kritischen Fahrsituationen anders beurteilen als naive Probanden. Dies konnte über alle vier Versuche gezeigt werden. Aufgrund der Unterschiede im Fahrverhalten, das durch die objektiven Kriterien nachgewiesen wurde, ist dieser Unterschied in der subjektiven Störungsbeurteilung zunächst nicht verwunderlich. Die Störungsbeurteilung bezieht sich schließlich immer auf den erlebten Verlauf der Situation, selbst wenn der Teilnehmer versucht, auf naive Probanden zu schließen. Hier zeigte sich jedoch keine regellose Abweichung, wie sie bei den objektiven Kriterien beobachtet wurde, sondern eine systematische Abhängigkeit. Schon anhand der Daten des ersten Versuchs wurde gezeigt, dass die Urteile naiver Probanden durch die Störungsurteile der Experten geschätzt werden konnten. Dieser Zusammenhang wurde bei der Auswertung der gesamten Versuchsreihe quantifiziert. Die Urteile der betrachteten 21 Experten liegen im Durchschnitt 0,75 Skalenpunkte auf der SBS über denen der Probanden und weisen keine signifikant verschiedene Streuung der Urteile auf.

Die Tatsache, dass bei über 30 betrachteten Szenarien kein Effekt der Beherrschbarkeitsexpertise auf die Streuung der Urteile über die Störung in kritischen Fahrsituationen beobachtet wurde, lässt den Schluss zu, dass die Experten den gleichen oder einen ähnlichen Mechanismus für die Bewertung der Störung verwenden. Da beide Gruppen auf der gleichen Skala urteilen, kann dies als plausibel gelten. Andererseits bestand die Aufgabe der Teilnehmer mit erhöhter Expertise darin, das subjektive Urteil naiver Probanden zu schätzen, die selbst jedoch nur ihr eigenes Urteil fällen mussten. Dieser Unterschied in der Instruktion führte jedoch nicht zu einer signifikanten Veränderung der Streuung der Urteile. Eine mögliche Erklärung, die durch exemplarische Verbalisierungen von Experten gestützt wird, besteht darin, dass die Experten ihr eigenes subjektives Urteil gemäß ihrer Vorstellung der Fahrfähigkeit von naiven Probanden um einen weitestgehend konstanten Wert nach oben korrigieren.

Einen weiteren Hinweis für die Ähnlichkeit des Mechanismus der Störungsbewertung zwischen Teilnehmern mit hoher und geringer Expertise gibt der beobachtete Reihenfolgeeffekt. Bei der Auswertung des ersten Versuchs wurde gezeigt, dass die Wiederholung von Szenarien unabhängig vom Niveau der Expertise einen ähnlichen Effekt auf die Störungsbeurteilung hat. Dies stützt nicht nur die Annahme über die Gleichheit des Mechanismus der subjektiven Störungsbewertung, sondern gibt auch konkrete Hinweise für die Gestaltung von Expertenstudien für die Störungsbewertung kritischer Fahrsituationen. Da Experten in gleicher Form von Wiederholungseffekten betroffen sind, folgt daraus, dass within-subject Versuchspläne diesen Effekt berücksichtigen müssen.

Bei der Schätzung des Intervalls der Kritikalitätsbewertung durch Experten in Schriftform wurde eine Qualität der geschätzten Intervalle beobachtet, die mit den Werten aus der Literatur in anderen Domänen vergleichbar ist. Der begrenzte Umfang der Szenarien, die in dieser Form geprüft wurden, lässt jedoch keine detailliertere Analyse des Zusammenhangs zu.

Durch eine Auswertung der subjektiven Urteile der Experten im Vergleich zu naiven Probanden über 30 Szenarien konnte gezeigt werden, dass bei gegebenen Urteilen von Experten die mögliche Spannbreite der mittleren Urteile von Probanden vorhergesagt werden kann. Damit wurde ein wichtiges Ziel dieser Dissertation erfüllt und der Weg geebnet, Expertenbewertungen für die frühzeitige Eingrenzung zu betrachtender Systemvarianten in der Entwicklung von Fahrerassistenzsystemen zu verwenden.

Der gleiche Datensatz wurde anschließend verwendet, um individuelle Unterschiede im Urteilsverhalten der Experten zu untersuchen. Dabei wurde zunächst gefunden, dass die

Urteile der Experten nur in geringem Maße über verschiedene Szenarien miteinander korreliert sind. Dies kann so interpretiert werden, dass die Experten die Szenarien tatsächlich unabhängig voneinander betrachten und nicht etwa jeweils ein von ihnen bevorzugtes Kritikalitätsniveau unabhängig von der Situation zum Ausdruck bringen. Die untersuchte Stichprobe ist also nicht davon dominiert, dass manche Experten durchgängig überkritisch antworten oder andere die Kritikalität von Situationen konsequent zu niedrig einschätzen.

Dies heißt jedoch nicht, dass alle Experten sich in ihrem Urteilsverhalten ähneln. Bei der individuellen Auswertung der subjektiven Urteile wurden signifikante Unterschiede zwischen den Experten gefunden. Dieser Umstand wurde zum Anlass genommen, die Konstruktion neuer Expertengruppen zu untersuchen. Dabei wurden für die beispielhafte Stichprobengröße von fünf Experten alle möglichen Expertengruppen numerisch anhand einer Kreuzvalidierung miteinander verglichen. Es wurde festgestellt, dass 192 oder 2,62% der Fünfergruppen einen geringeren mittleren quadrierten Fehler aufweisen, als dies bei Verwendung aller 21 Experten der Fall ist. Die individuelle Betrachtung der Experten kann also genutzt werden, um die Präzision der Prädiktion der Kritikalitätsurteile naiver Probanden zu optimieren.

Schließlich wurde eine lineare Gewichtungsmethode auf die Urteile der Experten angewendet, um eine weitere Verringerung der Streuung zu erzielen. Der Effekt dieser Maßnahme war jedoch nur gering. Dies liegt daran, dass bei der Gewichtungsmethode zunächst die Tendenz eines Experten anhand eines Teils seiner Urteile ermittelt wurde und der Erfolg der Gewichtung an anderen Szenarien gemessen wurde. Zuvor wurde jedoch bereits festgestellt, dass sich die Urteile der Experten nur in geringem Maße von einem Szenario auf das andere übertragen lassen. Während diese Unabhängigkeit der Urteile in den einzelnen Szenarien zwar eine positive Eigenschaft der Experten ist, schränkt dies die Effektivität der Gewichtungsmethode der Experten ein.

6 Ausblick

In der vorliegenden Dissertation konnten mehrere Konzepte nicht abschließend bewertet werden. So wurde die Reliabilität von Experten zwar beispielhaft betrachtet und quantifiziert, jedoch konnte diese nicht über verschiedenartige Szenarien verglichen werden. Dadurch ist die externe Validität der quantifizierten Größe nur eingeschränkt.

Ähnlich wurde zwar im dritten Versuch gezeigt, dass es möglich ist, die Wahrnehmbarkeit der Kritikalität einer Fahrsituation und die Durchführbarkeit der Gegenmaßnahme als Unterkonstrukte der Beherrschbarkeit durch Likert-Skalen zu operationalisieren, jedoch reicht der Umfang der getesteten Szenarien nicht für eine Validierung der entworfenen Skalen aus.

Diese Kritik lässt sich auch auf die Aussagen zu Intervallschätzungen der subjektiv empfundenen Kritikalität anwenden. Während durch den vierten Versuch zwar Hinweise auf einen Zusammenhang gefunden worden sind, wäre ein größerer Umfang der zugrundliegenden Datenbasis wünschenswert.

Diese Dissertation hat sich darauf beschränkt, die Effekte einer Rekrutierungsstrategie auf die Wahrnehmung der Kritikalität von kritischen Fahrsituationen von Fahrerassistenzsystemen zu betrachten. Dabei wurde darauf verzichtet, eine konkrete Abgrenzung von Experten und Nichtexperten zu definieren oder theoretisch zu begründen. Während in dieser Dissertation zwar Unterschiede im Urteilsverhalten der betrachteten Gruppen quantitativ beschrieben worden sind, bleibt es für weitere Arbeiten in diesem Bereich übrig, dichotome, leistungsbasierte Kriterien für die Zugehörigkeit zur Expertengruppe zu definieren. So können etwa aus einem akzeptablen Restrisiko für Fehlentscheidungen bezüglich der Kritikalität von Fahrsituationen quantitative Kriterien für Experten hergeleitet werden, die sich experimentell überprüfen lassen.

Der Ansatz, sich nur auf die Effekte der Rekrutierungsstrategie zu beschränken, führt auch dazu, dass sich in dieser Dissertation nur Hinweise auf funktionale Unterschiede im Urteilsverhalten der betrachteten Gruppen finden. Mit dem Vorwissen des quantitativen Effekts der Expertise im Bereich Beherrschbarkeit können zukünftige Arbeiten in diesem Bereich nun die qualitativen Unterschiede der subjektiven Bewertung, die durch die Expertise hervorgerufen werden, untersuchen.

