Autonome Fahrzeuge und moralische Dilemmas: Einflüsse der Perspektive, sozialer Erwünschtheit und der Handelnden

Inaugural-Dissertation

zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Maike M. Mayer
aus Neuss

Düsseldorf, Mai 2022

aus dem Institut für Experimentelle Psychologie
der Heinrich-Heine-Universität Düsseldorf

Gedruckt mit der Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf

Berichterstatter:

1. PD Dr. Raoul Bell
2. Prof. Dr. Axel Buchner

Tag der mündlichen Prüfung: 05.07.2022

Inhaltsverzeichnis

Zusammenfassung

Für eine erfolgreiche Einführung autonomer Fahrzeuge in den Straßenverkehr ist die gesellschaftliche Akzeptanz dieser Technologie erforderlich. Da autonome Fahrzeuge unabhängig von ihrer Leistungsfähigkeit nicht sämtliche Unfälle vermeiden können, könnte das Verhalten autonomer Fahrzeuge in Unfallsituationen deren Akzeptanz beeinflussen. Im Fokus der vorliegenden Arbeit standen zwei Aspekte, die die Wahrnehmung von Unfällen mit autonomen Fahrzeugen beeinflussen können: die präferierten Handlungen autonomer Fahrzeuge und die moralische Handlungsbewertung in Unfallsituationen. Es wurde untersucht, wie sich die Perspektive, aus der eine Unfallsituation betrachtet wird, auf die präferierte Handlung autonomer Fahrzeuge auswirkt (Experimente 1a bis 2b), ob Präferenzen zum Schutz einer größeren Gruppe auf Kosten des eigenen Lebens von sozialer Erwünschtheit beeinflusst werden (Experiment 3) und inwiefern sich die moralische Bewertung von Handlungen autonomer Fahrzeuge und menschlicher Fahrender unterscheidet (Experimente 4 und 5). In allen Experimenten wurden abstrakte Unfallszenarien in Form moralischer Dilemmas verwendet. Die Ergebnisse zeigten, dass die Perspektive die präferierte Handlung autonomer Fahrzeuge beeinflusste. Teilnehmende mit den Perspektiven der in den Unfall involvierten Parteien – Fahrzeuginsassen und Passanten – zeigten eine Tendenz, ihr eigenes Leben zu schützen. Allerdings zeigten sich utilitaristische Tendenzen zum Schutz größerer Gruppen unabhängig von der jeweiligen Perspektive und wirkten Selbstschutztendenzen zu einem gewissen Grad entgegen. Eine utilitaristische Tendenz, sich selbst zugunsten einer größeren Gruppe anderer Verkehrsteilnehmender zu opfern, schien außerdem nicht durch soziale Erwünschtheit beeinflusst zu werden. Des Weiteren beeinflussten utilitaristische Tendenzen auch die moralische Handlungsbewertung von autonomen Fahrzeugen und menschlichen Fahrenden: Je mehr Menschen durch eine Handlung gerettet wurden, desto moralisch vertretbarer wurde diese Handlung bewertet. Allerdings wurden die Handlungen menschlicher Fahrender als moralisch vertretbarer bewertet als die autonomer Fahrzeuge trotz identischer Handlungskonsequenzen. Eine Vermenschlichung autonomer Fahrzeuge führte jedoch zu einer Annäherung der Handlungsbewertungen von menschlichen Fahrenden und autonomen Fahrzeugen. Insgesamt wurden somit in der vorliegenden Arbeit potentielle Schwierigkeiten hinsichtlich der Akzeptanz autonomer Fahrzeuge sowie mögliche Lösungsansätze identifiziert.

Abstract

The successful introduction of autonomous vehicles into daily traffic requires societal acceptance of autonomous driving technologies. As autonomous vehicles—regardless of their overall performance—cannot avoid all accidents, the behaviour of autonomous vehicles in accident situations might affect their acceptance. The present studies focussed on the preferred actions of autonomous vehicles and the moral evaluation of actions in accident scenarios as two aspects that could influence the perception of accidents with autonomous vehicles. It was investigated how the perspective from which an accident scenario was evaluated influenced the preferred action of an autonomous vehicle (Experiments 1a to 2b), whether the preference to protect a larger group of people at the expense of one's life was influenced by social desirability (Experiment 3), and whether the moral evaluation of the actions of autonomous vehicles differed from the evaluation of the actions of human drivers in accident scenarios (Experiments 4 and 5). Abstract accident scenarios in form of moral dilemmas were used in all experiments. The results indicated that the preferred action of autonomous vehicles depended on the participants' perspective. Participants assigned to perspectives of parties involved in the accident—namely passengers and pedestrians—displayed a preference to protect their own lives. However, utilitarian tendencies to protect the larger group were observed irrespective of the assigned perspective and reduced preference differences among perspectives to some extent. Furthermore, a utilitarian tendency to self-sacrifice in order to protect a larger group of road users did not appear to be influenced by a social desirability bias. Utilitarian tendencies also influenced the moral evaluation of the actions of autonomous vehicles and human drivers in accident scenarios: The more people were saved by an action, the more morally justifiable the corresponding action was evaluated. Nevertheless, the actions of a human driver were constantly evaluated as more morally justifiable than the same actions of an autonomous vehicle despite identical action consequences. This evaluation difference between human drivers and autonomous vehicles was reduced by anthropomorphising the autonomous vehicle. In sum, potential difficulties regarding the acceptance of autonomous vehicles as well as possible approaches to encounter these difficulties were identified in the present studies.

Einleitung

»Diese Zeit gehört Dir« – unter diesem Leitspruch startete im Jahr 2015 eine großangelegte Kampagne der Deutschen Bahn, in deren Mittelpunkt die Botschaft stand, dass die Zeit während einer Bahnreise vielfältig und produktiv genutzt werden kann (Deutsche Bahn AG, 2019). Im Gegensatz dazu muss beispielsweise bei Autofahrten die Reisezeit für die eigentliche Fahraufgabe verwendet werden. Doch im Zuge der fortschreitenden Entwicklung automatisierter Fahrsysteme könnte bald auch die Reisezeit auf der Straße für andere Tätigkeiten als die eigentliche Fahraufgabe zur Verfügung stehen. Zumindest zählt die Möglichkeit, die Fahrzeit für andere Aufgaben wie Arbeiten, Lesen oder sogar Schlafen nutzen zu können, zu den möglichen Vorteilen, die typsicherweise mit automatisierten Fahrsystemen verbunden werden (z. B. Anderson et al., 2016; Bagloee et al., 2016; Koopman & Wagner, 2017).

Die Fahraufgabe zumindest vorübergehend an das Fahrzeug zu übergeben, ist erstmals bei Fahrzeugen möglich, deren Systeme Level 3 der Automatisierungsstufen von *SAE International* (2021) erreicht haben. Während bei Fahrzeugen der Level 0 bis 2 die Fahrzeugführenden die Fahraufgabe bzw. Teile davon übernehmen und von technischen Systemen dabei in verschiedenem Umfang unterstützt werden, können Fahrzeuge auf Level 3 unter bestimmten Umständen die Fahraufgabe eigenständig ausführen und somit vorübergehend selbständig fahren, ohne dass die Person im Fahrzeug aktiv an der Fahraufgabe beteiligt ist. Allerdings müssen die Fahrzeugführenden die Fahraufgabe bei einer Aufforderung durch das System oder bei Systemausfällen übernehmen. Bei Fahrzeugen auf Level 4 oder Level 5 ist dies hingegen nicht mehr zwingend erforderlich. Während Fahrzeuge auf Level 4 weiterhin auf bestimmte Anwendungskontexte und Einsatzszenarien beschränkt sind, können Fahrzeuge auf Level 5 unter allen Bedingungen selbstständig fahren. Gemäß dieser Definitionen sind derzeit Fahrzeuge auf Level 3 und Level 4 in Deutschland rechtlich zulässig (vgl. §§ 1a-1b, 1d-1f Straßenverkehrsgesetz)[1].

Vereinfacht können Level 4 und Level 5 der Fahrzeugautomatisierung zusammenfassend auch als autonomer Modus bezeichnet werden, da die Personen im

[1] Straßenverkehrsgesetz in der Fassung der Bekanntmachung vom 5. März 2003 (BGBl. I S. 310, 919), das zuletzt durch Artikel 1 des Gesetzes vom 12. Juli 2021 (BGBl. I S. 3108) geändert worden ist; abrufbar unter: https://www.gesetze-im-internet.de/stvg/BJNR004370909.html (zuletzt abgerufen am 17.05.2022).

Fahrzeug keine fahrbezogenen Aufgaben mehr übernehmen müssen, während bei Level 3 von einem automatisierten Modus gesprochen werden kann (Bundesanstalt für Straßenwesen, 2021), da hier bei Problemen die Person im Fahrzeug die Fahraufgabe übernehmen muss. Im Einklang mit der vereinfachten Definition der Bundesanstalt für Straßenwesen bezeichnet der Begriff »autonome Fahrzeuge« im Folgenden stets Fahrzeuge, die in der Lage sind, den Verkehr eigenständig zu bewältigen und keinen Eingriff durch eine Person im Fahrzeug erfordern (Level 4 und 5).

Neben der Möglichkeit, Autofahrten produktiv nutzen zu können, werden autonome Fahrzeuge meist mit weiteren Vorteilen, wie beispielsweise Treibstoff- und Emissionsreduktionen oder verbesserten Transportmöglichkeiten für Personen, die selbst nicht zur Fahrzeugführung befähigt sind, in Verbindung gebracht (z. B. Bagloee et al., 2016; Spieser et al., 2014; Waldrop, 2015). Vor dem Hintergrund, dass menschliches Versagen und Fehlverhalten zu den Hauptunfallfaktoren im Straßenverkehr zählen (z. B. National Highway Traffic Safety Administration, 2008; Statistisches Bundesamt, 2020), ist mit autonomen Fahrzeugen vor allem auch die Hoffnung auf eine erhöhte Verkehrssicherheit verknüpft (z. B. Anderson et al., 2016). Zwar weisen Auswertungen von Unfallberichten mit autonomen Fahrzeugen, die sich auf kalifornischen Straßen im Feldtest befinden, darauf hin, dass autonome Fahrzeuge derzeit noch im Schnitt häufiger verunfallen (durchschnittlich ca. 67 620 Kilometer pro Unfall) als nicht-autonome Fahrzeuge (durchschnittlich ca. 804 672 Kilometer pro Unfall), jedoch handelte es sich mehrheitlich um Unfälle bei geringen Geschwindigkeiten und um Auffahrunfälle, bei denen ein nicht-autonomes Fahrzeug von hinten auf ein autonomes Fahrzeug auffuhr (Favarò et al., 2017). Laut Favarò et al. legen die Daten der Unfallberichte daher insgesamt nahe, dass autonome Fahrzeuge bereits erfolgreich andere, schwerwiegendere Unfallarten vermeiden. Darüber hinaus galten in der Mehrheit der ausgewerteten Berichte nicht die autonomen Fahrzeuge, sondern andere Verkehrsparteien wie beispielsweise nicht-autonome Fahrzeuge oder nicht-motorisierte Verkehrsteilnehmende als Unfallverursachende (Favarò et al., 2017; Wang et al., 2020). Auch wenn dies die Leistungsfähigkeit autonomer Fahrzeuge in Hinblick auf die Unfallvermeidung stützt, verdeutlicht dieser Aspekt auch, dass autonome Fahrzeuge – unabhängig von ihrer Leistungsfähigkeit und Zuverlässigkeit – nicht alle Unfälle vermeiden können, unter anderem da sie mit anderen Verkehrsteilnehmenden wie Fußgängern oder Tieren, deren Verhalten schwer vorherzusagen ist, oder auch mit nicht-autonomen Fahrzeugen interagieren

(Awad et al., 2018; Goodall, 2014a; Koopman & Wagner, 2017; Lin, 2016; Nyholm, 2018). Hinzu kommt, dass kein technisches System stets fehlerfrei funktioniert (z. B. Gogoll & Müller, 2017; Lin, 2016).

Vor diesem Hintergrund wird deutlich, dass autonome Fahrzeuge auch für den Umgang mit Unfallsituationen programmiert werden müssen (Lin, 2016; Nyholm, 2018), um flächendeckend im Straßenverkehr eingesetzt werden zu können. Hierbei bietet sich ein möglicher Vorteil gegenüber von Menschen gesteuerten Fahrzeugen: Während menschliche Fahrende in Unfallsituationen innerhalb von Sekundenbruchteilen instinktiv reagieren müssen, bietet die Programmierung autonomer Fahrzeuge die Möglichkeit, vorab und ohne eine direkte Bedrohung sorgfältig zu überlegen, wie diese Fahrzeuge in kritischen Situationen reagieren sollen (Bonnefon et al., 2019; Faulhaber et al., 2019; Goodall, 2016a, 2016b; Shariff et al., 2017). Allerdings birgt die Programmierung autonomer Fahrzeuge auch die Gefahr, dass mögliche beabsichtigte oder unbeabsichtigte Verzerrungen in den Algorithmen durch ihren Einsatz in zahlreichen Fahrzeugen einen Einfluss auf Entscheidungen über Leben und Tod haben könnten (Bonnefon et al., 2019; Sütfeld et al., 2017). Dieser Aspekt gewinnt zusätzliche Relevanz vor dem Hintergrund, dass autonome Fahrzeuge auch für Unfallsituationen programmiert werden müssen, die moralische Aspekte umfassen (Awad et al., 2018; Goodall, 2016b; Greene, 2016; Lin, 2016) – also beispielsweise Entscheidungen darüber, wie potentielle Unfallschäden auf verschiedene Unfallparteien verteilt werden sollten bis hin zu möglichen Entscheidungen, wer in einem unvermeidlichen Unfall geopfert werden sollte.