Anhand der zur Verfügung stehenden Stichprobe an Experten konnte nicht getrennt untersucht werden, ob die gemessenen Effekte durch die Fahrausbildung, das Alter oder Geschlecht beeinflusst wurden. Dies ist eine Konsequenz der Tatsache, dass Experten nur Rekrutierung, nicht durch Manipulation der Expertise zur Verfügung standen, womit alle hier durchgeführten Studien lediglich Quasi-Experimente darstellen. Um diesen methodischen Mangel zu beheben, muss in zukünftigen Arbeiten eine Methode zur Manipulation der Expertise entwickelt werden. Da nun gezeigt worden ist, dass die Effekte der Expertise einen nutzbaren Effekt haben, kann dies als Motivation dafür gelten, ein Training von Experten zu untersuchen. Dieses kann dazu verwendet werden, neue Experten auszubilden oder die Fähigkeiten bestehende Experten zu verbessern oder aufrechtzuerhalten.

Im empirischen Teil dieser Dissertation wurde darauf verzichtet, Persönlichkeitsmerkmale der untersuchten Experten aufzunehmen und in die Analyse des Urteilsverhaltens mit einzubeziehen, um ethische und rechtliche Bedenken zu vermeiden. Unter Zuhilfenahme einer Methode zur Manipulation der Beherrschbarkeitsexpertise könnten diese umgangen werden, da dann zufällige Stichproben aus der allgemeinen Bevölkerung ohne beruflichen Bezug der Probanden zur Beherrschbarkeitsbewertung möglich sind.

Zusammenfassend ergeben sich aus dieser Dissertation Fragestellungen, die die Ursachen bzw. Grundlagen der beobachteten Effekte betreffen sowie solche, die deren Anwendung betreffen. Untersuchungen zum Beitrag von Persönlichkeitsmerkmalen, demographischer Merkmale oder funktionaler Unterschiede zwischen Experten und Nichtexperten zum Urteilsverhalten stellen Forschungen zu den Grundlagen der hier beobachteten Effekte dar. Untersuchungen zur Manipulation von Expertise und Ausbildung von Experten hingegen würden zur Anwendbarkeit der Methode beitragen.

Mit dieser Dissertation wurde Expertenbewertung erstmalig im Bereich Beherrschbarkeit von Fahrerassistenzsystemen angewendet. Deswegen ergeben sich aus den Beobachtungen eine große Anzahl an Fragestellungen, deren Antworten einen erheblichen Beitrag zur Entwicklung von zukünftigen Fahrerassistenzsystemen haben werden.

7 Zusammenfassung

Fahrerassistenzsysteme in modernen PKW erfüllen eine Vielzahl an Funktionen, um den Komfort und die Sicherheit der Fahrzeuginsassen und anderer Verkehrsteilnehmer zu erhöhen. Die Anzahl der verfügbaren Systeme wie auch ihre Komplexität ist dabei seit der Einführung von Fahrerassistenzsystemen stetig gewachsen. Es ist absehbar, dass sich dieser Trend auch in Zukunft fortsetzen wird. Dieses Wachstum verursacht eine Erhöhung des Aufwands der Bestimmung der Beherrschbarkeit der Fahrerassistenzsysteme, da nicht nur mehr Systeme untersucht werden müssen, sondern auch die Anzahl möglicher kritischer Szenarien bei jedem System wächst. Expertenbewertungen sind eine der möglichen Methoden, um dieser Entwicklung entgegenzutreten. Sie können ressourceneffizient eingesetzt werden und sind nicht auf einzelne Systeme beschränkt. Andererseits wurde diese Methode noch nicht im Bereich Beherrschbarkeit von Fahrerassistenzsystemen validiert.

Expertenbewertungen wurden bereits in vielen anderen Gebieten eingesetzt und analysiert. Der Forschungsbereich Naturalistic Decision Making untersucht die Effektivität erfahrener Individuen und dokumentiert Randbedingungen, unter denen Personen herausragende Leistungen erbringen können. Forschungen im Kontext der kognitiven Verzerrungen und Heuristiken haben andererseits ergeben, dass subjektive Urteile von Menschen unter vielen Umständen systematischen Abweichungen von normativen Lösungen unterliegen. Diese Abweichungen wurden in vielen unterschiedlichen Versuchssettings reproduziert und sind robust gegen Manipulationen. Auch große Erfahrung mit dem Untersuchungsgegenstand oder Schulung bezüglich der Effekte von kognitiven Verzerrungen können Heuristiken nicht kompensieren. Diese widersprüchlichen Perspektiven auf die Vertrauenswürdigkeit subjektiver Urteile motivieren eine genauere Analyse der Expertenbewertung im Bereich Beherrschbarkeit von Fahrerassistenzsystemen.

Die Beherrschbarkeit einer kritischen Fahrsituation kann nicht direkt gemessen werden. Sie kann in Probandenversuchen im Realverkehr, auf der Teststrecke oder im Simulator experimentell geschätzt werden. Da realistische Beherrschbarkeitswahrscheinlichkeiten jedoch um oder über 90% liegen, ist eine experimentelle Schätzung für eine große Anzahl an Szenarien unter realistischen Umständen ausgeschlossen. Ohne eine ausreichend große Anzahl gesicherte Zielwerte lassen sich jedoch Expertenurteile nur sehr einge-

schränkt validieren. Deswegen wurde in dieser Dissertation die Schätzung der Kritikalität von Fahrsituationen in den Vordergrund gerückt. Diese ist dafür geeignet, eine Vorauswahl relevanter Szenarien zu treffen und so Aufwände für experimentelle Beherrschbarkeitsnachweise auf die kritischsten Szenarien zu konzentrieren.

Um diesen Zusammenhang zu untersuchen, wurden Minimalanforderungen an mögliche Experten definiert und aufgrund dieser 21 Personen rekrutiert, deren angenommene Expertise anschließend untersucht wurde. Das Urteilsverhalten dieser Experten bezüglich der subjektiven Kritikalität wurde über vier Simulatorstudien in 30 Szenarien mit den Urteilen naiver Probanden verglichen. Dieser Datensatz erlaubt es, statistische Zusammenhänge zwischen den Urteilen der beiden Gruppen herzustellen. So konnten konkrete Grenzwerte ermittelt werden, die eine Schätzung der Ober- und Untergrenzen der Urteile naiver Probanden alleine aufgrund der Expertenurteile erlauben. Diese Grenzwerte wurden auch für Expertengruppen zwischen 1 und 10 Mitgliedern bestimmt. Diese Daten können dafür eingesetzt werden, Expertenstudien durchzuführen, um die kritischsten Szenarien eines Fahrerassistenzsystems zu identifizieren und eine Abschätzung der Kritikalität der Szenarien zu erhalten.

Weiterhin konnten einige Eigenschaften der hier untersuchten Expertenstichprobe ermittelt werden. Die Retest-Reliabilität der Urteile der Experten wurde bestimmt und trotz Rekrutierung der Experten durch Minimalanforderungen bei einem mittelgroßen Wert gefunden. Die Fähigkeit der Experten zur Schätzung des Streuintervalls der Probandenurteile wurde untersucht und in einem Bereich identifiziert, der auch in anderen Expertenstudien aus der Literatur berichtet worden ist. Es konnte gezeigt werden, dass die Experten im gleichen Maße wie naive Probanden von Gewöhnungseffekten durch wiederholte Bewertung identischer Szenarien betroffen sind und dass die Streuung der Urteile der Experten nicht geringer als bei naiven Probanden ist. Diese Erkenntnisse können in das Design von Expertenstudien einfließen und so Fehlentscheidungen vermeiden.

Anhand des Datensatzes wurde gezeigt, dass die Urteile der Experten über verschiedene Szenarien nicht miteinander korreliert sind. Daraus folgt, dass die Experten die Situationen weitestgehend unabhängig voneinander betrachten. Das Urteilsverhalten der betrachteten Experten wurde anhand statistischer Kennwerte quantifiziert. Damit konnten die Experten anhand objektiver Daten verglichen werden und signifikante Unterschiede zwischen ihnen identifiziert werden. Diese Unterschiede wurden in einem Verfahren berücksichtigt, das die jeweilige Urteilscharakteristik der Experten verwendet, um eine

optimierte Zusammenstellung von Expertengruppen zu ermitteln. Mit diesem Verfahren konnte die Genauigkeit bei der Abschätzung der Urteile naiver Probanden weiter verbessert werden.

Diese Dissertation hat sich nicht mit der Entstehung der Expertise beschäftigt. Es wurden lediglich Minimalkriterien für Experten definiert und die resultierende Expertengruppe untersucht. Die Entstehung von Expertise setzt im Allgemeinen Rückmeldung über die Qualität der eigenen Leistung voraus. Da diese Rückmeldung im Bereich Beherrschbarkeit von Fahrerassistenzsystemen nur sehr eingeschränkt verfügbar ist, ist die erfolgreiche Entwicklung von individueller Expertise nicht selbstverständlich. Weiterhin hat sich diese Dissertation nicht damit beschäftigt, wie der Effekt der hier angenommenen Expertise wirkt und wodurch er beeinflusst wird. So lässt sich vermuten, dass Persönlichkeitsmerkmale einen Einfluss auf die Neigung zur Über- oder Unterschätzung der Kritikalität haben können oder dass der Kontext der Befragung einen Effekt haben kann.

Mit dieser Dissertation wurde erstmalig die Expertenbewertung im Bereich Beherrschbarkeit von Fahrerassistenzsystemen untersucht und Randbedingungen für die Anwendung identifiziert. Für die Bewertung der Kritikalität von Fahrsituationen wurden anhand einer Stichprobe konkrete Grenzwerte ermittelt. Weiterhin wurden verschiedene Eigenschaften der Expertenbewertung ermittelt, die das Design von Expertenstudien unterstützen können. Damit wurde ein Beitrag zur Entwicklung zukünftiger Fahrerassistenzsysteme geleistet.

Literaturverzeichnis

ACEA. (2009). *ACEA Endorses Response Code of Practicer for Advanced Driver Assistance Systems.* Zugriff am 23.11.2012. Verfügbar unter http://www.acea.be/news/news_detail/acea_endorses_response_code_of_practice_for_advanced_driver_assistance_syst

Ashton, R. H. (2012). Reliability and Consensus of Experienced Wine Judges: Expertise Within and Between? *Journal of Wine Economics, 7* (1), 70–87. Verfügbar unter http://journals.cambridge.org/action/displayAbstract?fromPage=online&aid=8651753

Australian Law Reform Commission. (2000). *Managing Justice: A Review of the Federal Justice System*.

Ayanian, J. Z., Landrum, M. B., Normand, S.-L. T., Guadagnoli, E. & McNeil, B. J. (1998). Rating the Appropriateness of Coronary Angiography - Do Practicing Physicians Agree with an Expert Panel and with Each Other? *The New England Journal of Medicine, 338* (26), 1896–1904. Zugriff am 11.03.2014. Verfügbar unter http://www.nejm.org/doi/pdf/10.1056/NEJM199806253382608

Ayyub, B. M. (2001). *Elicitation of Expert Opinion for Uncertainty and Risks:* CRC Press.

Bedford, T. & Cooke, R. (2001). *Probabilistic Risk Analysis: Foundations and Methods:* Cambridge University Press.

Bedinger, M., Walker, G., Piecyk, M., Greening, P. & Krupenia, S. (2015). A hierarchical task analysis of commercial distribution driving in the UK. In *Procedia Manufacturing: 6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015) and the Affiliated Conferences* .

Bengler, K., Dietmayer, K., Färber, B., Maurer, M., Stiller, C. & Winner, H. (2012). *Die Zukunft der Fahrerassistenz. Ein Strategiepapier der Uni-DAS:* Uni-DAS e.V.

Bortz, J. & Döring, N. (2006). *Forschungsmethoden und Evaluation. Für Human- und Sozialwissenschaftler* (Springer-Lehrbuch, 4. Aufl.). Heidelberg: Springer.

Braun, P. A. & Yaniv, I. (1992). A Case Study of Expert Judgment: Economists' Probabilities versus Base-rate Model Forecasts. *Journal of Behavioral Decision Making* (5), 217–231.

Burgman, M., Fidler, F., McBride, M., Walshe, T. & Wintle, B. (2006). Eliciting Expert Judgments: Literature Review.

Buttersworth, B. (2006). Mathematical Expertise. In K. A. Ericsson, N. Charness, P. J. Feltovich & R. R. Hoffman (Hrsg.), *The Cambridge Handbook of Expertise and Expert Performance.* Cambridge, New York: Cambridge University Press. Zugriff am 21.10.2014. Verfügbar unter http://www.mathematicalbrain.com/pdf/2006BBCHEEP.PDF

Cabantous, L., Hilton, D., Kunreuther, H. & Michel-Kerjan, E. (2011). Is imprecise knowledge better than conflicting expertise? Evidence from insurers' decisions in the United States. *Journal of Risk and Uncertainty, 42* (3), 211–232.

Camerer, C. F. & Johnson, E. J. (1991). The process-performance paradox in expert judgment. How can experts know so much and predict so badly? In A. Ericsson & J. Smith (Hrsg.), *Toward a General Theory of Expertise: Prospects and Limits.* New York: Cambridge University Press. Zugriff am 30.06.2014. Verfügbar unter http://www0.gsb.columbia.edu/mygsb/faculty/research/pubfiles/1152/Process_and_performance_of_experts.pdf

Caputo, A. (2013). A literature review of cognitive biases in negotiation processes. *International Journal of Conflict Management, 24 (4),* 374–398.

Carbonell, K. B., Stalmeijer, R. E., Könings, K. D., Segers Mien & van Marrienboer, J. J. (2014). How experts deal with novel situations: A review of adaptive expertise. *Educational Research Review, 12,* 14–29.

Chapman, G. B. (2004). The Psychology of Medical Decision Making. In D. J. Koehler & N. Harvey (Hrsg.), *Blackwell handbook of judgment and decision making* (1. Aufl., S. 585–603). Oxford: Blackwell Pub.

Chapman, G. B. & Johnson, E. J. (2002). Incorporating the irrelevant: Anchors in judgments of belief and value. In T. Gilovich, D. Griffin & D. Kahneman (Hrsg.), *Heuristics and Biases: The Psychology of Intuitive Judgment* (S. 120–138). Cambridge University Press.

Chi, M. T. (2006). Two Approaches to the Study of Experts' Characteristics. In K. A. Ericsson, N. Charness, P. J. Feltovich & R. R. Hoffman (Hrsg.), *The Cambridge Handbook of Expertise and Expert Performance.* Cambridge, New York: Cambridge University Press. Zugriff am 21.10.2014. Verfügbar unter

http://learnlab.org/uploads/mypslc/publications/chi%20two%20approaches%20chapter%202006.pdf

Chi, M. T., Glaser, R. & Rees, E. (1982). Expertise in problem solving. In R. J. Sternberg (Hrsg.), *Advances in the psychology of human intelligence* (Bd. 1). Hillsdale, NJ: Erlbaum.

Clemen, R. T. & Winkler, R. L. (1999). Combining Probability Distributions From Experts in Risk Analysis. *Risk Analysis, 19* (2), 187–203. Verfügbar unter https://faculty.fuqua.duke.edu/~clemen/bio/Published%20Papers/28.CombiningDistributions-Clemen&Winkler-RA-99.pdf

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Continental AG. (2015). *Continental-Mobilitätsstudie 2015*.

Cooke, R. & Goossens, L. H. J. (2000). Procedures guide for structured expert judgement in accident consequence modelling. *Radiation Protection Dosimetry, 90* (3), 303–309.

Cooke, R. M. (1991). *Experts in uncertainty: opinion and subjective probability in science.* New York: Oxford University Press.

De Bondt, W. F. M. (1991). What do economists know about the stock market? *Journal of Portfolio Management, 17* (2), 84–91.

Donges, E. (1982). Aspekte der aktiven Sicherheit bei der Führung von Personenkraftwagen. *Automobil-Industrie,* 183–190.

Einicke, G. A. (2012). *Smoothing, Filtering and Prediction: Estimating the Past, Present and Future:* Intech.

Englich, B. & Mussweiler, T. (2001). Sentencing Under Uncertainty: Anchoring Effects in the Courtroom. *Journal of applied Social Psychology, 31* (7), 1535–1551.

Englich, B., Mussweiler, T. & Strack, F. (2006). Playing Dice With Criminal Sentences: The Influence of Irrelevant Anchors on Experts' Judicial Decision Making. *Personality and Social Psychology Bulletin, 32* (2), 188–200.

Epley, N. & Gilovich, T. (2004). Are Adjustments Insuficient. *Personality and Social Psychology Bulletin, 30* (4), 447–460.

Ericsson, K. A. (2008). Deliberate practice and acquisition of expert performance. *Academic Emergency Medicine, 15* (11), 988–994.

Ericsson, K. A., Krampe, R. T. & Tesch-Romer, C. (1993). The Role of Deliberate Practice in the Acquisition of Expert Performance. *Psychological Review, 100* (3), 363–406.

Europäisches Parlament und Rat. (2009). VERORDNUNG (EG) Nr. 661/2009 DES EUROPÄISCHEN PARLAMENTS UND DES RATES vom 13. Juli 2009 über die Typgenehmigung von Kraftfahrzeugen, Kraftfahrzeuganhängern und von Systemen, Bauteilen und selbstständigen technischen Einheiten für diese Fahrzeuge hinsichtlich ihrer allgemeinen Sicherheit (661).

Fastenmeier, W. (1995). Autofahrer und Verkehrssituation: Neue Wege zur Bewertung von Sicherheit und Zuverlässigkeit moderner Straßenverkehrssysteme. Bonn: Deutscher Psychologen Verlag

Farrington-Darby, T. & Wilson, J. R. (2006). The nature of expertise: a review. *Applied ergonomics, 37* (1), 17–32.