Um moralische Entscheidungen im Kontext autonomer Fahrzeuge zu untersuchen, werden häufig Szenarien eingesetzt, die auf dem sogenannten *Trolley Problem* oder auch *Trolley Dilemma* (Foot, 1967; Thomson, 1976, 1985) basieren. In seiner Grundform behandelt dieses Gedankenexperiment einen außer Kontrolle geratenen Straßenbahnwagon. Die Person am Steuer der Straßenbahn kann diese nicht bremsen und hat daher nur die Möglichkeit, die Straßenbahn auf eins von zwei Gleisen zu lenken. Auf dem Gleis voraus befinden sich fünf Personen, auf einem Nebengleis nur eine Person. Was sollte die Person am Steuer der Straßenbahn in dieser Situation tun? Die Wahl welchen Gleises ist moralisch zulässig? Da sich solche und ähnliche Szenarien gut auf autonome Fahrzeuge übertragen lassen, werden Szenarien, die auf dem *Trolley Problem* basieren, in der aktuellen Forschung sowohl in abstrakter Form als Bilder, Textvignetten oder einer Kombination von beidem (z. B. Awad et al., 2018;

Bonnefon et al., 2016; Frank et al., 2019; Gill, 2020; Li et al., 2016) als auch in Form von virtuellen Simulationen (z. B. Kallioinen et al., 2019) verwendet, um moralische Dilemmas mit autonomen Fahrzeugen zu untersuchen (z. B. Bonnefon et al., 2019; Wolkenstein, 2018). Solche Szenarien dienen dabei nicht als Blaupausen für die Programmierung autonomer Fahrzeuge (z. B. Wolkenstein, 2018), sondern dazu, moralische Intuitionen und Überlegungen sowie Prozesse der moralischen Entscheidungsfindung zu untersuchen und verschiedene ethische Theorien gegenüberzustellen (z. B. Cushman & Greene, 2012; Goodall, 2016a; Hauser et al., 2007; Wolkenstein, 2018). Vor allem die Prinzipien der Deontologie und des Utilitarismus finden in der Literatur zu moralischer Entscheidungsfindung – auch im Zusammenhang mit autonomen Fahrzeugen – häufig Erwähnung (z. B. Faulhaber et al., 2019; Gawronski et al., 2017; Gogoll & Müller, 2017; Goodall, 2014b; Greene et al., 2004). Sie unterscheiden sich insbesondere anhand der Kriterien, nach denen die moralische Zulässigkeit einer Handlung beurteilt wird. Die Deontologie stellt übergeordnete moralische Richtlinien, beispielsweise bestimmte Rechte und Pflichten, in den Fokus (z. B. Kant, 1786/2011). Die moralische Zulässigkeit einer Handlung hängt von deren Vereinbarkeit mit den übergeordneten Richtlinien ab – entspricht eine Handlung den moralischen Richtlinien (z. B. »Du sollst nicht töten«) ist sie moralisch akzeptabel. Im Gegensatz dazu hängt die moralische Zulässigkeit einer Handlung nach dem Prinzip des Utilitarismus von den Konsequenzen der Handlung ab: Eine Handlung ist moralisch zulässig, wenn sie den größtmöglichen Nutzen bzw. das größtmögliche Gute erzielt (z. B. Bentham, 1789; Mill, 1871/2010), beispielsweise indem negative Unfallfolgen reduziert werden.

Des Weiteren ermöglichen auf dem *Trolley Problem* basierende Szenarien, moralisch relevante Faktoren und Eigenschaften verschiedener Unfallsituationen zu identifizieren (z. B. Hauser et al., 2007; Keeling, 2020), die etwa einen Einfluss auf die bevorzugte Handlung des autonomen Fahrzeugs in der dargestellten Situation haben könnten. Beispielsweise untersuchten Awad et al. (2018) in einer großangelegten Online-Studie den Einfluss von neun verschiedenen Faktoren auf die präferierte Handlung eines autonomen Fahrzeugs in verschiedenen dilemmatischen Unfallsituationen. Die Teilnehmenden konnten für jede Unfallsituation zwischen zwei konträren Handlungen des autonomen Fahrzeugs (z. B. Opferung der Personen auf der Straße oder der Personen im Fahrzeug) wählen. Insgesamt wurden fast 40 Millionen Entscheidungen von Teilnehmenden aus 233 Ländern erhoben. Trotz – unter anderem

kulturspezifischer – Varianz in den Handlungspräferenzen konnten die Autoren bei einer Analyse über die verschiedenen Länder hinweg drei Präferenzen für moralische Dilemmas mit autonomen Fahrzeugen identifizieren, die am stärksten ausgeprägt waren: eine Präferenz zum Schutz menschlichen Lebens gegenüber Tieren, eine Präferenz zum Schutz von jüngeren gegenüber älteren Personen sowie eine Präferenz zum Schutz einer größeren gegenüber einer kleineren Personenanzahl. Die Tendenz zum Schutz möglichst vieler Personen steht im Einklang mit dem oben beschriebenen Prinzip des Utilitarismus und kann daher auch als utilitaristische Tendenz betrachtet werden, die oft in moralischen Dilemmas im Kontext autonomer Fahrzeuge beobachtet werden kann (z. B. Awad et al., 2018; Bergmann et al., 2018; Bonnefon et al., 2016; Faulhaber et al., 2019).

Die Untersuchung möglicher Einflussfaktoren auf die Handlungspräferenzen für autonome Fahrzeuge und auf die Bewertung verschiedener Handlungen in Unfallsituationen kann Hinweise geben, wie die Gesellschaft beispielsweise auf bestimmte Ereignisse mit autonomen Fahrzeugen reagiert (z. B. Goodall, 2016a), um so mögliche Schwierigkeiten hinsichtlich der Akzeptanz autonomer Fahrzeuge antizipieren zu können. Dies ist relevant, da eine breite gesellschaftliche Akzeptanz autonomer Fahrzeuge entscheidend für deren erfolgreiche Implementierung im Straßenverkehr ist (z. B. Awad et al., 2018; Bergmann et al., 2018; Bonnefon et al., 2016; Faulhaber et al., 2019; Lin, 2016; Wolkenstein, 2018). Die potentiellen Vorteile autonomer Fahrzeuge, wie beispielsweise eine erhöhte Verkehrssicherheit oder ein verbesserter Verkehrsfluss, setzen voraus, dass autonome Fahrzeuge in großer Anzahl in den Straßenverkehr eingebunden werden (z. B. Franklin et al., 2021; Liu & Liu, 2021). Ohne eine ausreichende öffentliche Akzeptanz ist dies jedoch kaum zu erreichen, da Akzeptanz und Vertrauen die Nutzung einer Technologie beeinflussen (z. B. Lee et al., 2015; Lee & Moray, 1994; Lee & See, 2004; Muir & Moray, 1996; Parasuraman & Riley, 1997; Parasuraman et al., 2008). Die Programmierung autonomer Fahrzeuge für ihr Verhalten im Straßenverkehr und insbesondere hinsichtlich des Umgangs mit Unfallsituationen erfordert daher eine sorgfältige Abwägung dessen, was die Gesellschaft zu akzeptieren bereit ist.

In der vorliegenden Arbeit wurden daher in insgesamt sieben Experimenten verschiedene Einflussfaktoren sowohl auf die präferierte Handlung autonomer Fahrzeuge in dilemmatischen Verkehrssituationen (Experimente 1a bis 3) als auch auf die Bewertung verschiedener bereits geschehener Handlungen untersucht (Experimente

4 und 5). Zunächst wurde ein möglicher Einfluss der persönlichen Perspektive auf die präferierte Handlung in dilemmatischen Verkehrssituationen mit autonomen Fahrzeugen geprüft (Experimente 1a bis 2b), indem eine potentiell mit der eingenommenen Perspektive einhergehende Tendenz zum Selbstschutz verglichen mit anderen Verkehrsteilnehmenden möglichen utilitaristischen Präferenzen gegenübergestellt wurde. Weiterführend wurde in Experiment 3 unter Berücksichtigung der persönlichen Perspektive untersucht, inwiefern die oft beobachteten utilitaristischen Präferenzen möglicherweise auf soziale Erwünschtheit zurückzuführen sind. Die Experimente 4 und 5 konzentrierten sich schließlich auf die moralische Bewertung bereits geschehener Handlungen in Unfallsituationen. Hierbei wurde untersucht, ob die moralische Bewertung der Handlungen davon beeinflusst wird, ob menschliche Fahrende oder autonome Fahrzeuge handeln.

Einfluss der Perspektive

Autonome Fahrzeuge so zu konzipieren, dass möglichst wenig Schaden verursacht und im Ernstfall die größtmögliche Anzahl an Menschenleben geschützt wird, könnte gesamtgesellschaftlich betrachtet sinnvoll erscheinen. Ein solcher Ansatz entspräche dem Prinzip des Utilitarismus und korrespondierende Präferenzen lassen sich in vielen Studien, die moralische Dilemmas im Kontext autonomer Fahrzeuge untersuchen, beobachten (z. B. Awad et al., 2018; Bergmann et al., 2018; Bonnefon et al., 2016; Faulhaber et al., 2019). Es gibt jedoch Hinweise darauf, dass Präferenzen in Dilemmas nicht nur von moralischen Überlegungen – wie beispielsweise den in der vorliegenden Arbeit schwerpunktmäßig untersuchten utilitaristischen Überlegungen – geleitet werden, sondern auch selbstschützende Tendenzen die Entscheidungsfindung beeinflussen können (Volz et al., 2017). Daraus ließe sich eine Präferenz für Handlungen ableiten, die das eigene Leben schützen.

Im Kontext autonomer Fahrzeuge würde dies bedeuten, dass Nutzende Fahrzeuge präferieren, die den Schutz des eigenen Lebens gegenüber dem Schutz des Lebens anderer Verkehrsteilnehmender priorisieren. Im Einklang damit beobachteten Bonnefon et al. (2016) eine Diskrepanz zwischen der Handlungspräferenz für autonome Fahrzeuge und der Bereitschaft, diese zu kaufen: In mehreren Experimenten zeigten die Versuchsteilnehmenden eine Präferenz für einen utilitaristischen Algorithmus in autonomen Fahrzeugen. Nichtsdestotrotz waren sie weniger bereit ein

utilitaristisches Fahrzeug zu kaufen als ein Fahrzeug, dass die Personen im Fahrzeug schützt. Auch Liu und Liu (2021) beobachteten eine höhere Bereitschaft, ein autonomes Fahrzeug zu nutzen, das die Personen im Fahrzeug schützt, verglichen mit einem utilitaristischen autonomen Fahrzeug. Es zeigten sich außerdem Hinweise darauf, dass die Teilnehmenden eher bereit wären, einen Aufpreis für ein autonomes Fahrzeug zu bezahlen, das sie selbst schützen würde, als für ein autonomes Fahrzeug mit einem utilitaristischen Algorithmus.

Diese Datenmuster lassen sich als Hinweis auf ein soziales Dilemma bei der Programmierung autonomer Fahrzeuge interpretieren (Bonnefon et al., 2016). Ein soziales Dilemma bezeichnet einen Konflikt zwischen den (Eigen-)Interessen des Individuums und den Interessen der Gemeinschaft. Wenn Individuen ihren eigenen Interessen folgen – also unabhängig von den sozialen Konsequenzen auf den eigenen Vorteil bedacht agieren – führt dies zu einem schlechteren Ergebnis für die gesamte Gesellschaft als ohne Fokus auf den eigenen Vorteil (Dawes, 1980; Messick & Brewer, 1983). Demnach könnte es zwar aus Sicht der Gesellschaft sinnvoll sein, den Schaden bei Unfällen insgesamt möglichst gering zu halten und durch eine utilitaristische Programmierung die Anzahl der gefährdeten Personen zu reduzieren, jedoch könnten Käuferinnen und Käufer autonomer Fahrzeuge – unabhängig von der Anzahl gefährdeter Personen – Präferenzen zum Schutz ihres eigenen Lebens zeigen (Faulhaber et al., 2019; Gogoll & Müller, 2017). In der Folge könnte dies dazu führen, dass sich autonome Fahrzeuge auf dem Markt etablieren oder durchsetzen, die stets die Personen im Fahrzeug schützen und damit deren Leben wichtiger gewichten als das Leben anderer Verkehrsteilnehmender.

Eine alternative Position zu dem prioritären Schutz der Personen im Fahrzeug nimmt die Ethik-Kommission des Bundesministeriums für Verkehr und digitale Infrastruktur ein (Bundesministerium für Verkehr und digitale Infrastruktur, 2017). Die Ethik-Kommission erarbeitete im Auftrag des Bundesministeriums ethische Richtlinien für automatisierte bis hin zu autonomen Fahrzeugen, die unter anderem auch die Programmierung autonomer Fahrzeuge in unausweichlichen Unfallsituationen umfassen. Laut der Ethik-Kommission ist bei Unfällen dem Schutz menschlichen Lebens höchste Priorität zuzuschreiben und Personenschäden sind unbedingt zu vermeiden (vgl. Regel 7; Bundesministerium für Verkehr und digitale Infrastruktur, 2017). Interessanterweise differenziert die Ethik-Kommission explizit zwischen Sicherheitsinteressen verschiedener Verkehrsteilnehmender: »(...) Die an der

Erzeugung von Mobilitätsrisiken Beteiligten dürfen Unbeteiligte nicht opfern.« (Bundesministerium für Verkehr und digitale Infrastruktur, 2017, Regel 9, S. 11). Personen, die autonome Fahrzeuge nutzen und damit in den Straßenverkehr einführen, sind für die mit der Nutzung einhergehenden Risiken – auch für andere Verkehrsteilnehmende – verantwortlich. Daher dürfen Verkehrsteilnehmende außerhalb des autonomen Fahrzeugs nach Dafürhalten der Ethik-Kommission nicht zugunsten der Personen im Fahrzeug geopfert werden.

Die von der Ethik-Kommission bezogene, zu möglichen Selbstschutztendenzen der Personen im Fahrzeug konträre Position verdeutlicht das Konfliktpotential zwischen verschiedenen Gruppen im Straßenverkehr. Ein Interessenskonflikt zwischen Verkehrsteilnehmenden hinsichtlich der bevorzugten Programmierung autonomer Fahrzeuge für den Umgang mit Unfallsituationen, dadurch dass alle Parteien auf ihren jeweiligen Selbstschutz bestehen, könnte sich negativ auf die Akzeptanz autonomer Fahrzeuge auswirken. Derzeit gibt es jedoch nur wenige Studien, die untersuchen, in welchem Ausmaß sich moralische Präferenzen von motorisierten und nicht-motorisierten Verkehrsteilnehmenden für die Handlungen autonomer Fahrzeuge in unvermeidbaren Unfallsituationen unterscheiden. Die bisherige Forschung konzentrierte sich hauptsächlich auf die Perspektive der Personen im Fahrzeug und auf die Perspektive von Beobachtenden (z. B. Awad et al., 2018; Bonnefon et al., 2016; Li et al., 2016). Jedoch sind auch andere Verkehrsteilnehmende – darunter vor allem Passantinnen und Passanten als größte nicht-motorisierte Gruppe (Nobis & Kuhnimhof, 2019) – unmittelbar von der Programmierung autonomer Fahrzeuge betroffen und könnten sich in ihren Präferenzen – unter anderem auch hinsichtlich möglicher Tendenzen zum Selbstschutz – von Personen im Fahrzeug oder von unbeteiligten Beobachtenden unterscheiden. Eine Studie von Kallioinen et al. (2019) weist beispielsweise darauf hin, dass Passantinnen und Passanten selbstschützende Tendenzen haben könnten. Verglichen mit einer Gefährdung der Personen auf der Straße zeigten Teilnehmende, die dilemmatische Unfallsituationen aus Sicht von Passantinnen oder Passanten beurteilten, eine Präferenz für die Opferung der Personen im Fahrzeug – vor allem in Situationen, in denen das Leben der Personen im Fahrzeug und das der Personen auf der Straße gegeneinander abgewogen wurde.