Ferrell, W. R. (1994). Discrete subjective probabilities and decision analysis: elicitation, calibration and combination. In G. Wright & P. Ayton (Hrsg.), *Subjective Probability*. New York: John Wiley and Sons.

Fischoff, B. (1981). *Debiasing* (PTR-1092-81-3). : Perceptronics.

Floater, M. S. (1991). *Derivatives of rational Bezier curves.* Zugriff am 09.01.2014. Verfügbar unter http://www.mn.uio.no/math/english/people/aca/michaelf/papers/bez.pdf

Galaske, P., Farid, M., & Bengler, K. (2014). Influence of Expertise on the Judgment of Controllability of Advanced Driver Assistance Systems. In Proceedings of the 5th International Conference on Applied Human Factors and Ergonomics AHFE 2014, 2128–2135.

Galaske, P., Reisenauer, R., Farid, M., & Bengler, K. (2015). Diagnostische Bewertung der Beherrschbarkeit von Fahrerassistenzsystemen. In VerANTWORTung für die Arbeit der Zukunft, GfA-Press, Dortmund.

Galaske, P., Weinbeer, V., Farid, M., & Bengler, K. (2015). Confidence and Reliability in Driving Simulator Based Expert Reviews of ADAS Controllability. In Procedia Manufacturing, 3, 2974–2981.

Gasser, T. M., Arzt, C., Ayoubi, M. & Bartels, A. (2012). *Rechtsfolgen zunehmender Fahrzeugautomatisierung*. : Bundesanstalt für Straßenwesen.

Gasser, T. M., Seeck, A. & Smith, B. W. (2015). Rahmenbedingungen für die Fahrerassistenzentwicklung. In H. Winner, S. Hakuli, F. Lotz & C. Singer (Hrsg.), *Handbuch Fahrerassistenzsysteme* (3. Aufl.). Springer Fachmedien Wiesbaden.

Goldberg, L. R. (1970). Man Versus Model of Man. A Rationale, Plus some Evidence, for a Method of Improving on Clinical Inferences. *Psychological Bulletin, 73* (6), 422–432.

Green, D. M. & Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. New York: Wiley.

Green, K. C., Armstrong, J. S. & Graefe, A. (2007). Methods to elicit forecasts from groups: Delphi and prediction markets compared. *Foresight: The International Journal of Applied Forecasting, 8*, 17–20.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E. & Nelson, C. (2000). Clinical vs. mechanical prediction: A meta-analysis. *Psychological Assessment, 12*, 19–30.

Harper, R. P. & Cooper, G. E. (1986). Handling Qualities and Pilot Evaluation. *Journal of Guidance, Control and Dynamics, 9* (5), 515–529.

Hart, A. (1986). *Knowledge acquisition for expert systems*. New York: Mc Graw-Hill Book Co.

Hassard, S. T. (2009). The Variations of Recognition Primed Decision-Making and How it Informs Design Decision-Making. *Proceedings of NDM9, the 9th International Conference on Naturalistic Decision Making*, 339.

Helmer, T. (2015). *Development of a methodology for the evaluation of active safety using the example of preventive pedestrian protection. Doctoral thesis accepted by Technische Universität Berlin, Germany* (Springer theses): Springer Theses.

Helmreich, R. L., Merrit, A. C. & Sherman, P. J. (1996). Human Factors and National Culture in Aircraft. *Human Factors and National Culture. ICAO Journal, 51* (8), 14–16.

Hodge, T. & Deakin, J. M. (1998). Deliberate practice and expertise in the martial arts: The role of context in motor recall. *Journal of Sport & Exercise Psychology, 20*, 260–279.

Hoffman, R. R. (1996). How Can Expertise be Defined? Implications of research From Cognitive Psychology. In R. Williams, W. Falkner & J. Fleck (Hrsg.), *Exploring Expertise* (S. 81–100). Edinburgh: University of Edinburgh Press.

Kahneman, D. (2012). *Thinking, fast and slow*. London: Penguin Books.

Kahneman, D. & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist, 64* (6), 515–526.

Kirwan, B. (1994). *A Guide to Practical Human Reliability Assessment:* CRC Press.

Klein, G. (1998). *Sources of Power: How People Make Decisions*. Cambridge, MA: MIT Press.

Klein, G. (2008). Naturalistic Decision Making. *Human Factors, 50* (3), 456–460.

Kompaß, K., Helmer, T., Wang, L. & Kates, R. (2015). Gesamthafte Bewertung der Sicherheitsveränderung durch FAS / HAF im Verkehrssystem: Der Beitrag von Simulation. *Fahrerassistenz und Aktive Sicherheit: Wirksamkeit - Beherrschbarkeit - Absicherung*.

Krengel, U. (2005). *Einführung in die Wahrscheinlichkeitstheorie und Statistik* (8. Aufl.): Vieweg.

Kunreuther, H., Pauly, M. V. & McMorrow, S. (2013). *Insurance and Behavioral Economics. Improving Decisions in the Most Misunderstood Industry*. Cambridge: Cambridge University Press.

Larrick, R. P. (2004). Debiasing. In D. J. Koehler & N. Harvey (Hrsg.), *Blackwell handbook of judgment and decision making* (1. Aufl., S. 316–337). Oxford: Blackwell Pub.

Lewis, H. W., Budnitz, R. J., Rowe, W. D., Kouts, H. J. C., Hippel, F. von, Loewenstein, W. B. et al. (1979). Risk assessment review group report to the U.S. Nuclear Regulatory Commission. *IEEE Transactions on Nuclear Science, 26* (5), 4686–4690.

Lin, S.-w. & Bier, V. M. (2008). A study of expert confidence, *93* (5), 711–721.

Lipshitz, R., Klein, G., Orasanu, J. & Salas, E. (2001). Taking stock of naturalistic decision making. *Journal of Behavioral Decision Making, 14* (5), 331–352.

Malone, S. & Brünken, R. (2015). Hazard perception assessment - How much ecological validity is necessary? *6th International Conference on Applied Human Factors and Ergonomics*.

Maybeck, P. S. (1979). *Stochastic models, estimation and control* (Mathematics in science and engineering, Bd. 141, 3 Bände). New York: Academic Press.

Meehl, P. E. (1954). *Clinical versus Statistical Prediction. A Theoretical Analysis and a Review of the Evidence*.

Meyer, M. A. & Booker, J. M. (1990). *Eliciting and Analyzing Expert Judgment*. Washington, DC: US Nuclear Regulatory Commission, Office of Nuclear Regulatory Research, Division of Systems Research.

Moore, D. A., Tenney, E. R. & Haran, U. (2016). Overprecision in Judgment. In G. Keren & G. Wu (Hrsg.), *Blackwell Handbook of Judgment and Decision Making* (2. Aufl.). New York: Wiley; Wiley-Blackwell. Zugriff am 12.08.2014. Verfügbar unter http://learnmoore.org/mooredata/140404HOC.pdf

Morgan, M. G. (2014). Use (and abuse) of expert elicitation in support of decision making for public policy. *Proceedings of the National Academy of Sciences, 111* (20), 7176–7184.

Mussweiler, T., Englich, B. & Strack, F. (2004). Anchoring Effect. In R. Pohl (Hrsg.), *Cognitive Illusions: A Handbook of Fallacies and Biases in Thinking, Judgment and Memory* (S. 183–200). Psychology Press.

Neukum, A. (2010). Controllability of erroneous steering torque interventions: Driver reactions and influencing factors.

Neukum, A. & Krüger, H.-P. (2003). Fahrerreaktionen bei Lenksystemstörungen - Untersuchungsmethodik und Bewertungskriterien. *VDI-Berichte* (1791), 297–318.

Neukum, A., Lübbeke, T., Krüger, H.-P., Mayser, C. & Steinle, J. (2008) ACC-Stop&Go: Fahrerverhalten an funktionalen Systemgrenzen. In M. Maurer & C. Stiller (Hrsg.), *5. Workshop Fahrerassistenzsysteme* (S. 141–150). Karlsruhe.

Neukum, A. & Reinelt, W. (2005). Bewertung der Funktionssicherheit aktiver Lenksysteme: ein Human Factors Ansatz. *VDI-Berichte Nr 1919*, 161–176.

Neukum, A., Paulig, J., Frömmig, L., Henze, R. (2009). Untersuchung zur Wahrnehmung von Lenkmomenten bei PKW. *FAT-Schriftenreihe 222.*

NHTSA. (2013). *U.S. Department of Transportation Releases Policy on Automated Vehicle Development*. Verfügbar unter http://www.nhtsa.gov/About+NHTSA/Press+Releases/U.S.+Department+of+Transportation+Releases+Policy+on+Automated+Vehicle+Development

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J. et al. (2006). *Uncertainty Judgments: Eliciting Expert Probabilities*. New York: Wiley.

Ouchi, F. (02.2014). *A Literature Review on the Use of Expert Opinion in Probabilistic Risk Analysis.* : World Bank.

Phillips, J. K., Klein, G. & Sieck, W. R. (2004). Expertise in Judgment and Decision Making: A Case for Training Intuitive decision Skills. In D. J. Koehler & N. Harvey (Hrsg.), *Blackwell handbook of judgment and decision making* (1. Aufl., S. 297–315). Oxford: Blackwell Pub.

PReVENT. (2009). Code of Practice for the Design and Evaluation of ADAS. V5.0.

Rangra, S., Sallak, M., Schön, W. & Vanderhaegen, F. (2015). Human reliability assessment under uncertainty. *6th International Conference on Applied Human Factors and Ergonomics*, 2633–2640.

Rasmussen, J. (1983). Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. *IEEE Transactions on Systems, Man, Cybernetics, SMC-13* (3), 257–266.

Rasmussen, N. C. & et al. (1975). *Reactor safety study. An assessment of accident risk in U.S. commersial nuclear power plants.* WASH-1400 (NUREG-75/014). Rockville, MS: U.S. Nuclear Regulatory Commission.

RESPONSE 3. (2009). *Code of Practice for the Design and Evaluation of ADAS* (5.0. Aufl.).

ISO, 26262-3 (2011). *Road vehicles – Functional safety - Part 3: Concept Phase.*

Rohrmann, B. (1978). Empirische Studien zur Entwicklung von Antwortskalen fuer die sozial-wissenschaftliche Forschung. *Zeitschrift fuer Sozialpsychologie* (9), 222–245.

Sauvageaot, F., Urdapilleta, I. & Peyron, D. (2006). Within and between variations of texts elicited from nine wine experts. *Food and Quality Preference, 17,* 429–444.

ISO, 5492 (2008). *Sensory analysis—Vocabulary.*

Shanteau, J. (1992). Competence in experts: The role of task characteristics. *Organisational Behavior and Human Decision Processes, 53,* 252–262.

Simmermacher, D. & Winner, H. (2011). Beherrschbarkeit von Gierstörungen durch ein Fahrerkollektiv. *ATZ - Automobiltechnische Zeitschrift, 113* (9), 696–701.

Simon, H. A. (1992). What is an “explanation” of behavior? *Psychological Science, 3* (3), 150–161.

Soll, J. B. & Klayman, J. (2004). Overconfidence in Interval Estimates. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30* (2), 299–314. Verfügbar unter http://www.chicagocdr.org/cdrpubs/pdf_index/cdr_559.pdf

Stewart, T. R., Moninger, W. R., Grassia, J., Brady, R. H. & Merrem, F. H. (1989). Analysis of Expert Judgment in a Hail Forecasting Experiment. *Weather and Forecasting* (4), 24–32.

Swain, A. & Guttmann, H. (1983). *Handbook of Human Reliability Analysis with Emphasis on Nuclear Power Plant Applications:* USNRC.

SAE, J3016 (2014). *Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems.*

Tetlock, P. E. (2005). *Expert political judgment. How good is it? How can we know?* Princeton: Princeton University Press.

Timmerbeil, S. (2003). The Role of Expert Witnesses in German and U.S. Civil Litigation. *Annual Survey of International & Comparative Law, 9* (1), 164–187.

Tversky, A. & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science, 185* (4157), 1124–1131.

Ullmann, S. (2006). *Der Normalfahrer als Messgröße für die Optimierung und Absicherung aktiver fahrdynamischer Regelsysteme*, Tagung Aktive Sicherheit durch Fahrerassistenzsysteme, München.

Vasyukova, E. E. (2012). The Nature of Chess Expertise: Knowledge of Search? *Psychology in Russia: State of the Art,* 511–528.

Ward, P., Hodges, N. J., Starkes, J. L. & Williams, M. A. (2007). The road to excellence: Deliberate practice and the development of expertise. *High Ability Studies, 18* (2), 119–153. s.

Weitzel, A. & Winner, H. (2012). Ansatz zur Kontrollierbarkeitsbewertung von Fahrerassistenzsystemen vor dem Hintergrund der ISO 26262. In K. Dietmayer (Hrsg.), *8. Workshop Fahrerassistenzsysteme.* Darmstadt: Uni-Das.

Wesp, A. (2011). *Analyse fahrerwirksamer Systemauslegungen und -störungen eines Fahrzeugs mit Hinterradlenkung bei gleichzeitiger Fahrerbeanspruchung durch eine Fahraufgabe*. Dissertation, TU Darmstadt. Darmstadt.

Wilhelm, U., Ebel, S. & Weitzel, A. (2015). Funktionale Sicherheit und ISO 26262. In H. Winner, S. Hakuli, F. Lotz & C. Singer (Hrsg.), *Handbuch Fahrerassistenzsysteme* (3. Aufl., S. 85–104). Springer Fachmedien Wiesbaden.

Wilson, T. D., Houston, C., Etling, K. M. & Brekke, N. (1996). A new look at anchoring effects: Basic anchoring and its antecedents. *Journal of Experimental Psychology, 4*, 387–402.

World Health Organization. (2013). *Health Risk Assessment from the nuclear accident after the 2011 great east japan earthquake and tsunami*.

Anhang

Übersicht der verwendeten Basisszenarien

In diesem Unterkapitel werden die Fahrszenarien der vier durchgeführten Fahrsimulatorstudien aufgeführt. Jedes Szenario wird zunächst in Anlehnung an die Systematik von Fastenmeier (1995) klassifiziert. Anschließend stellt ein Schaubild eine visuelle Übersicht über das Szenario und die beteiligten Aktoren dar. Danach wird der typische Ablauf des Szenarios in Textform dargestellt. Die konkreten Fahrsituationen der vier durchgeführten Versuche basieren jeweils auf einem der hier beschriebenen Szenarien. Gegebenenfalls vorgenommen Anpassungen oder Parametrisierungen können der jeweiligen Versuchsbeschreibung entnommen werden.

Tabelle 40 Klassifikation von Szenario 1

Merkmal	Ausprägung
Straßentyp	Landstraße einspurig
Horizontalverlauf	Kurve
Vertikalverlauf	Kuppe
Knotenpunkte	Keine
Engstellen	Stehendes Hindernis
Assistenzsystem	Längs- und Querführung des Fahrzeugs in der Spur
Automatisierungslevel (BaST)	3
Kritische Systemgrenze	Begrenzte automatische Verzögerung

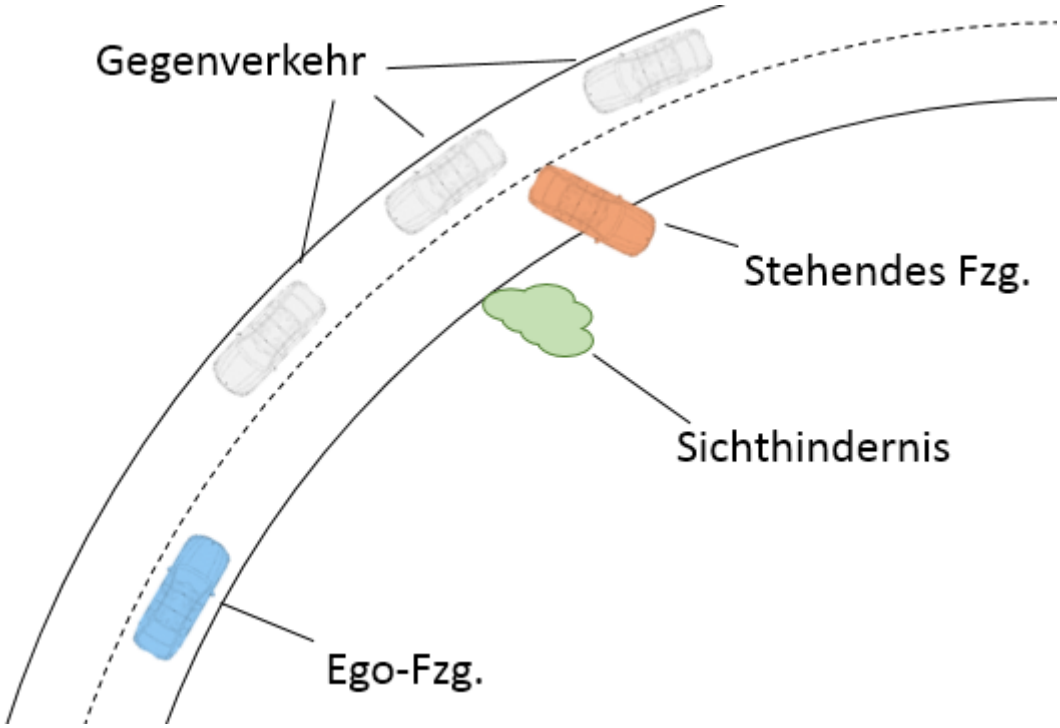

Abbildung 28 Übersichtsgrafik für Szenario 1

Das Egofahrzeug bewegt sich mit auf einer einspurigen Landstraße mit Gegenverkehr. Das Fahrzeug übernimmt die Längs- und Querführung des Fahrzeugs während der Fahrer durch eine Nebentätigkeit visuell sowie kognitiv abgelenkt ist. In einer Rechtskurve mit einer Kuppe erfolgt ein Warnton durch das Assistenzsystem. Kurz darauf wird hinter einem Sichthindernis ein stehendes Fahrzeug in der Spur des Egofahrzeugs sichtbar. Gegenverkehr versperrt die gegenüberliegende Spur und zwischen Gegenverkehr und dem stehenden Hindernis ist nicht genügend Platz, um dazwischen durchzufahren. Die automatische Verzögerung des Assistenzsystems ist nicht ausreichend, um eine Kollision mit dem stehenden Hindernis zu verhindern. Der Fahrer muss selbst die Bremse betätigen, um das Egofahrzeug vor dem Hindernis zum Stehen zu bringen.