Allerdings ist bei den Ergebnissen von Kallioinen et al. (2019) zu beachten, dass der Einfluss der Perspektive in immersiven virtuellen Umgebungen untersucht wurde, wodurch die Bedrohung für die eigene Perspektive besonders salient war

und selbstschützende Tendenzen der Versuchsteilnehmenden verstärkt worden sein könnten. Die Ergebnisse von Frank et al. (2019) weisen darauf hin, dass auch bei der Beurteilung von abstrakten moralischen Dilemmas mit autonomen Fahrzeugen selbstschützende Tendenzen vorliegen könnten: Teilnehmende, die die Perspektive der Person im Fahrzeug einnahmen, waren eher bereit, die Person auf der Straße zu opfern als Teilnehmende, die sich in die Rolle einer Passantin bzw. eines Passanten hineinversetzen sollten. Allerdings zeigte die Mehrheit der Teilnehmenden eine Präferenz zum Schutz der Passantin bzw. des Passanten gegenüber der Person im Fahrzeug, selbst wenn die Teilnehmenden sich in die Person im Fahrzeug hineinversetzen sollten.

Neben möglichen Tendenzen zum Schutz des eigenen Lebens finden sich in den Ergebnissen von Kallioinen et al. (2019) und Frank et al. (2019) auch erste Hinweise darauf, dass es möglicherweise Prinzipien geben könnte, die Selbstschutztendenzen verschiedener Verkehrsteilnehmender entgegenwirken. So zeigten sowohl Teilnehmende, die die Perspektive der Passantinnen bzw. Passanten einnahmen, und Teilnehmende mit der Perspektive der Fahrzeuginsassinnen bzw. -insassen in der Studie von Kallioinen et al. (2019) Tendenzen, möglichst viele Leben zu retten und somit utilitaristische Präferenzen. Ähnliches konnten auch Frank et al. (2019) beobachten: Bei einer Manipulation der Personenzahl auf der Straße und im Fahrzeuginneren zeichneten sich utilitaristische Präferenzen zum Schutz der größeren Gruppe ab. Auch bei Faulhaber et al. (2019) finden sich Hinweise darauf, dass Menschen nicht um jeden Preis ihr eigenes Leben schützen wollen. Sie beobachteten eine steigende Bereitschaft zur Selbstopferung, je mehr Personen durch diese Handlung gerettet werden konnten.

Zusammengenommen legen die Ergebnisse von Faulhaber et al. (2019), Frank et al. (2019) und Kallioinen et al. (2019) Grenzen für den Einfluss von Selbstschutztendenzen auf die präferierte Handlung von autonomen Fahrzeugen in moralischen Dilemmas nahe. Inwiefern die beobachteten utilitaristischen Präferenzen als Gegengewicht zu selbstschützenden Tendenzen verschiedener Verkehrsteilnehmender fungieren können, ist bisher jedoch unklar. In den Experimenten 1a bis 2b der vorliegenden Arbeit wurde daher der Einfluss der Perspektive in einem moralischen Dilemma mit autonomen Fahrzeugen auf die Präferenzen für die Handlungen des autonomen Fahrzeugs untersucht. Zum einen wurde geprüft, ob sich verschiedene Perspektiven

(Passanten-, Beobachter-, oder Fahrzeuginsassen-Perspektive)[2] in ihren Handlungspräferenzen für autonome Fahrzeuge unterscheiden (Frank et al., 2019; Kallioinen et al., 2019). Zum anderen wurden mit der eingenommenen Perspektive einhergehende Tendenzen zum Selbstschutz systematisch utilitaristischen Überlegungen gegenübergestellt, um zu untersuchen, inwiefern utilitaristische Tendenzen den Selbstschutztendenzen verschiedener Verkehrsteilnehmender entgegen wirken können. Dazu wurde sowohl die Anzahl der Personen auf der Straße (Experimente 1a und 1b) als auch die Anzahl der Personen im Fahrzeug (Experimente 2a und 2b) manipuliert. Die Versuchsteilnehmenden wurden in allen vier Experimenten randomisiert einer der drei Perspektiven zugeteilt, aus der sie die jeweiligen Szenarien beurteilen und ihre präferierte Handlung des autonomen Fahrzeugs angeben sollten. Um die Robustheit der jeweiligen Befunde zu überprüfen (vgl. Open Science Collaboration, 2015) und auf eine breitere empirische Basis zu stellen, wurden die Experimente 1a und 2a (hauptsächlich studentische Teilnehmende) in den Experimenten 1b und 2b (Stichproben aus Forschungspanels) repliziert (für zusätzliche Informationen zu den Stichproben der Experimente 1a bis 2b siehe Online-Materialien zu Mayer et al., 2021). Da es sich um direkte Replikationen handelt und somit das verwendete Material und der experimentelle Ablauf jeweils identisch sind, werden die Experimente 1a und 1b sowie die Experimente 2a und 2b im Folgenden zusammen betrachtet.

Experimente 1a und 1b

Um zu untersuchen, inwieweit Selbstschutztendenzen und utilitaristische Überlegungen die präferierte Handlung eines autonomen Fahrzeugs in einer dilemmatischen Verkehrssituation beeinflussen, wurden die Teilnehmenden in den Experimenten 1a und 1b gebeten, verschiedene unvermeidliche Unfallsituationen mit autonomen Fahrzeugen zu beurteilen. Den Teilnehmenden wurden abstrakte Bilder präsentiert (z. B. Awad et al., 2018; Frank et al., 2019), in denen verschiedene Situationen aus der Vogelperspektive abgebildet wurden. In allen Szenarien fuhr ein autonomes Fahrzeug mit einer Person im Fahrzeuginneren auf einer einspurigen Straße, auf der sich ein Hindernis und eine Person oder eine Personengruppe bestehend aus zwei,

[2] Der Übersichtlichkeit halber sind im Folgenden die Perspektiven bzw. die Experimentalbedingungen mit der maskulinen Wortform bezeichnet (z. B. »Passanten-Perspektive«) ebenso wie die Verhältnisse der verschiedenen Personengruppen zueinander (»Fahrzeuginsassen-zu-Passanten-Verhältnis«). Die gewählte Bezeichnung schließt explizit Personen jeglichen Geschlechts ein.

fünf oder zehn Personen befand. Die Aufgabe der Teilnehmenden bestand darin, für jedes Szenario anzugeben, ob das autonome Fahrzeug in der dargestellten Situation entweder die Person im Fahrzeug oder die Person/en auf der Straße opfern sollte. Aus den Häufigkeiten, mit denen die beiden Optionen in den verschiedenen Szenarien gewählt wurden, wurde für jede Perspektive und jedes Fahrzeuginsassen-zu-Passanten-Verhältnis die Wahrscheinlichkeit, die Person im Fahrzeug zu opfern, geschätzt und mithilfe einfacher Entscheidungsbäume verglichen (für Details siehe Mayer et al., 2021). Für alle im Folgenden berichteten statistischen Analysen (auch in den Experimenten 2a bis 5) lag das kritische Alpha-Niveau bei $\alpha = .05$. Bei Mehrfachvergleichen wurde eine Bonferroni-Holm-Korrektur (Holm, 1979) des Alpha-Niveaus vorgenommen. Dem Verfahren von Holm (1979) folgend wurden die Einzelvergleiche sequentiell basierend auf der Größe der jeweiligen *p*-Werte gegen ein individuelles kritisches Alpha-Niveau ($\leq .05$) auf Signifikanz geprüft.

Sofern die beiden in den Unfall involvierten Perspektiven eine Selbstschutztendenz zeigen, sollte sich dies in einer Präferenz zur Opferung der jeweils anderen Unfallpartei bzw. in einem Widerwillen zur Selbstopferung äußern. Teilnehmende, die die Fahrzeuginsassen-Perspektive einnahmen, sollten in diesem Fall beispielsweise eine geringere Wahrscheinlichkeit zur Opferung der Person im Fahrzeug zeigen als Teilnehmende, die die Passanten-Perspektive einnahmen. Die deskriptiven Statistiken für die Wahrscheinlichkeit, die Person in Fahrzeug zu opfern, sind in Abbildung 1 dargestellt. Hypothesenkonform zeigten sowohl in Experiment 1a als auch in Experiment 1b Teilnehmende, die die Fahrzeuginsassen-Perspektive einnahmen, über alle Fahrzeuginsassen-zu-Passanten-Verhältnisse hinweg eine signifikant geringere Wahrscheinlichkeit, die Person im Fahrzeug (also sich selbst) zu opfern als Teilnehmende, die die Passanten-Perspektive [Experiment 1a: $G^2(4) = 326.60$, $p < .001$, $w = .26$; Experiment 1b: $G^2(4) = 292.34$, $p < .001$, $w = .23$] oder die Beobachter-Perspektive einnahmen [Experiment 1a: $G^2(4) = 146.58$, $p < .001$, $w = .17$; Experiment 1b: $G^2(4) = 149.58$, $p < .001$, $w = .17$]. Teilnehmende, die die Passanten-Perspektive einnahmen, zeigten hingegen in beiden Experimenten die höchste Wahrscheinlichkeit, die Person im Fahrzeug zu opfern, die signifikant über der der Beobachter-Perspektive lag [Experiment 1a: $G^2(4) = 43.01$, $p < .001$, $w = .09$; Experiment 1b: $G^2(4) = 30.58$, $p < .001$, $w = .07$]. Somit lässt sich ein signifikanter Effekt der Perspektive auf die Wahrscheinlichkeit, die Person im Fahrzeug zu opfern, beobachten. Der deutliche Unterschied zwischen den involvierten Perspektiven der Fahrzeuginsassen und der

Passanten in der Wahrscheinlichkeit, die Person im Fahrzeug zu opfern, den Abbildung 1 illustriert, stützt die Hypothese, dass Selbstschutztendenzen die präferierten Handlungen autonomer Fahrzeuge in moralischen Dilemmas beeinflussen.

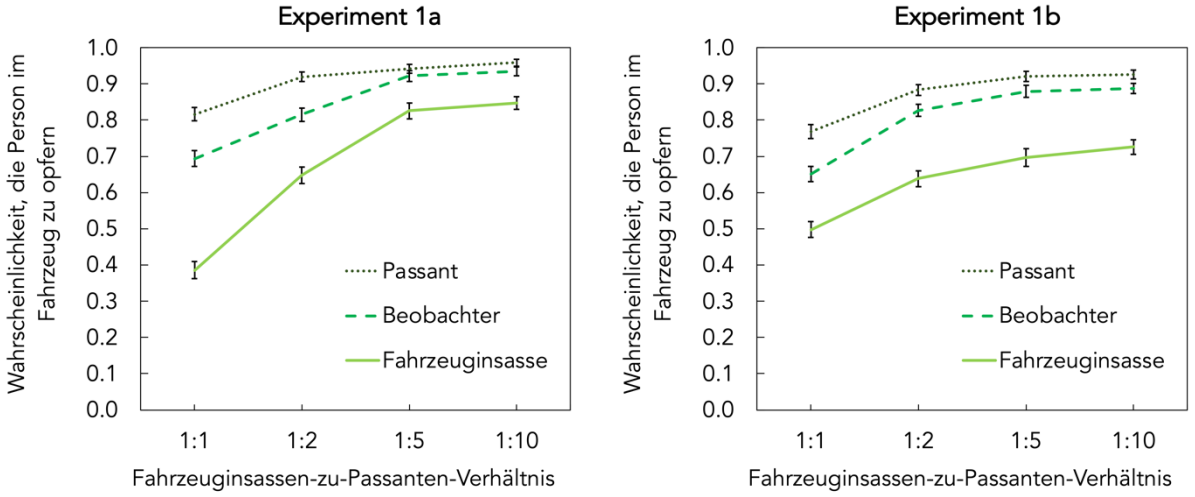


Abbildung 1. Die Wahrscheinlichkeit, die Person im Fahrzeug anstatt der Person/en auf der Straße zu opfern, als Funktion des Fahrzeuginsassen-zu-Passanten-Verhältnisses (1:1, 1:2, 1:5, 1:10) und der eingenommenen Perspektive (Passant, Beobachter, Fahrzeuginsasse) für die Experimente 1a und 1b. Die Fehlerbalken repräsentieren die Standardfehler (bootstrapped).

Die Manipulation der Personenanzahl auf der Straße ermöglichte es zusätzlich zu dem Einfluss von Selbstschutztendenzen auch utilitaristische Präferenzen der verschiedenen Perspektiven zu untersuchen. Eine Präferenz für utilitaristische Handlungen sollte sich in einer mit zunehmender Personenanzahl auf der Straße steigenden Wahrscheinlichkeit zur Opferung der Person im Fahrzeug äußern. Sofern Utilitarismus ein Prinzip darstellt, auf das sich die Teilnehmenden perspektivübergreifend einigen können, sollten sich die Unterschiede in der Wahrscheinlichkeit, die Person im Fahrzeug zu opfern, zwischen den Perspektiven mit steigender Anzahl an Personen auf der Straße verringern oder gar ausbleiben. Im Einklang mit dieser Annahme ist in Abbildung 1 zu erkennen, dass die Wahrscheinlichkeit, die Person im Fahrzeug zu opfern, mit steigender Personenanzahl auf der Straße für alle Perspektiven zunahm. Dies legt utilitaristische Präferenzen für alle Perspektiven nahe. Die Vergleiche der Perspektiven für die einzelnen Fahrzeuginsassen-zu-Passanten-Verhältnisse getrennt zeigten in beiden Experimenten jedoch, dass die Unterschiede zwischen den involvierten Perspektiven nicht vollständig durch die beobachteten utilitaristischen

Präferenzen eliminiert wurden [für die Vergleiche zwischen Fahrzeuginsassen- und Passanten-Perspektive: alle $p < .001$ in beiden Experimenten; für Details siehe Tabellen 1 und 2 in Mayer et al., 2021]. Auch für das extremste Fahrzeuginsassen-zu-Passanten-Verhältnis von 1:10 unterschieden sich die Präferenzen der Fahrzeuginsassen- und der Passanten-Perspektive. Lediglich die Beobachter- und die Passanten-Perspektive näherten sich soweit an, dass ab einem Fahrzeuginsassen-zu-Passanten-Verhältnis von 1:5 keine signifikanten Unterschiede zwischen den beiden Perspektiven mehr bestanden [Fahrzeuginsassen-zu-Passanten-Verhältnis 1:5: Experiment 1a: $G^2(1) = 1.04$, $p = .308$, $w = .01$; Experiment 1b: $G^2(1) = 3.55$, $p = .060$, $w = .03$; Fahrzeuginsassen-zu-Passanten-Verhältnis 1:10: Experiment 1a: $G^2(1) = 2.54$, $p = .111$, $w = .02$; Experiment 1b: $G^2(1) = 4.31$, $p = .038$, $w = .03$; alle anderen $p < .015$]. Dies stützt den Einfluss von Selbstschutztendenzen auf die Präferenzen in der Beurteilung von moralischen Dilemmas mit autonomen Fahrzeugen. Zusätzlich weisen die Ergebnisse darauf hin, dass utilitaristische Tendenzen die Präferenzunterschiede zwischen den Perspektiven reduzieren, jedoch nicht vollständig eliminieren können.