Tabelle 41 Klassifikation von Szenario 2

Merkmal	Ausprägung
Straßentyp	Landstraße einspurig
Horizontalverlauf	Kurve
Vertikalverlauf	Kuppe
Knotenpunkte	Keine
Engstellen	Stehendes Hindernis
Assistenz	Längs- und Querführung des Fahrzeugs in der Spur
Automatisierungslevel (BaST)	2
Kritische Systemgrenze	Begrenzte automatische Verzögerung

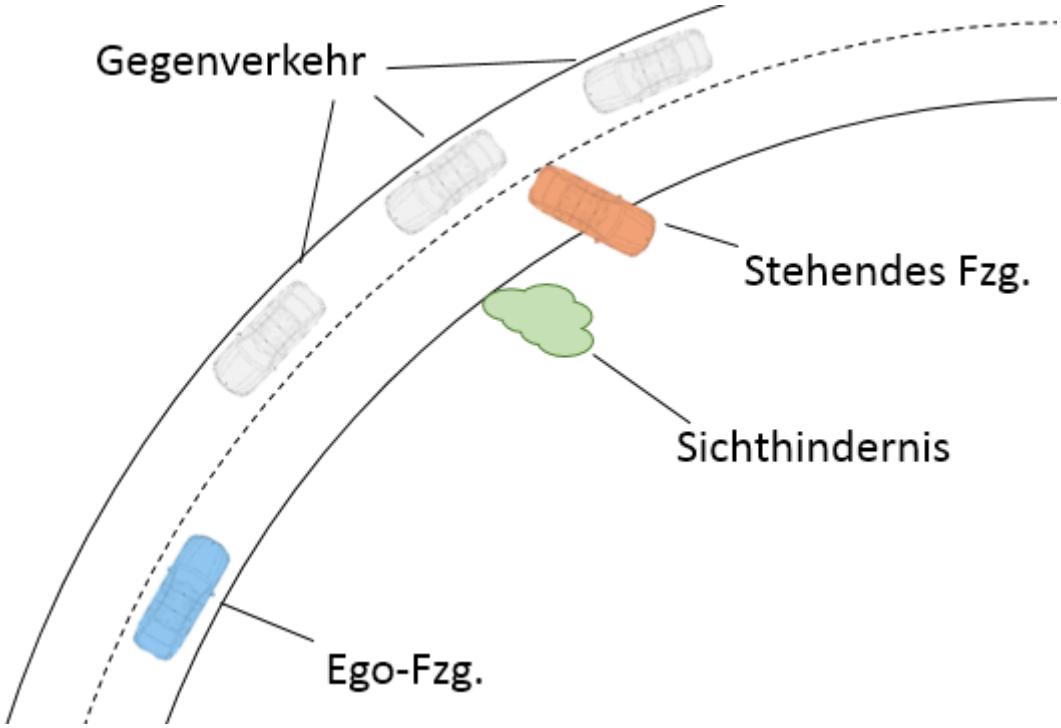

Abbildung 29 Übersichtsgrafik für Szenario 2

Das Egofahrzeug bewegt sich mit auf einer einspurigen Landstraße mit Gegenverkehr. Das Fahrzeug übernimmt die Längs- und Querführung des Fahrzeugs während der Fahrer das System überwacht. In einer Rechtskurve mit einer Kuppe erfolgt ein Warnton durch das Assistenzsystem. Kurz darauf wird hinter einem Sichthindernis ein stehendes Fahrzeug in der Spur des Egofahrzeugs sichtbar. Gegenverkehr versperrt die gegenüberliegende Spur und zwischen Gegenverkehr und dem stehenden Hindernis ist nicht genügend Platz, um dazwischen durchzufahren. Die automatische Verzögerung des Assistenzsystems ist nicht ausreichend, um eine Kollision mit dem stehenden Hindernis zu

verhindern. Der Fahrer muss selbst die Bremse betätigen, um das Egofahrzeug vor dem Hindernis zum Stehen zu bringen.

Tabelle 42 Klassifikation von Szenario 3

Merkmal	Ausprägung
Straßentyp	Landstraße einspurig
Horizontalverlauf	Gerade
Vertikalverlauf	Eben
Knotenpunkte	Keine
Engstellen	Keine
Assistenz	Längs- und Querführung des Fahrzeugs in der Spur
Automatisierungslevel (BaST)	2
Kritische Systemgrenze	Unangebrachtes Lenkmoment zum Fahrbahnrand

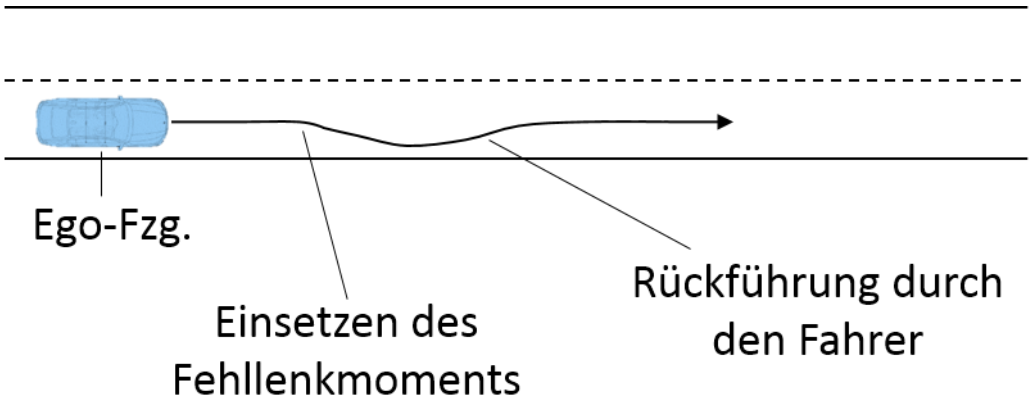

Abbildung 30 Übersichtsgrafik für Szenario 3

Das Egofahrzeug bewegt sich auf einer einspurigen Landstraße ohne Gegenverkehr. Das Fahrzeug übernimmt die Längs- und Querführung des Fahrzeugs während der Fahrer das System überwacht. Auf einem ebenen sowie geraden Streckenabschnitt erfolgt unvermittelt ein Lenkmoment zum rechten Fahrbahnrand für zwei Sekunden. Der Fahrer muss das Fehllenkmoment abstützen, um zu verhindern, dass das Fahrzeug die Fahrspur verlässt. Nach dem Ende des Fehllenkmoments muss der Fahrer das Fahrzeug weiterhin in der Spur stabilisieren.

Tabelle 43 Klassifikation von Szenario 4

Merkmal	Ausprägung
Straßentyp	Autobahn zweispurig
Horizontalverlauf	Gerade
Vertikalverlauf	Eben
Knotenpunkte	Keine
Engstellen	Keine
Assistenz	Längs- und Querführung des Fahrzeugs in der Spur
Automatisierungslevel (BaST)	2
Kritische Systemgrenze	Verspätete Erkennung von einscherenden Fahrzeugen

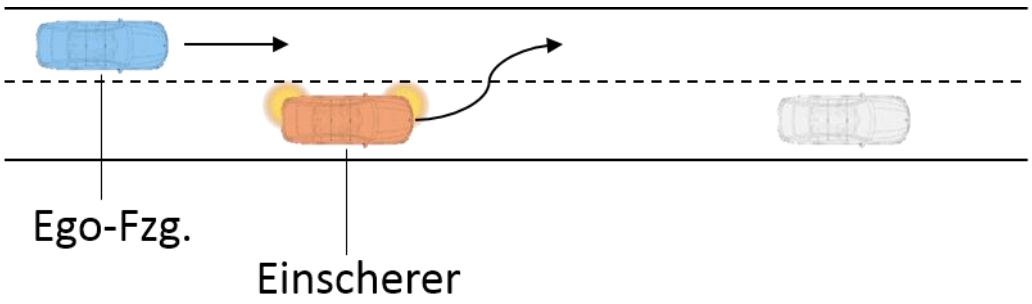

Abbildung 31 Übersichtsgrafik für Szenario 4

Das Egofahrzeug bewegt sich auf der linken von zwei Spuren auf der Autobahn. Das Fahrzeug übernimmt die Längs- und Querführung des Fahrzeugs während der Fahrer das System überwacht. Auf einem ebenen sowie geraden Streckenabschnitt schert ein langsameres Fremdfahrzeug von der rechten Spur in die Spur des Egofahrzeugs ein. Der Fahrer muss die Bremse betätigen, um eine Kollision mit dem einscherenden Fahrzeug zu verhindern.