In den ersten beiden Experimenten wurde allerdings nur die Anzahl der Personen auf der Straße manipuliert und einer einzelnen Person im Fahrzeuginneren gegenübergestellt. Wie sich eine Manipulation der Anzahl von Personen im Fahrzeug auf die Präferenzen der verschiedenen Perspektiven und den generellen Einfluss der Perspektive auswirkt, wurde daher in den Experimenten 2a und 2b untersucht.

Experimente 2a und 2b

Vor allem vor dem Hintergrund der Richtlinien der Ethik-Kommission des Bundesministeriums für Verkehr und digital Infrastruktur, dass an der Erzeugung von Mobilitätsrisiken beteiligte Personen (z. B. Fahrzeuginsassinnen und -insassen) Personen ohne Verantwortung für die Mobilitätsrisiken autonomer Fahrzeuge (z. B. Passantinnen und Passanten) nicht opfern dürfen (Bundesministerium für Verkehr und digitale Infrastruktur, 2017), ist interessant, inwiefern die Präferenzen von moralischen Laien der Einschätzung der Ethik-Kommission entsprechen. Zeigen Teilnehmende mit der Passanten-Perspektive bei einer variierenden Anzahl an Personen im Fahrzeug eine ähnliche Bereitschaft, sich selbst zugunsten einer größeren Gruppe anderer Verkehrsteilnehmender zu opfern, wie Teilnehmende mit der Fahrzeuginsassen-Perspektive bei einer variierenden Anzahl an Personen auf der Straße in den Experimenten 1a und 1b? Oder zeigt sich der Empfehlung der Ethik-Kommission

folgend – vor allem in der Passenten-Perspektive – eine persistierende Präferenz zum Schutz der Personen auf der Straße unabhängig von der Anzahl der Personen im Fahrzeug?

Um dies zu untersuchen wurde das methodische Vorgehen der Experimente 1a und 1b übernommen, aber nun die Anzahl der Personen im Fahrzeug variiert, während sich stets nur eine Person auf der Straße befand. Sofern der Einfluss der Perspektive stabil ist, sollte sich die Wahrscheinlichkeit, die Person/en im Fahrzeug zu opfern, erneut zwischen der Fahrzeuginsassen- und der Passanten-Perspektive unterscheiden. Für die Fahrzeuginsassen-Perspektive sollte sich im Sinne einer Selbstschutztendenz die geringste und für die Passanten-Perspektive die höchste Wahrscheinlichkeit zur Opferung der Person/en im Fahrzeug zeigen. Wenn des Weiteren auch für Personengruppen im Fahrzeug utilitaristische Tendenzen greifen und einen Einfluss auf die perspektivenspezifischen Präferenzen haben, sollte sich ein im Vergleich zu den Experimenten 1a und 1b umgekehrter Kurvenverlauf zeigen: Mit einer steigenden Personenanzahl im Fahrzeug sollte die für die Perspektiven beobachtete Wahrscheinlichkeit, die Person/en im Fahrzeug zu opfern, abnehmen, da sich die größere – und nach dem utilitaristischen Prinzip somit zu schützende – Gruppe nun innerhalb statt wie in den Experimenten 1a und 1b außerhalb des autonomen Fahrzeugs befindet.

Abbildung 2 illustriert, dass – wie in den Experimenten 1a und 1b – Teilnehmende, die die Fahrzeuginsassen-Perspektive einnahmen, in beiden Experimenten durchgehend die geringste Wahrscheinlichkeit, die Person/en im Fahrzeug zu opfern, zeigten, während Teilnehmende, die die Passanten-Perspektive einnahmen, konstant die höchste Wahrscheinlichkeit und somit die stärkste Präferenz zur Opferung der Person/en im Fahrzeug aufwiesen. Die Präferenzen der Teilnehmenden mit der Fahrzeuginsassen- und der Passanten-Perspektive unterschieden sich auch inferenzstatistisch signifikant voneinander [Experiment 2a: $G^2(4) = 243.30$, $p < .001$, $w = .22$; Experiment 2b: $G^2(4) = 422.06$, $p < .001$, $w = .26$]. Dies deutet erneut auf Selbstschutztendenzen der beiden involvierten Perspektiven hin. Darüber hinaus unterschied sich sowohl die für die Passanten-Perspektive [Experiment 2a: $G^2(4) = 69.74$, $p < .001$, $w = .12$; Experiment 2b: $G^2(4) = 372.76$, $p < .001$, $w = .25$], als auch für die Fahrzeuginsassen-Perspektive [Experiment 2a: $G^2(4) = 59.58$, $p < .001$, $w = .11$; Experiment 2b: $G^2(4) = 13.89$, $p = .008$, $w = .05$] beobachtete Wahrscheinlichkeit, die Person/en im Fahrzeug zu opfern, signifikant von der für die Beobachter-

Perspektive, die zwischen den Wahrscheinlichkeiten der beiden involvierten Perspektiven lag.

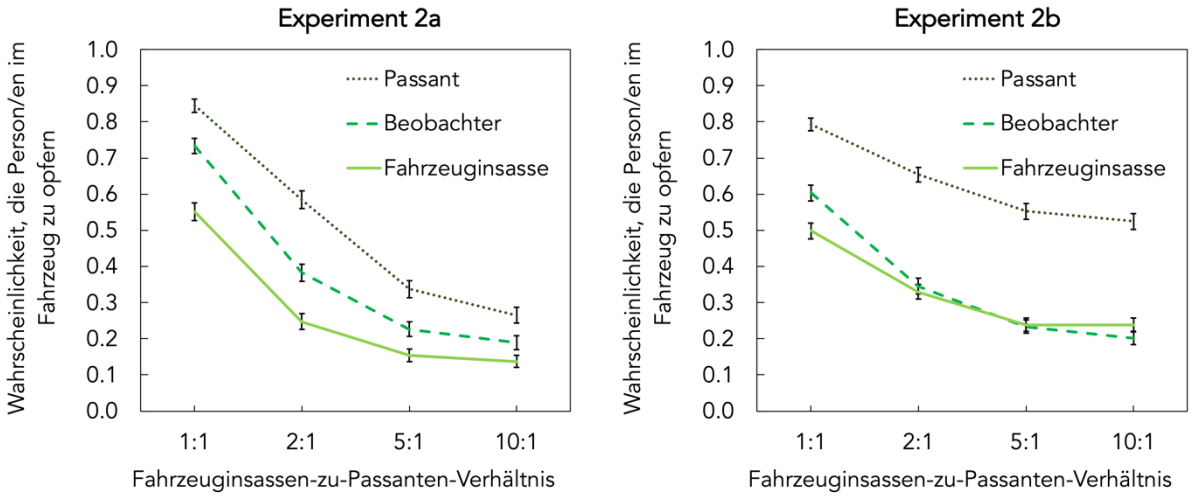


Abbildung 2. Die Wahrscheinlichkeit, die Person/en im Fahrzeug anstatt der Person auf der Straße zu opfern, als Funktion des Fahrzeuginsassen-zu-Passanten-Verhältnisses (1:1, 2:1, 5:1, 10:1) und der eingenommenen Perspektive (Passant, Beobachter, Fahrzeuginsasse) für die Experimente 2a und 2b. Die Fehlerbalken repräsentieren die Standardfehler (bootstrapped).

Des Weiteren ist in Abbildung 2 ein im Vergleich zu den Experimenten 1a und 1b umgekehrter Kurvenverlauf zu erkennen – mit steigender Personenanzahl im Fahrzeug sank für alle drei Perspektiven die Präferenz für die Opferung der Personen im Fahrzeug zugunsten der Person auf der Straße. Dies legt utilitaristische Tendenzen nahe, die bereits in den ersten beiden Experimenten – dort allerdings zugunsten der Personen auf der Straße – beobachtet wurden. In Teilen weicht das Datenmuster somit von einer situationsunabhängigen Präferenz zum Schutz der Passantinnen und Passanten gegenüber Fahrzeuginsassinnen und -insassen ab, die die Richtlinien der Ethik-Kommission des Bundesministeriums für Verkehr und digitale Infrastruktur vielleicht nahegelegt hätten (Bundesministerium für Verkehr und digitale Infrastruktur, 2017): Obwohl bei einem Fahrzeuginsassen-zu-Passanten-Verhältnis von 1:1 die Mehrheit der Teilnehmenden – über alle Perspektiven hinweg betrachtet – im Einklang mit den Richtlinien der Ethik-Kommission eine Präferenz zum Schutz der einzelnen Person auf der Straße zeigte, sank mit einer zunehmenden Personenanzahl im Fahrzeug die Bereitschaft, diese zugunsten der Person auf der Straße zu opfern. Allerdings eliminierten diese utilitaristischen Tendenzen, die sich erneut

unabhängig von der Perspektive beobachten ließen, auch hier die Unterschiede zwischen den Perspektiven der beiden involvierten Parteien – selbst bei dem extremsten Fahrzeuginsassen-zu-Passanten-Verhältnis von 10:1 – nicht vollständig [für alle Vergleiche der Fahrzeuginsassen- und Passanten-Perspektive: alle $p < .001$; siehe Tabellen 3 und 4, Mayer et al., 2021]. Lediglich in Experiment 2b näherten sich die Beobachter- und die Fahrzeuginsassen-Perspektive soweit an, dass sie sich ab einem Fahrzeuginsassen-zu-Passanten-Verhältnis von 2:1 nicht mehr signifikant voneinander unterschieden [Fahrzeuginsassen-zu-Passanten-Verhältnis 2:1: $G^2(1) = 0.29$, $p = .590$, $w = .01$; Fahrzeuginsassen-zu-Passanten-Verhältnis 5:1: $G^2(1) = 0.03$, $p = .854$, $w < .01$; Fahrzeuginsassen-zu-Passanten-Verhältnis 10:1: $G^2(1) = 2.08$, $p = .149$, $w = .02$; alle anderen $p < .050$]. Insgesamt stützt das Datenmuster jedoch die Stabilität der Präferenzunterschiede zwischen den beiden involvierten Perspektiven.

Diskussion der Experimente 1a bis 2b

Zusammengefasst sind die Ergebnisse der Experimente 1a, 1b, 2a und 2b sehr konsistent. Die Befunde zeigen, dass die Perspektive, aus der moralische Dilemmas mit autonomen Fahrzeugen betrachtet werden, einen Einfluss auf die bevorzugte Handlung des Fahrzeugs in den dargestellten Situationen hat. Vor allem die Handlungspräferenzen der involvierten Parteien unterschieden sich in den vorliegenden Experimenten deutlich voneinander: Teilnehmende, die die Szenarien aus der Passanten-Perspektive beurteilten, zeigten insgesamt die höchste Wahrscheinlichkeit und somit die stärkste Präferenz zur Opferung der Person/en im Fahrzeug. Teilnehmende, die die Fahrzeuginsassen-Perspektive einnahmen, zeigten hingegen insgesamt die geringste Wahrscheinlichkeit, die Person/en im Fahrzeug (und damit sich selbst) zu opfern. Die ausgeprägte Divergenz der für die beiden involvierten Parteien beobachteten Handlungspräferenzen lässt sich als Tendenz zum Selbstschutz interpretieren. Die Ergebnisse stützen somit die Befunde von Frank et al. (2019) und Kallioinen et al. (2019).

Trotz der ausgeprägten Perspektivunterschiede in den Handlungspräferenzen für autonome Fahrzeuge legen die Befunde eine Annäherung der perspektivenspezifischen Präferenzen mit steigender Personenanzahl und daher mit einem zunehmend unausgeglichenen Fahrzeuginsassen-zu-Passanten-Verhältnis nahe: Je mehr Menschen durch die Opferung einer einzelnen Person gerettet werden konnten, desto höher war die Wahrscheinlichkeit, die größere Gruppe zu retten, für alle Perspektiven.

Dies lässt sich als utilitaristische Präferenz interpretieren, welche sich auch in anderen Studien zeigt (z. B. Awad et al., 2018; Bonnefon et al., 2016; Faulhaber et al., 2019). Des Weiteren weist die für alle Perspektiven beobachtete utilitaristische Tendenz darauf hin, dass eine gewisse Annäherung zwischen den verschiedenen Perspektiven erreicht werden kann. Bis zu einem gewissen Grad wirken utilitaristische Präferenzen den Selbstschutztendenzen der involvierten Parteien entgegen.

Insgesamt lässt sich das beobachtete Befundmuster daher als durch Selbstschutztendenzen beeinflusster Utilitarismus beschreiben: Viele Teilnehmende zeigten eine Präferenz, möglichst viele Leben zu retten (siehe z. B. auch Awad et al., 2018; Bergmann et al., 2018; Bonnefon et al., 2016; Faulhaber et al., 2019), auch wenn dies damit einherging, sich selbst opfern zu müssen. Diese utilitaristischen Präferenzen wurden jedoch von einer Tendenz zum Schutz des eigenen Lebens beeinflusst. Zumeist mussten mehrere Menschenleben in Gefahr sein, bevor eine Selbstopferung mehrheitlich gegenüber selbstschützenden Alternativen präferiert wurde (Faulhaber et al., 2019). Dieses Datenmuster ist nicht auf die Fahrzeuginsassen-Perspektive beschränkt, was vor allem mit Blick auf die Richtlinien der Ethik-Kommission des Bundesministeriums für Verkehr und digitale Infrastruktur interessant ist: Basierend auf diesen Richtlinien sollten Personen auf der Straße in keinem Szenario zugunsten der Personen im autonomen Fahrzeug geopfert werden (Bundesministerium für Verkehr und digitale Infrastruktur, 2017). Jedoch wurde mit einer steigenden Personenanzahl in dem Fahrzeug in den Experimenten 2a und 2b in allen Perspektiven eine zunehmende Tendenz zum Schutz der größeren Gruppe und damit der Personen in dem autonomen Fahrzeug sichtbar. Zusammen mit den Ergebnissen der Experimente 1a und 1b lässt sich somit bei beiden involvierten Parteien die Bereitschaft beobachten, sich für eine größere Gruppe der jeweils anderen Unfallpartei zu opfern.

Einfluss von sozialer Erwünschtheit

Werden in Befragungen Merkmale erfasst, für die soziale Normen existieren, die beschreiben, welche Merkmalsausprägungen von einer Gesellschaft als angemessen oder wünschenswert angesehen werden, neigen manche Befragte – unabhängig von ihrer tatsächlichen Merkmalsausprägung – dazu, eher im Einklang mit diesen Normen zu antworten als Antworten zu geben, die der zugrundeliegenden Norm widersprechen (Sudman & Bradburn, 1974; siehe z. B. Tourangeau & Yan, 2007 für

einen Überblick zu sozial erwünschtem Antwortverhalten). Diese Tendenz zu sozial erwünschten Antworten kann in der Folge zu unzuverlässigen Prävalenzschätzungen der dazugehörigen sensiblen Merkmale führen (z. B. Krumpal, 2013; Tourangeau & Yan, 2007), da manche Befragte ihre tatsächliche Einstellung oder bestimmte Handlungen ungerne angeben, sofern sie damit die Verletzung einer sozialen Norm zugeben müssten.