Tabelle 44 Klassifikation von Szenario 5

Merkmal	Ausprägung
Straßentyp	Landstraße einspurig
Horizontalverlauf	Gerade
Vertikalverlauf	Eben
Knotenpunkte	Keine
Engstellen	Keine
Assistenz	Längs- und Querführung des Fahrzeugs in der Spur
Automatisierungslevel (BaST)	2
Kritische Systemgrenze	Unangebrachter Lenkruck nach rechts

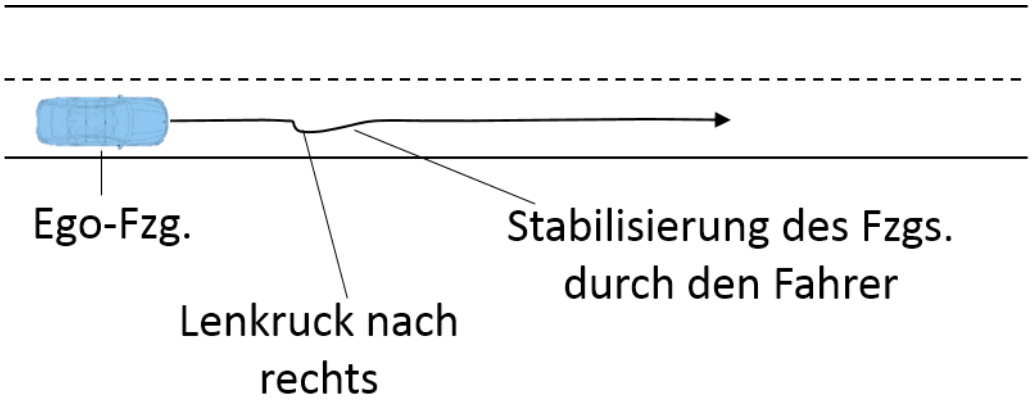

Abbildung 32 Übersichtsgrafik für Szenario 5

Das Egofahrzeug bewegt sich auf einer einspurigen Landstraße ohne Gegenverkehr. Das Fahrzeug übernimmt die Längs- und Querführung des Fahrzeugs während der Fahrer das System überwacht. Auf einem ebenen sowie geraden Streckenabschnitt erfolgt unvermittelt ein Lenkruck zum rechten Fahrbahnrand für 300 ms. Der Fahrer muss das stabilisieren, um zu verhindern, dass das Fahrzeug die Fahrspur verlässt.

Tabelle 45 Klassifikation von Szenario 6

Merkmal	Ausprägung
Straßentyp	Landstraße einspurig
Horizontalverlauf	Gerade
Vertikalverlauf	Eben
Knotenpunkte	Keine
Engstellen	Keine
Assistenz	Längs- und Querführung des Fahrzeugs in der Spur
Automatisierungslevel (BaST)	2
Kritische Systemgrenze	Begrenzte automatische Verzögerung

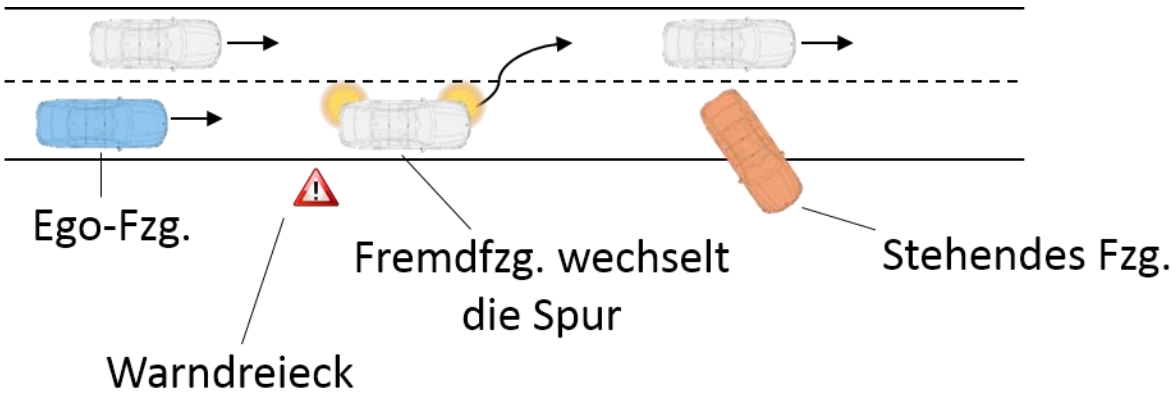

Abbildung 33 Übersichtsgrafik für Szenario 6

Das Egofahrzeug bewegt sich auf einer zweispurigen Autobahn hinter einem Vorderfahrzeug. Das Ego-Fahrzeug übernimmt die Längs- und Querführung des Fahrzeugs während der Fahrer das System überwacht. Auf einem ebenen sowie geraden Streckenabschnitt wechselt das Vorderfahrzeug die Spur und gibt die Sicht auf ein stehendes Hindernis frei. Der Abstand zum Hindernis, in dem das Vorderfahrzeug die Spur wechselt, ist in zwei Stufen konfigurierbar. Das Vorhandensein eines Warndreiecks auf dem Standstreifen ist ebenfalls konfigurierbar. Fremdfahrzeuge auf der Nebenspur verhindern einen Fahrspurwechsel des Egofahrzeugs nach links. Das Hindernis kann nicht nach rechts umfahren werden. Der Fahrer muss eine Kollision durch einen Bremseingriff verhindern.

Tabelle 46 Klassifikation von Szenario 7

Merkmal	Ausprägung
Straßentyp	Innerorts einspurig
Horizontalverlauf	Gerade
Vertikalverlauf	Eben
Knotenpunkte	Keine
Engstellen	Keine
Assistenz	Abstandstempomat
Automatisierungslevel (BaST)	1
Kritische Systemgrenze	Begrenzte automatisierte Verzögerung

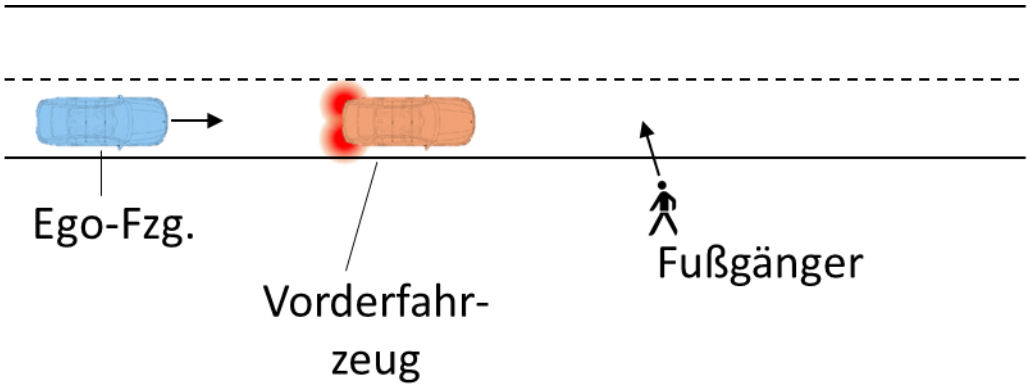

Abbildung 34 Übersichtsgrafik für Szenario 7

Das Egofahrzeug bewegt sich auf einer einspurigen Straße innerorts ohne Gegenverkehr. Das Fahrzeug übernimmt die Längsführung des Fahrzeugs während der Fahrer die Querführung übernimmt. Das Egofahrzeug folgt einem Vorderfahrzeug. Auf einem ebenen sowie geraden Streckenabschnitt überquert ein Fußgänger vor dem Vorderfahrzeug die Fahrbahn und das Vorderfahrzeug leitet eine Vollbremsung ein. Der Fahrer muss durch einen Bremseingriff eine Kollision mit dem Vorderfahrzeug verhindern.