Auch im Kontext autonomer Fahrzeuge ist ein Einfluss von sozialer Erwünschtheit denkbar: Gerade in dilemmatischen Situationen, in denen das eigene Leben dem Leben anderer Verkehrsteilnehmender gegenübergestellt wird, könnten Personen beispielsweise eine utilitaristische Präferenz zur Selbstopferung entgegen ihrer eigentlichen Überzeugungen angeben, um zu vermeiden, dass sie aufgrund ihrer Bereitschaft, andere für ihr eigenes Überleben zu opfern, selbstsüchtig oder kaltherzig erscheinen. Es gibt beispielsweise Hinweise darauf, dass es als moralisch weniger akzeptabel betrachtet wird, eine andere Person zugunsten einer Gruppe zu opfern als sich selbst zu opfern (Sachdeva et al., 2015). Allerdings beobachteten Sütfeld et al. (2019) einen limitierten Einfluss sozialer Erwünschtheit auf Präferenzen in moralischen Dilemmas. Jedoch wurden in dieser Untersuchung weder mögliche Selbstopferungen noch utilitaristische Präferenzen einbezogen.

Sofern Versuchsteilnehmende die Wahl der utilitaristischen Alternative in moralischen Dilemmas mit autonomen Fahrzeugen als soziale Norm empfinden – worauf die häufige Beobachtung utilitaristischer Präferenzen hindeuten könnte (z. B. Awad et al., 2018; Bonnefon et al., 2016; Kallioinen et al., 2019) – wären sie weniger gewillt ihre Präferenz zum Selbstschutz in Befragungen offen anzugeben. In diesem Fall würden die Ergebnisse der Experimente 1a bis 2b – aber möglicherweise auch anderer Studien – die Präferenz für utilitaristische Handlungen überschätzen, während der Einfluss sozial unerwünschter Selbstschutztendenzen unterschätzt würde. Um mögliche Probleme hinsichtlich der Akzeptanz autonomer Fahrzeuge antizipieren zu können, ist es wichtig, die tatsächliche Einstellung gegenüber verschiedenen Handlungsoptionen in moralischen Dilemmas mit autonomen Fahrzeugen zu erfassen. Inwiefern soziale Erwünschtheit einen Einfluss auf die Bewertung moralischer Dilemmas mit autonomen Fahrzeugen hat, ist jedoch unklar. Daher wurde in Experiment 3 untersucht, ob soziale Erwünschtheit das Antwortverhalten hinsichtlich der präferierten Handlungen autonomer Fahrzeuge in moralischen Dilemmas beeinflusst. Insbesondere war von Interesse, ob die Präferenz für utilitaristische

Handlungen in moralischen Dilemmas, die eine potentielle Selbstopferung enthalten, durch sozial erwünschte Antworttendenzen überschätzt wird.

Um den Einfluss sozialer Erwünschtheit zu untersuchen, wurde in Experiment 3 das *Extended-Crosswise*-Modell (Heck et al., 2018) verwendet. Dabei handelt es sich um eine indirekte Befragungsmethode. Indirekte Befragungsmethoden wurden dafür entwickelt, sozial erwünschten Antworttendenzen in Befragungen entgegenzuwirken. Indirekte Befragungsmethoden, die wie das *Extended-Crosswise*-Modell auf der *Randomized-Response*-Technik (Warner, 1965) basieren, erhöhen die Vertraulichkeit der Antworten der Teilnehmenden, indem den Daten zufälliges Rauschen hinzugefügt wird. Dadurch kann auf individueller Ebene nicht mehr nachvollzogen werden, welche Antwort eine Person in Bezug auf ein zu erfassendes sensibles Merkmal gegeben hat.

Das *Extended-Crosswise*-Modell basiert auf dem *Crosswise*-Modell (Yu et al., 2008), bei dem das zufällige Rauschen dadurch hinzugefügt wird, dass die Teilnehmenden zeitgleich zu zwei Aussagen Stellung nehmen sollen. Eine der Aussagen bezieht sich auf das zu erfassende sensible Merkmal (z. B. »Das autonome Fahrzeug sollte in der dargestellten Situation utilitaristisch handeln.«). Die andere Aussage bezieht sich auf ein nicht-sensibles Merkmal mit bekannter Prävalenz wie beispielsweise den eigenen Geburtsmonat, dessen Prävalenz sich mithilfe offizieller Geburtsstatistiken schätzen lässt (z. B. »Ich bin im November oder Dezember geboren.«). Um zu beiden Aussagen gemeinsam Stellung zu nehmen, geben die Teilnehmenden entweder an, beiden Aussagen zuzustimmen bzw. nicht zuzustimmen (z. B. »Ich stimme beiden Aussagen oder keiner der beiden Aussagen zu.«) oder nur einer Aussage zuzustimmen (z. B. »Ich stimme nur einer Aussage (egal welcher) zu.«). Während bei einer direkten Befragung (z. B. »Sollte das autonome Fahrzeug in der dargestellten Situation utilitaristisch handeln?« mit den Antwortoptionen »Ja« und »Nein«) aus der gegebenen Antwort ein Rückschluss auf die Einstellung der antwortenden Person gezogen werden kann, ist ein entsprechender Rückschluss bei der Verschlüsselung durch das *Crosswise*-Modell nicht möglich: Durch die Antwortoptionen ist nicht ersichtlich, ob eine einzelne Person der sensiblen Aussage zugestimmt hat. Mithilfe der bekannten Prävalenz des nicht-sensiblen Merkmals kann jedoch auf Gruppenebene geschätzt werden, wie viele Personen der sensiblen Aussage zugestimmt haben. Durch die im Vergleich zu einer direkten Befragung erhöhte Vertraulichkeit der Antworten kann davon ausgegangen werden, dass die Befragten eher

bereit sind, ehrliche Angaben zu sensiblen Themen zu machen als bei einer direkten Befragung, wodurch die Prävalenz eines sensiblen Merkmals valider geschätzt werden kann (Lensvelt-Mulders et al., 2005).

Das in Experiment 3 verwendete *Extended-Crosswise*-Modell (Heck et al., 2018) hat gegenüber dem *Crosswise*-Modell den Vorteil, dass sich mithilfe dieses Modells zusätzlich prüfen lässt, ob sich die Versuchsteilnehmenden an die Instruktionen der Befragung gehalten haben. Dies wird durch zwei indirekte Befragungsgruppen ermöglicht. Wie bei dem *Crosswise*-Modell werden den Befragten in beiden indirekten Befragungsgruppen eine Aussage zu dem zu erfassenden sensiblen Merkmal und eine Aussage zu einem nicht-sensiblen Merkmal mit bekannter Prävalenz präsentiert, zu denen die Teilnehmenden zeitgleich Stellung nehmen sollen. Die beiden indirekten Befragungsbedingungen unterscheiden sich lediglich in den nicht-sensiblen Aussagen, die komplementär zueinander formuliert sind. Erhält die eine Gruppe beispielsweise die nicht-sensible Aussage »Ich bin im November oder Dezember geboren.«, erhält die andere Gruppe dementsprechend die Aussage »Ich bin zwischen Januar und Oktober geboren.«. Da die Versuchsteilnehmenden randomisiert auf die beiden indirekten Befragungsgruppen aufgeteilt werden, kann angenommen werden, dass sich weder die Prävalenz der Geburtsmonate noch die Prävalenz des sensiblen Merkmals zwischen den beiden Befragungsgruppen unterscheidet. Die ermittelten Prävalenzschätzungen des sensiblen Merkmals sollten sich dementsprechend ebenfalls nicht zwischen den Befragungsgruppen unterscheiden, sofern sich die Versuchsteilnehmenden an die Instruktionen gehalten haben. Signifikante Unterschiede in den Prävalenzschätzungen zwischen den beiden indirekten Befragungsgruppen weisen dementsprechend darauf hin, dass die Daten nicht vertrauenswürdig sind und nicht interpretiert werden sollten (Heck et al., 2018). Das *Extended-Crosswise*-Modell konnte beispielsweise bereits den Einfluss sozialer Erwünschtheit im Kontext von Islamophobie (Meisters et al., 2020) und bei Angaben zu der eigenen Handhygiene während der Anfangsphase der Covid-19-Pandemie (Mieth et al., 2021) zeigen.

Experiment 3

Um zu überprüfen, ob die utilitaristische Präferenz, sich selbst zu opfern, um das Leben anderer zu retten, in moralischen Dilemmas mit autonomen Fahrzeugen von sozialer Erwünschtheit beeinflusst wird, wurden die Teilnehmenden in Experiment 3 randomisiert auf drei Befragungsbedingungen aufgeteilt: Gemäß dem

Extended-Crosswise-Modell (Heck et al., 2018) gab es zwei indirekte Befragungsbedingungen sowie eine direkte Befragungsbedingung als Vergleichsgruppe. In allen Befragungsbedingungen wurde das gleiche moralische Dilemma präsentiert, in dem zwischen der Opferung von fünf Personen in einem autonomen Fahrzeug und der Opferung einer einzelnen Person auf der Straße abgewogen werden musste. Alle Teilnehmenden wurden gebeten, sich in die Person auf der Straße hineinzuversetzen und sollten zu der sensiblen Aussage »In der dargestellten Situation sollte das autonome Fahrzeug mich als Passantin oder Passanten schützen und die fünf Fahrzeuginsassen opfern.« Stellung nehmen. In den indirekten Befragungsbedingungen erfolgte die Stellungnahme zu der sensiblen Aussage zusammen mit einer nicht-sensiblen Aussage (Gruppe 1: »Ich bin im November oder Dezember geboren.«, Gruppe 2: »Ich bin zwischen Januar und Oktober geboren.«).

Mithilfe des in Abbildung 3 abgebildeten multinomialen Verarbeitungsbaummodells wurde für die jeweiligen Befragungsbedingungen die Prävalenz π der Personen geschätzt, die den Schutz der eigenen Person der utilitaristischen Handlung, die fünf Personen im Fahrzeug zu schützen, vorziehen. In der direkten Befragungsbedingung entspricht die Zustimmung zu der sensiblen Aussage einer Präferenz zum Selbstschutz, während eine Ablehnung der sensiblen Aussage einer utilitaristischen Präferenz entspricht (vgl. oberer Baum in Abbildung 3). Die Prävalenzen der beiden indirekten Befragungsbedingungen (π_{IB1} und π_{IB2}) werden hingegen mithilfe der bekannten Prävalenz des Merkmals im November oder Dezember geboren zu sein (Parameter p_{NovDez}) geschätzt (siehe Mayer et al., 2021 für weitere Details des Vorgehens). Wenn soziale Erwünschtheit utilitaristische Präferenzen beeinflusst, sollten in den indirekten Befragungsgruppen weniger Teilnehmende eine (utilitaristische) Präferenz zur Selbstopferung angeben und somit insgesamt mehr Zustimmung für die sensible Aussage zeigen, dass das autonome Fahrzeug sie selbst schützen sollte, als Teilnehmende in der direkten Befragungsbedingung.

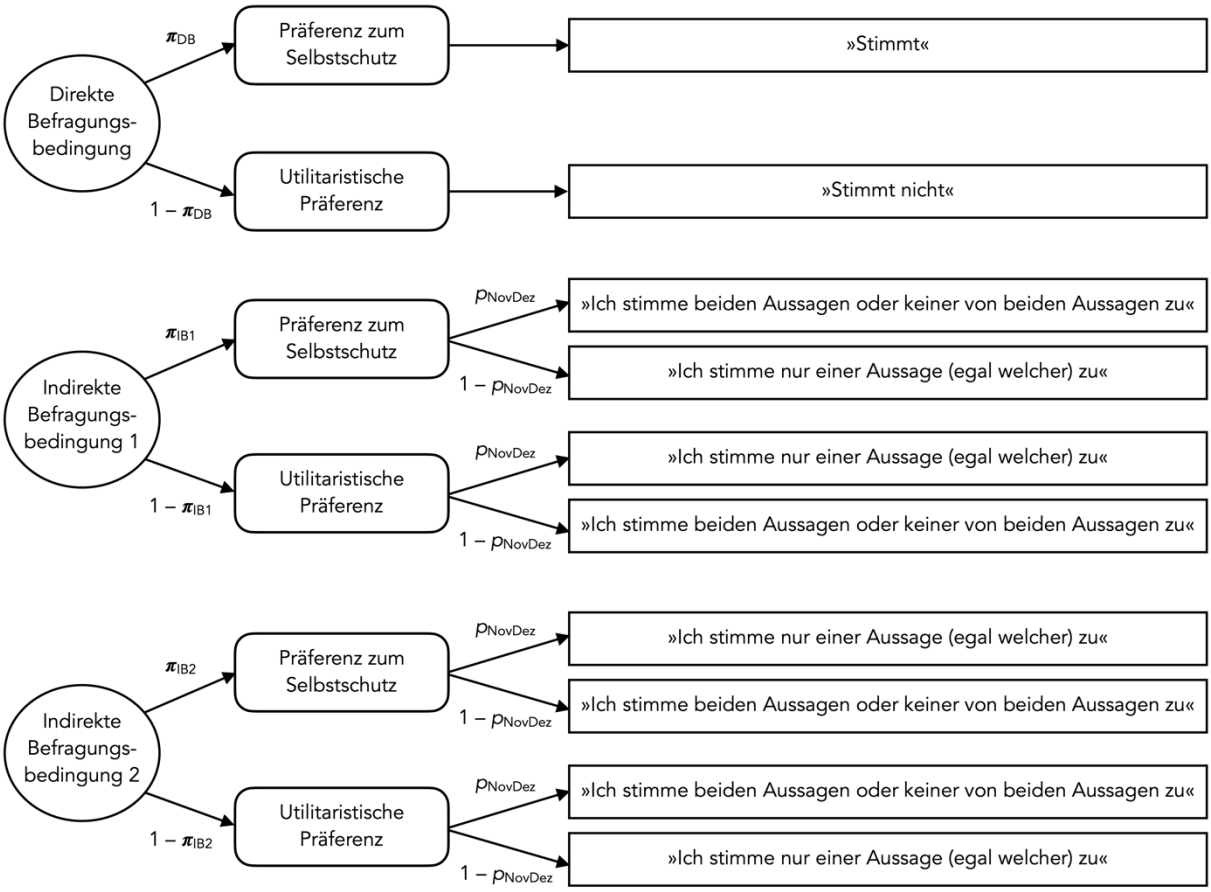


Abbildung 3. Das kombinierte multinomiale Verarbeitungsbaummodell für die direkte Befragungsbedingung (oberer Baum) und die beiden indirekten Befragungsbedingungen (mittlerer und unterer Baum) des *Extended-Crosswise*-Modells in Experiment 3. In den rechts abgebildeten Rechtecken werden die Antwortkategorien der jeweiligen Befragungsbedingung dargestellt. Die Parameter π repräsentieren die Prävalenzschätzungen für die selbstschützende Präferenz, dass das autonome Fahrzeug fünf Personen im Fahrzeuginneren zugunsten der Person auf der Straße opfert, in den jeweiligen Befragungsbedingungen. Der Parameter p_{NovDez} repräsentiert die bekannte Prävalenz des nicht-sensiblen Merkmals, im November oder Dezember geboren zu sein.

Da sich die Prävalenzschätzungen für das sensible Merkmal – also die Bereitschaft, fünf Personen im Fahrzeug zum eigenen Vorteil zu opfern – nicht zwischen den beiden indirekten Befragungsbedingungen unterschied [$G^2(1) = 1.37$, $p = .243$, $w = .03$], wurden die beiden indirekten Befragungsbedingungen zusammengefasst und mit der direkten Befragungsbedingung verglichen. Dabei zeigte sich kein signifikanter Unterschied zwischen der Prävalenzschätzung des sensiblen Merkmals in der direkten Befragungsbedingung [41.0 %, $SE = 2.3$] und der Prävalenzschätzung der kombinierten indirekten Befragungsbedingung [40.1 %, $SE = 2.4$; $\Delta G^2(1) = 0.07$,

$p = .790$, $w = .01$]. Die Daten widersprechen somit der Annahme, dass soziale Erwünschtheit die utilitaristische Präferenz, sich selbst zugunsten einer größeren Gruppe zu opfern, in dem verwendeten moralischen Dilemma mit einem autonomen Fahrzeug beeinflusst. Dies steht im Einklang mit den Befunden von Sütfeld et al. (2019), die ebenfalls Evidenz für den eingeschränkten Einfluss sozialer Erwünschtheit auf das Antwortverhalten in moralischen Dilemmas beobachteten. Des Weiteren stützen die Ergebnisse die Validität der oft in moralischen Dilemmas im Kontext autonomer Fahrzeuge beobachteten utilitaristischen Präferenz (z. B. Awad et al., 2018; Bergmann et al., 2018; Bonnefon et al., 2016; Faulhaber et al., 2019), die sich ebenfalls in Experiment 3 beobachten ließ: Unabhängig von der Befragungsbedingung und im Einklang mit den Ergebnissen der Experimente 1a bis 2b, zeigte die Mehrheit der Teilnehmenden (ca. 60 %) eine utilitaristische Präferenz zur Selbstopferung, um fünf andere Personen zu retten. Nachdem im Fokus der Experimente 1a bis 3 die präferierten Handlungen autonomer Fahrzeuge standen, konzentrierten sich die Experimente 4 und 5 auf die moralische Bewertung von Handlungen in unvermeidlichen Unfallsituationen.

Einfluss der Handelnden

Dass autonome Fahrzeuge unabhängig von ihrer Leistungsfähigkeit nicht alle Unfälle vermeiden können, weil sie mit Verkehrsteilnehmenden, deren Verhalten schwer vorherzusagen ist, interagieren (z. B. Koopman & Wagner, 2017; Lin, 2016; Nyholm, 2018), wird nicht zuletzt durch reale Unfälle mit Fahrzeugen, die mit automatisierten Fahrsystemen ausgestattet waren, verdeutlicht. Der Unfall mit einem Tesla im Jahr 2016 (National Transportation Safety Board, 2017), bei dem der Fahrer des Teslas ums Leben kam, und der Unfall mit einem Volvo der Firma Uber im Jahr 2018 (National Transportation Safety Board, 2019), bei dem eine Passantin getötet wurde, gehören vermutlich zu den bekanntesten Vorfällen. Solche Unfälle können, vor allem in der Anfangsphase der Einführung autonomer Fahrzeuge und nicht zuletzt durch die Neuheit autonomer Fahrsysteme, eine erhöhte Medienaufmerksamkeit auf sich ziehen (z. B. Jelinski et al., 2021; Shariff et al., 2017). Eine (besonders) kritische Berichterstattung über tödliche Unfälle mit autonomen Fahrzeugen könnte die öffentliche Akzeptanz autonomer Fahrzeuge negativ beeinflussen (Anania et al., 2018; Shariff et al., 2017) und dadurch für die Einführung autonomer Fahrzeuge

problematisch sein, zumal die öffentliche Meinung zu autonomen Fahrzeugen bisher gemischt zu sein scheint (Becker & Axhausen, 2017). Hinsichtlich der Skepsis gegenüber autonomen Fahrzeugen geben Personen oft an, die Kontrolle über die Fahraufgabe nicht vollständig oder endgültig an das Fahrzeug abgeben zu wollen bzw. einen Kontrollverlust zu befürchten (Smith & Anderson, 2017; Winkler et al., 2019; Wolf, 2016). Dass autonome Fahrsysteme auch Entscheidungen treffen könnten, in deren Folge Menschen verletzt oder sogar getötet werden, könnte zu der Abneigung beitragen, die Kontrolle an ein autonomes Fahrzeug abzugeben (Bigman & Gray, 2018; Li et al., 2016; Malle et al., 2016).

Vor diesem Hintergrund ist die Wahrnehmung von Unfällen mit autonomen Fahrzeugen und die Bewertung von Handlungen in Unfallsituationen besonders relevant. Zu untersuchen, wie Menschen Handlungen autonomer Fahrzeuge im Vergleich zu Handlungen menschlicher Fahrender in tödlichen Unfallszenarien moralisch bewerten, könnte dabei helfen, potentielle Probleme hinsichtlich der Akzeptanz autonomer Fahrzeuge zu antizipieren. Was als moralisch vertretbares Handeln betrachtet wird, könnte durchaus davon abhängen, ob autonome Fahrzeuge oder menschliche Fahrende handeln. Zwar weisen einige Studien darauf hin, dass von autonomen Fahrzeugen und menschlichen Fahrenden ähnliche Handlungen gewünscht sind (z. B. Bonnefon et al., 2016; Kallioinen et al., 2019; Li et al., 2016; Young & Monroe, 2019), jedoch könnte sich die moralische Bewertung einer bereits durchgeführten Handlung mit feststehenden Konsequenzen zwischen verschiedenen Handelnden unterscheiden. In mehreren Studien untersuchten Malle und Kollegen beispielsweise, wie die Handlungen von verschiedenen Maschinen wie Robotern oder Drohnen verglichen mit Handlungen von Menschen in verschiedenen dilemmatischen Situationen bewertet wurden (Malle et al., 2019; Malle et al., 2015; Malle et al., 2016). Die Ergebnisse deuten teilweise darauf hin, dass es Unterschiede in der Bewertung von moralisch aufgeladenen Handlungen von Menschen und Maschinen geben könnte. Beispielsweise finden sich in der Studie von Malle et al. (2015) Hinweise darauf, dass von Robotern utilitaristische Entscheidungen erwartet werden und utilitaristische Handlungen für Roboter zulässiger sein könnten als für Menschen. Des Weiteren gibt es Evidenz für eine Präferenz von menschlichen anstelle von maschinellen Entscheidenden bei Entscheidungen über Leben und Tod (Bigman & Gray, 2018) sowie für eine Abneigung, moralische Aufgaben an Maschinen zu übergeben (Gogoll & Uhl, 2018). Die mögliche Aversion gegenüber Maschinen, die moralische

Entscheidungen treffen, könnte zu einer kritischeren Bewertung der Handlungen autonomer Fahrzeuge verglichen mit der Bewertung der Handlungen menschlicher Fahrender führen. Darauf könnte auch die von Young und Monroe (2019) beobachtete Tendenz hinweisen, autonomen Fahrzeugen mehr Schuld für geschehene Handlungen in Unfallsituationen zuzuschreiben als menschlichen Fahrenden.

In den Experimenten 4 und 5 der vorliegenden Arbeit wurde daher untersucht, inwiefern die Handlungen autonomer Fahrzeuge und menschlicher Fahrender moralisch unterschiedlich bewertet werden. Dafür wurde die moralische Bewertung der Handlungen von autonomen Fahrzeugen und menschlichen Fahrenden in verschiedenen tödlichen Unfallszenarien gegenübergestellt. Insbesondere waren dabei zwei Aspekte von Interesse: die zugrundeliegenden moralischen Prinzipien sowie eine mögliche Tendenz, die Handlungen von Menschen positiver zu bewerten als die Handlungen autonomer Fahrzeuge. Ob die gleichen moralischen Prinzipien herangezogen werden, um die Handlungen von Menschen und autonomen Fahrzeugen zu bewerten oder ob verschiedene Prinzipien für unterschiedliche Handelnde relevant sind, wurde am Beispiel utilitaristischer Handlungen untersucht. Die Ergebnisse von Malle et al. (2015) könnten darauf hindeuten, dass utilitaristische Handlungen autonomer Fahrzeuge im Vergleich zu den gleichen Handlungen menschlicher Fahrender wohlwollender bewertet werden. Um dies in den Experimenten 4 und 5 überprüfen zu können wurde die Anzahl der Personen auf der Straße in den präsentierten Unfallszenarien variiert. Des Weiteren legen die Ergebnisse von Bigman und Gray (2018) und Gogoll und Uhl (2018) eine Abneigung gegenüber Maschinen nahe, die moralische Entscheidungen treffen. Dies könnte zu einer kritischeren Bewertung der Handlungen autonomer Fahrzeuge verglichen mit den Handlungen menschlicher Fahrender führen. Dahingehend wurde in Experiment 5 zusätzlich untersucht, inwiefern die Vermenschlichung eines autonomen Fahrzeugs eine Möglichkeit darstellen könnte, Unterschieden in der moralischen Bewertung von Handlungen autonomer Fahrzeuge und menschlicher Fahrender entgegenzuwirken.

Experiment 4

Um zu untersuchen, ob die Handlungen autonomer Fahrzeuge und menschlicher Fahrender in dilemmatischen Unfallsituationen moralisch unterschiedlich bewertet werden, wurden den Teilnehmenden verschiedene abstrakt dargestellte Unfallszenarien präsentiert. In allen Szenarien war der Unfall bereits geschehen. Somit

bestand keine Unsicherheit hinsichtlich der Handlungskonsequenzen. Die Szenarien unterschieden sich in der Anzahl der Personen auf der Straße (eine Person, zwei oder fünf Personen) und der Handlung, die von den Handelnden vorgenommen wurde (Person im Fahrzeug opfern oder Person/en auf der Straße opfern). Die Teilnehmenden sollten entweder die Handlungen eines autonomen Fahrzeugs oder eines Menschen aus einer moralischen Perspektive auf einer sechsstufigen Skala von »sehr verwerflich« (1) bis »sehr vertretbar« (6) bewerten. Die mittlere moralische Bewertung der jeweiligen Handlungen wurde zwischen den Handelnden und den verschiedenen Unfallszenarien verglichen. Wenn Menschen eine Aversion gegenüber Maschinen haben, die moralische Entscheidungen treffen (Bigman & Gray, 2018; Gogoll & Uhl, 2018), sollten die Handlungen eines autonomen Fahrzeugs als moralisch verwerflicher bewertet werden als die korrespondierenden Handlungen eines Menschen. Des Weiteren sollte die Bewertung der Handlungen des autonomen Fahrzeugs stärker von der Anzahl der Personen auf der Straße abhängen als die Bewertung der Handlungen eines Menschen, sofern utilitaristische Handlungen bei autonomen Fahrzeugen positiver bewertet werden als bei Menschen (Malle et al., 2015).

Zunächst kann festgehalten werden, dass die Handlung, die Person im Fahrzeug zu opfern – unabhängig davon, ob ein Mensch oder ein autonomes Fahrzeug handelte – als moralisch vertretbarer bewertet wurde als die Handlung, die Person/en auf der Straße zu opfern [$F(1,358) = 340.82$, $p < .001$, $\eta_p^2 = .49$]. Die deskriptiven Statistiken der moralischen Bewertung sind in Abbildung 4 dargestellt. In der Mehrheit der präsentierten Szenarien stellt die Handlung, die Person im Fahrzeug zu opfern, die utilitaristische Handlung dar, sodass dieses Datenmuster auf eine Präferenz für utilitaristische Handlungen hindeutet. Tatsächlich beeinflusste die Anzahl der Personen auf der Straße die moralische Bewertung der Handlungen [$F(2,357) = 23.94$, $p < .001$, $\eta_p^2 = .12$]. Die Person im Fahrzeug zu opfern wurde als moralisch vertretbarer bewertet, je mehr Personen sich auf der Straße befanden, während die Handlung, die Person/en auf der Straße zu opfern, als moralisch verwerflicher bewertet wurde, je mehr Personen sich auf der Straße befanden [$F(2,357) = 187.20$, $p < .001$, $\eta_p^2 = .51$]. Jeder Anstieg der Personenanzahl führte für die Handlung, die Person im Fahrzeug zu opfern zu einer signifikanten Zunahme und für die Handlung, die Person/en auf der Straße zu opfern, zu einer signifikanten Abnahme der moralischen Bewertung [einfache Haupteffektvergleiche, alle $p < .001$].

Die Auswirkung der Personenanzahl auf der Straße war für autonome Fahrzeuge und menschliche Fahrende gleich [$F(2,357) = 2.73$, $p = .067$, $\eta_p^2 = .02$].

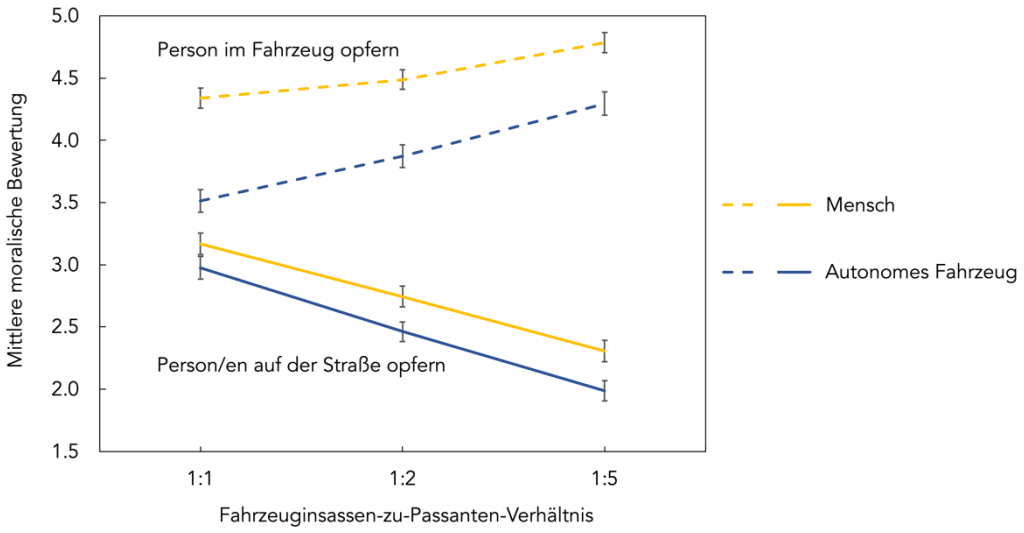


Abbildung 4. Die mittlere moralische Bewertung der Handlung, die Person im Fahrzeug zu opfern (gestrichelte Linien), und der Handlung, die Person/en auf der Straße zu opfern (durchgezogene Linien), als Funktion des Fahrzeuginsassen-zu-Passanten-Verhältnisses (1:1, 1:2, 1:5) und der Handelnden (Mensch, autonomes Fahrzeug). Die Bewertungsskala reichte von »sehr verwerflich« (1) bis »sehr vertretbar« (6). Die Fehlerbalken repräsentieren die Standardfehler der Mittelwerte.