Tabelle 47 Klassifikation von Szenario 8

Merkmal	Ausprägung
Straßentyp	Innerorts einspurig
Horizontalverlauf	Gerade
Vertikalverlauf	Eben
Knotenpunkte	Keine
Engstellen	Fahrradfahrer
Assistenz	Längs- und Querführung des Fahrzeugs in der Spur
Automatisierungslevel (BaST)	2
Kritische Systemgrenze	Unangebrachtes Lenkmoment nach rechts

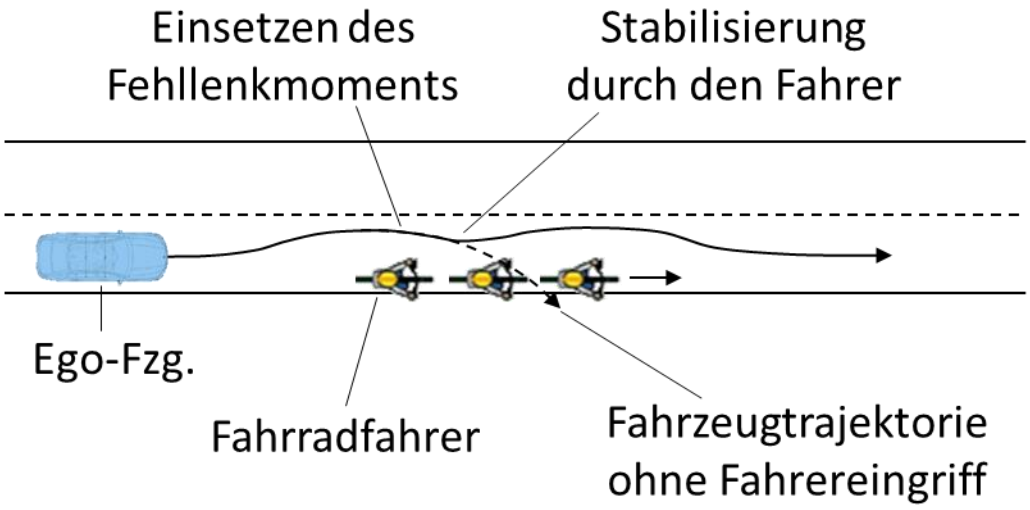

Abbildung 35 Übersichtsgrafik für Szenario 8

Das Egofahrzeug bewegt sich auf einer einspurigen Straße innerorts. Das Fahrzeug übernimmt die Längs- und Querführung des Fahrzeugs in der Engstelle während der Fahrer das System überwacht. Auf einem ebenen sowie geraden Streckenabschnitt passiert der Fahrer eine Gruppe Fahrradfahrer, die am rechten Rand der Spur fährt. Während dem Überholvorgang führt das Assistenzsystem ein unangebrachtes Lenkmoment nach rechts aus. Der Fahrer muss das Fehllenkmoment abstützen, um zu verhindern, dass das Ego-Fahrzeug mit den Fahrradfahrern kollidiert. Nach dem Ende des Fehllenkmoments muss der Fahrer das Fahrzeug weiterhin in der Spur stabilisieren.

Tabelle 48 Klassifikation von Szenario 9

Merkmal	Ausprägung
Straßentyp	Landstraße einspurig
Horizontalverlauf	Gerade
Vertikalverlauf	Kurve
Knotenpunkte	Keine
Engstellen	Keine
Assistenz	Längs- und Querführung des Fahrzeugs in der Spur
Automatisierungslevel (BaST)	2
Kritische Systemgrenze	Fußgänger wird nicht erkannt

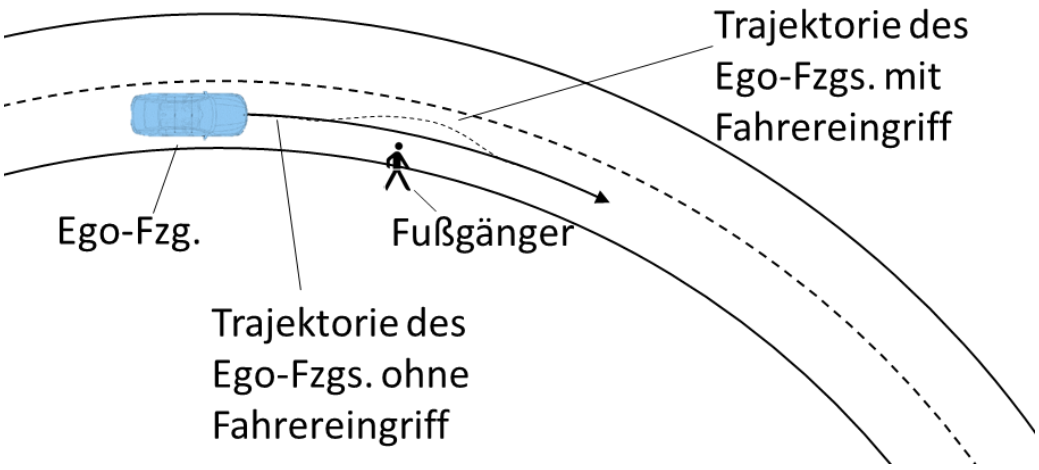

Abbildung 36 Übersichtsgrafik für Szenario 9

Das Egofahrzeug bewegt sich auf einer einspurigen Landstraße ohne Gegenverkehr. Das Fahrzeug übernimmt die Längs- und Querführung des Fahrzeugs während der Fahrer das System überwacht. In einer Rechtskurve geht ein Fußgänger entlang des rechten Fahrbahnrands. Das Assistenzsystem berücksichtigt den Fußgänger nicht. Der Fahrer muss mit der Lenkung und/oder Bremse eingreifen, um eine Kollision mit dem Fußgänger zu vermeiden.

Tabelle 49 Klassifikation von Szenario 10

Merkmal	Ausprägung
Straßentyp	Autobahn zweispurig
Horizontalverlauf	Gerade
Vertikalverlauf	Eben
Knotenpunkte	Keine
Engstellen	Autobahnbaustelle
Assistenz	Längs- und Querführung des Fahrzeugs in der Spur
Automatisierungslevel (BaST)	2
Kritische Systemgrenze	Unangebrachtes Lenkmoment nach rechts

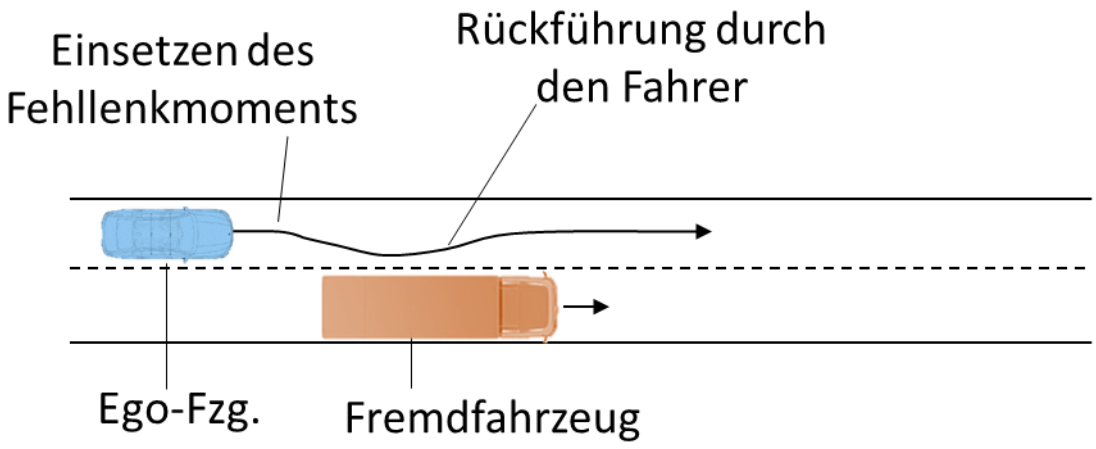

Abbildung 37 Übersichtsgrafik für Szenario 10

Das Egofahrzeug bewegt sich auf einer zweispurigen Autobahn in einem Baustellenbereich. Das Fahrzeug übernimmt die Längs- und Querführung des Fahrzeugs in der Engstelle während der Fahrer das System überwacht. Das Ego-Fahrzeug fährt auf der linken Spur. Auf einem ebenen sowie geraden Streckenabschnitt passiert der Fahrer einen LKW, der auf der rechten Spur fährt. Während dem Überholvorgang führt das Assistenzsystem ein unangebrachtes Lenkmoment nach rechts aus. Der Fahrer muss das Fehllenkmoment abstützen, um zu verhindern, dass das Ego-Fahrzeug mit dem Fremdfahrzeug kollidiert. Nach dem Ende des Fehllenkmoments muss der Fahrer das Fahrzeug weiterhin in der Spur stabilisieren.

Zuordnung der Fahrsituationen zu den Basisszenarien

Jede Fahrsituation aus den vier durchgeführten Versuchen basiert auf einer der zuvor beschriebenen Szenarien. Die folgenden Tabellen geben eine Übersicht, welche Fahrsituation auf welchem Szenario basiert. Wiederholungen des kompletten Satzes an Fahrsituationen werden nicht gesondert aufgeschlüsselt und können dem jeweiligen Versuchsdesign entnommen werden.

Tabelle 50 Zuordnung der Fahrsituationen in Versuch 1 zu den Szenarien

Nummer	1	2	3	4
Szenario	1	2	2	1

Tabelle 51 Zuordnung der Fahrsituationen in Versuch 2 zu den Szenarien

Nummer	1	2	3	4	5	6	7
Szenario	3	4	3	2	4	2	5

Tabelle 52 Zuordnung der Fahrsituationen in Versuch 3 zu den Szenarien

Nummer	1	2	3	4	5
Szenario	3	2	6	4	6

Tabelle 53 Zuordnung der Fahrsituationen in Versuch 4 zu den Szenarien

Nummer	1	2	3	4	5
Szenario	7	3	8	9	10