Nichtsdestotrotz wurden die Handlungen des Menschen als moralisch vertretbarer bewertet als die Handlungen des autonomen Fahrzeugs [$F(1,358) = 40.51$, $p < .001$, $\eta_p^2 = .10$]. Dies stützt die Hypothese, dass Menschen eine Aversion gegen Maschinen haben, die moralische Entscheidungen treffen (Bigman & Gray, 2018; Gogoll & Uhl, 2018). Auch die Interaktion zwischen dem Faktor der Handelnden und der gewählten Handlung war signifikant [$F(1,358) = 4.72$, $p = .030$, $\eta_p^2 = .01$]. Der Bewertungsunterschied zwischen den Handlungen des autonomen Fahrzeugs und den Handlungen des Menschen war für die Handlung, die Person im Fahrzeug zu opfern, ausgeprägter [einfache Haupteffektvergleiche: $\eta_p^2 = .08$, $p < .001$] als für die Handlung, die Person/en auf der Straße zu opfern [einfache Haupteffektvergleiche: $\eta_p^2 = .02$, $p = .016$].

Des Weiteren zeigte sich eine signifikante Dreifach-Interaktion zwischen den Handelnden, der gewählten Handlung und der Personenanzahl auf der Straße

[$F(2,357) = 4.54$, $p = .011$, $\eta_p^2 = .02$]. Auch für beide Handelnde getrennt betrachtet ließ sich sowohl der Effekt der gewählten Handlung [Mensch: $F(1,186) = 241.01$, $p < .001$, $\eta_p^2 = .56$; autonomes Fahrzeug: $F(1,172) = 117.40$, $p < .001$, $\eta_p^2 = .41$] als auch der Effekt der Personenanzahl auf der Straße [Mensch: $F(2,185) = 18.70$, $p < .001$, $\eta_p^2 = .17$; autonomes Fahrzeug: $F(2,171) = 6.79$, $p = .001$, $\eta_p^2 = .07$] replizieren. Die Interaktion der gewählten Handlung und der Personenanzahl auf der Straße wurde ebenso für beide Handelnden signifikant [Mensch: $F(2,185) = 90.23$, $p < .001$, $\eta_p^2 = .49$; autonomes Fahrzeug: $F(2,171) = 96.98$, $p < .001$, $\eta_p^2 = .53$]. Zusammen mit der signifikanten Interaktion zwischen den Handelnden und der gewählten Handlung, könnte die signifikante Dreifach-Interaktion nahelegen, dass die Bewertung der Handlungen des Menschen weniger von der Anzahl der Personen auf der Straße beeinflusst wurde als die Bewertung der Handlungen des autonomen Fahrzeugs. Auch wenn dies die Hypothese zu stützen scheint, dass utilitaristische Handlungen bei autonomen Fahrzeugen positiver bewertet werden als bei menschlichen Fahrenden (Malle et al., 2015), kann dieses Datenmuster ebenfalls damit erklärt werden, dass die Entscheidung des Menschen, sich selbst zugunsten anderer zu opfern, bereits bei nur einer Person auf der Straße positiv bewertet wurde und diese positive Einschätzung durch weitere Personen auf der Straße kaum zu steigern war. Zusätzlich weisen die beiden Interaktionen eine geringe Effektstärke auf [Interaktion der Handelnden und der gewählten Handlung: $\eta_p^2 = .01$; Dreifach-Interaktion: $\eta_p^2 = .02$]. Somit ist fraglich, ob sich diese Ergebnisse replizieren lassen. Dies wurde in Experiment 5 getestet. Des Weitern wurde in Experiment 5 untersucht, ob eine Vermenschlichung des autonomen Fahrzeugs eine Möglichkeit darstellen könnte, um die Bewertung der Handlungen autonomer Fahrzeuge und menschlicher Fahrender einander anzunähern.

Experiment 5

Die Suche nach möglichen Maßnahmen, um die in Experiment 4 beobachtete negativere Bewertung der Handlungen autonomer Fahrzeuge im Vergleich zu denen menschlicher Fahrender zu reduzieren, ist vor dem Hintergrund der für eine Einführung autonomer Fahrzeuge erforderlichen öffentlichen Akzeptanz besonders interessant. Einen möglichen Ansatzpunkt hierfür könnte eine Vermenschlichung autonomer Fahrzeuge darstellen. Die Vermenschlichung unbelebter Objekte oder Entitäten durch die Zuschreibung menschlicher Charakteristika und Eigenschaften wird als Anthropomorphismus bezeichnet (Bartneck et al., 2009; Epley et al., 2007). Es gibt

Hinweise darauf, dass die Anthropomorphisierung technischer Systeme deren Wahrnehmung positiv beeinflussen kann und beispielsweise zu erhöhter Vertrauenswürdigkeit oder zu mehr Vertrauen in anthropomorphisierte technische Systeme verglichen mit nicht-anthropomorphisierten Systemen führt (z. B. Gong, 2008; Lee et al., 2015; Niu et al., 2018; Pak et al., 2012). Waytz et al. (2014) konnten außerdem zeigen, dass einem autonomen Fahrzeug mit einem menschlichen Namen, einer Stimme und einem Geschlecht mehr vertraut wurde und es bei einem fremdverschuldeten Unfall weniger Schuld zugewiesen bekam als ein autonomes Fahrzeug ohne anthropomorphe Eigenschaften. Daher scheint es plausibel, dass Anthropomorphismus auch die moralische Bewertung der Handlungen von autonomen Fahrzeugen beeinflussen könnte. Hinweise hierauf finden sich auch in der Studie von Young und Monroe (2019): Wurde der Entscheidungsprozess eines autonomen Fahrzeugs ähnlich zu dem eines Menschen beschrieben (z. B. Gedanken, Gefühle), verringerte dies tendenziell die Schuld, die einem autonomen Fahrzeug für Handlungen zugeschrieben wurde, verglichen mit einem autonomen Fahrzeug mit einem eher mechanisch beschriebenen Entscheidungsprozess. Des Weiteren beobachteten Malle et al. (2016) einen Einfluss der äußeren Erscheinung von Robotern auf die Schuldzuweisung für Handlungen in moralischen Dilemmas. Der Unterschied in der Schuldzuweisung war für einen humanoiden Roboter im Vergleich zu einem Menschen geringer als für einen mechanisch dargestellten Roboter verglichen mit einem Menschen.

Um zu überprüfen, inwieweit Anthropomorphismus eine Möglichkeit darstellen könnte, die moralische Bewertung von Handlungen autonomer Fahrzeuge und menschlicher Fahrender anzunähern, wurde in Experiment 5 zusätzlich zu einem autonomen Fahrzeug und einem Menschen ein anthropomorphisiertes autonomes Fahrzeug einbezogen. Das Fahrzeug wurde durch die Zuschreibung eines Vornamens anthropomorphisiert (z. B. Hong et al., 2020; Waytz et al., 2014). Abgesehen von der zusätzlichen Experimentalbedingung wurde das methodische Vorgehen sowie das Material aus Experiment 4 übernommen, um zu überprüfen, ob sich die zentralen Ergebnisse von Experiment 4 replizieren lassen. Parallel zu der Datenauswertung in Experiment 4 wurde die mittlere moralische Bewertung der jeweiligen Handlungen für den Menschen, das anthropomorphisierte autonome Fahrzeug und das autonome Fahrzeug und für die verschiedenen Unfallsituationen miteinander verglichen. Die deskriptiven Statistiken der moralischen Bewertung sind in Abbildung 5 dargestellt.

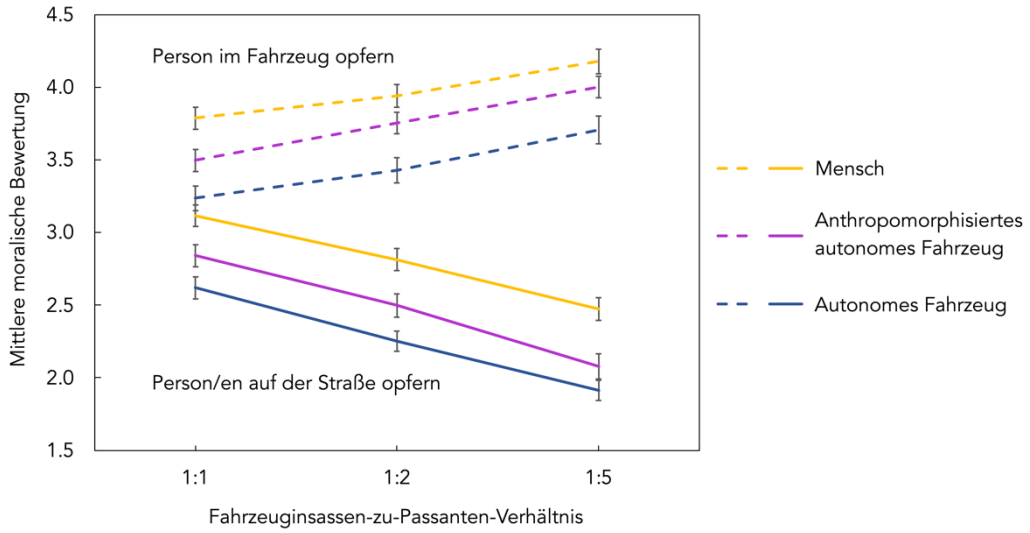


Abbildung 5. Die mittlere moralische Bewertung der Handlung, die Person im Fahrzeug zu opfern (gestrichelte Linien), und der Handlung, die Person/en auf der Straße zu opfern (durchgezogene Linien), als Funktion des Fahrzeuginsassen-zu-Passanten-Verhältnisses (1:1, 1:2, 1:5) und der Handelnden (Mensch, anthropomorphisiertes autonomes Fahrzeug, autonomes Fahrzeug). Die Bewertungsskala reichte von »sehr verwerflich» (1) bis »sehr vertretbar« (6). Die Fehlerbalken repräsentieren die Standardfehler der Mittelwerte.

Ob ein Mensch, ein anthropomorphisiertes autonomes Fahrzeug oder ein autonomes Fahrzeug handelte, beeinflusste – im Einklang mit den Ergebnissen des Experiments 4 – die moralische Bewertung der Handlung [$F(2,752) = 24.72$, $p < .001$, $\eta_p^2 = .06$]. Auch in Experiment 5 wurden die Handlungen des Menschen als moralisch vertretbarer bewertet als die Handlungen beider autonomen Fahrzeuge [$F(1,752) = 37.76$, $p < .001$, $\eta_p^2 = .05$]. Dies stützt erneut die Hypothese, dass Menschen eine Abneigung gegen Maschinen haben, die moralische Entscheidungen treffen (Bigman & Gray, 2018; Gogoll & Uhl, 2018). Allerdings wurden die Handlungen des anthropomorphisierten autonomen Fahrzeugs als moralisch vertretbarer bewertet als die Handlungen des nicht-anthropomorphisierten autonomen Fahrzeugs [$F(1,752) = 11.35$, $p = .001$, $\eta_p^2 = .01$]. Eine Anthropomorphisierung des autonomen Fahrzeugs hob die Unterschiede in den Handlungsbewertungen zwischen autonomen Fahrzeugen und menschlichen Fahrenden nicht vollständig auf, trug jedoch dazu bei, die Bewertungsunterschiede zu verringern.

Des Weiteren wurde erneut die Handlung, die Person im Fahrzeug zu opfern, als moralisch vertretbarer bewertet als die Handlung, die Person/en auf der Straße zu opfern, unabhängig davon, ob ein Mensch, ein anthropomorphisiertes autonomes

Fahrzeug oder ein autonomes Fahrzeug handelte [$F(1,752) = 399.57$, $p < .001$, $\eta_p^2 = .35$]. Der Effekt der Anzahl der Personen auf der Straße konnte ebenfalls repliziert werden [$F(2,751) = 28.18$, $p < .001$, $\eta_p^2 = .07$]. Die Richtung dieses Effekts war erneut abhängig von der gewählten Handlung [$F(2,751) = 219.12$, $p < .001$, $\eta_p^2 = .37$]: Je mehr Personen sich auf der Straße befanden, desto vertretbarer wurde die Handlung, die Person im Fahrzeug zu opfern, bewertet und desto verwerflicher wurde die Handlung, die Person/en auf der Straße zu opfern, eingeschätzt [einfache Haupteffektvergleiche, alle $p < .001$]. Die Auswirkung der Personenanzahl auf der Straße unterschied sich nicht zwischen den Handelnden [$F(4,1502) = 1.10$, $p = .353$, $\eta_p^2 < .01$]. Auch in Experiment 5 wurden demnach Handlungen positiver bewertet, die im Einklang mit dem utilitaristischen Prinzip zum Schutz der größtmöglichen Personenanzahl stehen. Im Gegensatz zu Experiment 4 zeigte sich jedoch weder eine Interaktion zwischen den Handelnden und der gewählten Handlung [$F(2,752) = 0.30$, $p = .742$, $\eta_p^2 < .01$], noch eine Dreifach-Interaktion zwischen den Handelnden, der gewählten Handlung und der Personenanzahl auf der Straße [$F(4,1502) = 0.86$, $p = .485$, $\eta_p^2 < .01$]. Dies spricht gegen die Hypothese, dass utilitaristische Handlungen eher von autonomen Fahrzeugen als von menschlichen Fahrenden erwartet werden.

Diskussion der Experimente 4 und 5

In beiden Experimenten zeigte sich eine deutliche Tendenz, die Handlungen autonomer Fahrzeuge in tödlichen Unfallszenarien moralisch kritischer zu bewerten als die gleichen Handlungen menschlicher Fahrender. Dies korrespondiert mit der Beobachtung von Young und Monroe (2019), dass autonomen Fahrzeugen mehr Schuld für ihre Handlungen in Unfallsituationen zugewiesen wurde als menschlichen Fahrenden. Des Weiteren stützen die Ergebnisse die basierend auf den Studien von Bigman und Gray (2018) und Gogoll und Uhl (2018) formulierte Hypothese, dass Menschen eine Aversion gegenüber Maschinen haben, die moralische Entscheidungen treffen.

Eine Möglichkeit, der kritischeren Bewertung der Handlungen autonomer Fahrzeuge verglichen mit der Bewertung der Handlungen menschlicher Fahrender entgegenzuwirken, könnte basierend auf den Ergebnissen von Experiment 5 eine Anthropomorphisierung autonomer Fahrzeuge darstellen. Dies steht im Einklang mit anderen Studien, die eine positive Wirkung von Anthropomorphismus auf die Wahrnehmung technischer Systeme zeigen konnten (z. B. Gong, 2008; Lee et al.,

2015; Niu et al., 2018; Pak et al., 2012). Die in Experiment 5 gewählte Anthropomorphisierung durch die Zuschreibung eines Namens (z. B. Hong et al., 2020; Waytz et al., 2014) konnte die Unterschiede in der Bewertung der Handlungen autonomer Fahrzeuge und menschlicher Fahrender zwar nicht vollständig eliminieren, allerdings scheint es plausibel, dass zusätzliche Formen der Anthropmorphisierung die positive Tendenz verstärken könnten. Bei der Auswahl anthropomorpher Eigenschaften empfiehlt es sich jedoch, darauf zu achten, dass die Anthropomorphisierung der Maschine bei Nutzenden keine Erwartungen an die Leistungsfähigkeit der Maschine erzeugt, die von der Maschine nicht erfüllt werden können. Die daraus resultierende Erwartungsverletzung könnte die Akzeptanz der Maschine beeinträchtigen (Malle et al., 2016).

Unabhängig von den beobachteten Bewertungsunterschieden zwischen den Handlungen autonomer Fahrzeuge und menschlicher Fahrender, legen die Ergebnisse beider Experimente nahe, dass utilitaristische Prinzipien eine Rolle bei der Handlungsbewertung sowohl von autonomen Fahrzeugen als auch von menschlichen Fahrenden spielen. Unabhängig davon, ob ein Mensch oder ein autonomes Fahrzeug handelte, wurden Handlungen als zunehmend moralisch vertretbarer bewertet, je mehr Menschenleben gerettet wurden und als zunehmend moralisch verwerflicher betrachtet, je mehr Personen durch die Handlung ums Leben kamen. Dies steht im Einklang mit den oft beobachteten utilitaristischen Präferenzen sowohl für autonome Fahrzeuge als auch für menschliche Fahrende in kritischen Unfallsituationen (z. B. Awad et al., 2018; Bergmann et al., 2018; Faulhaber et al., 2019; Kallioinen et al., 2019; Li et al., 2016). Allerdings finden sich in den Experimenten 4 und 5 keine klaren Hinweise darauf, dass die Anzahl der involvierten Personen die moralische Handlungsbewertung bei autonomen Fahrzeugen stärker beeinflusst als bei menschlichen Fahrenden. Sowohl die Interaktion der Handelnden und der gewählten Handlung als auch die Dreifach-Interaktion der Handelnden, der gewählten Handlung und der Personenanzahl auf der Straße, deren statistische Signifikanz auf qualitative Unterschiede in der Bewertung der Handlungen von autonomen Fahrzeugen und menschlichen Fahrenden hinweisen könnte, waren in Experiment 4 mit geringen Effektstärken assoziiert und konnten in Experiment 5 nicht repliziert werden. Zusammengenommen legen die Ergebnisse beider Experimente daher nahe, dass die Interaktionen zu vernachlässigen sind.

Allgemeine Diskussion

In der vorliegenden Arbeit wurden verschiedene Einflussfaktoren auf die Handlungspräferenzen für autonome Fahrzeuge und auf die moralische Handlungsbewertung in unvermeidlichen Unfallsituationen untersucht. Mit der Einführung autonomer Fahrzeuge in den Straßenverkehr werden viele Hoffnungen verbunden (für einen Überblick siehe z. B. Anderson et al., 2016; Bagloee et al., 2016) – insbesondere Hoffnungen auf eine erhöhte Verkehrssicherheit. Unabhängig von ihrer Leistungsfähigkeit können autonome Fahrzeuge jedoch nicht alle Unfälle vermeiden, da technische Störungen nicht auszuschließen sind und die Fahrzeuge mit Verkehrsteilnehmenden interagieren, deren Verhalten schwer vorherzusagen ist und zuweilen von den Verkehrsregeln abweicht (z. B. Awad et al., 2018; Goodall, 2014a; Koopman & Wagner, 2017; Lin, 2016; Nyholm, 2018). Dementsprechend müssen autonome Fahrzeuge auch auf den Umgang mit Unfallsituationen vorbereitet werden (z. B. Lin, 2016; Nyholm, 2018). Dies schließt auch Unfallsituationen ein, die moralische Aspekte enthalten (z. B. Awad et al., 2018; Lin, 2016). Da die erfolgreiche Einführung autonomer Fahrzeuge eine breite gesellschaftliche Akzeptanz voraussetzt (z. B. Awad et al., 2018; Bergmann et al., 2018; Bonnefon et al., 2016; Lin, 2016), ist relevant welche Verhaltensweisen Menschen – also die potentiellen Nutzenden autonomer Fahrzeuge sowie andere Verkehrsteilnehmende – von autonomen Fahrzeugen in unvermeidlichen Unfallsituationen präferieren und wie sie auf geschehene Handlungen in Unfallsituationen reagieren. Forschung in diesen Bereichen kann dazu beitragen, die Reaktionen der Öffentlichkeit auf kritische Situationen mit autonomen Fahrzeugen abzuschätzen (z. B. Goodall, 2016a) und so Schwierigkeiten hinsichtlich der Akzeptanz autonomer Fahrzeuge zu antizipieren. Ziel der vorliegenden Arbeit war es daher, verschiedene Einflussfaktoren auf die Handlungspräferenzen für autonome Fahrzeuge und Handlungsbewertungen in unvermeidlichen Unfallsituationen zu untersuchen. Dafür wurden abstrakte Unfallszenarien in Form von moralischen Dilemmas verwendet. Auch wenn solche Unfallsituationen potentiell selten vorkommen, sind sie dennoch emotional salient (Bonnefon et al., 2016). Des Weiteren scheinen potentielle Kundinnen und Kunden moralische Dilemmas als relevante Herausforderung hinsichtlich der Einführung autonomer Fahrzeuge zu betrachten (Gill, 2021).

Im Fokus der Experimente 1a bis 2b stand der Einfluss der Perspektive, aus der Unfallszenarien mit einem autonomen Fahrzeug betrachtet werden, auf die präferierten Handlungen des autonomen Fahrzeugs. Tatsächlich ließ sich ein deutlicher Einfluss der Perspektive beobachten: Die Wahrscheinlichkeit, die Personen/en im Fahrzeug zu opfern, unterschied sich konstant zwischen den in den Unfall involvierten Perspektiven, während die Wahrscheinlichkeit, die Person/en im Fahrzeug zu opfern, die für die Beobachter-Perspektive beobachtet wurde, zumeist zwischen den Wahrscheinlichkeiten der involvierten Perspektiven lag. Die deutliche Divergenz in der Wahrscheinlichkeit, die Personen/en im Fahrzeug zu opfern, zwischen den beiden involvierten Perspektiven weist auf eine Tendenz zum Schutz der jeweils eigenen Perspektive und somit zum Selbstschutz hin. Dies steht im Einklang mit den Ergebnissen von Kallioinen et al. (2019) und von Frank et al. (2019), die in ihren Studien erste Hinweise auf Tendenzen zum Selbstschutz beobachten konnten. Interessanterweise scheinen die in der vorliegenden Arbeit beobachteten Selbstschutztendenzen ausgeprägter zu sein als in der Studie von Kallioinen et al. (2019), in der immersive virtuelle Umgebungen verwendet wurden. In ihrem ersten Experiment beobachteten Kallioinen et al. (2019) einen Konflikt zwischen den Perspektiven nur in einem spezifischen Szenario, in dem schwerwiegende Verletzungen für die Personen im Fahrzeug sehr wahrscheinlich waren. Eine mögliche Erklärung für die in der vorliegenden Arbeit ausgeprägteren Selbstschutztendenzen könnte sein, dass in den vorliegenden Experimenten die tödlichen Unfallfolgen für die jeweiligen Unfallparteien eindeutig kommuniziert wurden. Szenarien, in denen eine der beiden Unfallparteien in Folge des Unfalls stirbt, bergen möglicherweise ein stärkeres Konfliktpotential zwischen den involvierten Parteien als Szenarien mit weniger eindeutigen Unfallfolgen. Dass sich jedoch Selbstschutztendenzen sowohl in virtuellen Umgebungen als auch in abstrakten Szenarien beobachten lassen, legt den Schluss nahe, dass es sich um einen persistierenden kognitiven Bias handelt, der die Präferenzen in moralischen Dilemmas mit autonomen Fahrzeugen beeinflusst. In Folge dessen könnte der Fokus auf eine einzelne Interessensgruppe wie etwa Fahrzeuginsassinnen und -insassen, beispielsweise im Rahmen der Programmierung autonomer Fahrzeuge, für die Akzeptanz autonomer Fahrzeuge problematisch sein. Ein sorgfältiges Abwägen verschiedener Interessen erscheint daher sinnvoll.

Aus den Hinweisen auf einen Interessenskonflikt verschiedener Verkehrsteilnehmender ergibt sich die Frage, wie sich die verschiedenen Interessen einander

annähern lassen. Die Ergebnisse der Experimente 1a bis 2b weisen darauf hin, dass utilitaristische Präferenzen den Selbstschutztendenzen der involvierten Parteien entgegenwirken können. Alle Perspektiven zeigten eine zunehmende Präferenz zum Schutz der größeren Gruppe, je mehr Personen gefährdet wurden – unabhängig davon, ob es sich dabei um Personen im Fahrzeug oder auf der Straße handelte. Zum einen stützt dies Befunde, die Präferenzen für utilitaristische Handlungen nahelegen (z. B. Awad et al., 2018; Faulhaber et al., 2019; Kallioinen et al., 2019; Li et al., 2016), zum anderen lassen diese Ergebnisse den Schluss zu, dass sich eine gewisse Einigkeit zwischen verschiedenen Interessensgruppen durch utilitaristische Überlegungen erreichen lassen könnte. Weder Fahrzeuginsassinnen und -insassen noch Passantinnen und Passanten scheinen stets ihr eigenes Leben schützen zu wollen.

Ein möglicher Kritikpunkt an der Annahme, dass utilitaristische Überlegungen zu einer Annäherung verschiedener Perspektiven beitragen können, könnte darin bestehen, dass die utilitaristischen Präferenzen und die damit einhergehende zunehmende Bereitschaft, sich selbst für eine Gruppe anderer zu opfern, nicht den realen Einstellungen der verschiedenen Verkehrsteilnehmenden entspricht. Stattdessen könnte die zunehmende utilitaristische Bereitschaft zur Selbstopferung ein Artefakt individueller Bemühungen sein, nicht selbstsüchtig oder kaltherzig zu erscheinen weil das eigene Leben priorisiert wird. In Experiment 3 gaben jedoch sowohl in einer klassischen direkten Befragung als auch in einer indirekten Befragung ca. 60 % der Teilnehmenden an, dass ein autonomes Fahrzeug sie als Passantin oder Passanten zugunsten von fünf Personen im Fahrzeug opfern sollte. Utilitaristische Präferenzen scheinen daher nicht von den Teilnehmenden angegeben worden zu sein, um nicht zugeben zu müssen, dass sie mehrere andere Personen für ihr eigenes Überleben opfern würden. Somit konnte kein Einfluss sozialer Erwünschtheit auf die Handlungspräferenzen für autonome Fahrzeuge in moralischen Dilemmas festgestellt werden, wodurch die Validität der Befunde aus den Experimenten 1a bis 2b, aber auch aus anderen Studien, die auf utilitaristische Präferenzen hinweisen (z. B. Awad et al., 2018; Kallioinen et al., 2019), gestützt wird.

Die Untersuchung von Handlungspräferenzen für autonome Fahrzeuge, die in den Experimenten 1a bis 3 betrachtet wurden, ermöglicht Rückschlüsse darüber, welche Handlungen von autonomen Fahrzeugen in verschiedenen Situationen erwartet werden oder erwünscht sind. Dies wiederum kann Hinweise darauf geben, welche Handlungsansätze für autonome Fahrzeuge sozial akzeptabel und welche

Handlungsansätze problematisch sein könnten. Neben diesen Handlungspräferenzen für autonome Fahrzeuge ist für die Akzeptanz autonomer Fahrzeuge auch bedeutsam, wie die jeweiligen Handlungen bewertet werden, wenn sie bereits geschehen sind. Diese Handlungsbewertung könnte für die Wahrnehmung von Unfällen mit autonomen Fahrzeugen – insbesondere im Vergleich zu der Wahrnehmung von Unfällen mit menschlichen Fahrenden – besonders relevant sein. Selbst wenn sowohl für menschliche Fahrende als auch für autonome Fahrzeuge a priori die gleichen Handlungen in einer Unfallsituation präferiert werden (z. B. Bonnefon et al., 2016; Kallioinen et al., 2019), kann sich die retrospektive Beurteilung der Handlungen unterscheiden (siehe auch Scheutz & Malle, 2021). Daher wurde weiterführend in den Experimenten 4 und 5 die moralische Bewertung von Handlungen autonomer Fahrzeuge und menschlicher Fahrender verglichen. Die Ergebnisse zeigten, dass utilitaristische Prinzipien die moralische Bewertung der Handlungen autonomer Fahrzeuge und menschlicher Fahrender beeinflussten. Mit einer zunehmenden Anzahl an Personen auf der Straße wurde die Handlung, diese zu retten, zunehmend als moralisch vertretbar bewertet und die komplementäre Handlung, die Person im Fahrzeug zu retten, als zunehmend moralisch verwerflich. Dies weist darauf hin, dass in den untersuchten Szenarien ähnliche Kriterien für die Bewertung der Handlungen beider Handelnden herangezogen wurden und korrespondiert darüber hinaus mit den utilitaristischen Tendenzen, die in den Experimenten 1a bis 3 beobachtet wurden. Nichtsdestotrotz wurden die Handlungen des Menschen durchgehend als moralisch vertretbarer bewertet als die Handlungen des autonomen Fahrzeugs. Eine Vermenschlichung des autonomen Fahrzeugs reduzierte in Experiment 5 die Bewertungsunterschiede, auch wenn diese nicht vollständig eliminiert wurden.

Die negativere Bewertung der Handlungen autonomer Fahrzeuge im Vergleich zu der Bewertung der Handlungen menschlicher Fahrender steht im Einklang mit einer Abneigung gegenüber Maschinen, die moralische Entscheidungen treffen (Bigman & Gray, 2018; Gogoll & Uhl, 2018), und könnte zu dem teilweise beobachteten Unwillen beitragen, die Kontrolle über die Fahraufgabe an das Fahrzeug abzugeben (z. B. König & Neumayr, 2017; Wolf, 2016). Darüber hinaus korrespondieren die vorliegenden Ergebnisse mit den Befunden von Young und Monroe (2019), dass autonomen Fahrzeugen mehr Schuld für ihre Handlungen zugewiesen wird als menschlichen Fahrenden. Zusätzlich gibt es Hinweise darauf, dass Unfälle mit autonomen Fahrzeugen als schwerwiegender eingeschätzt werden als Unfälle mit

menschlichen Fahrenden (Liu et al., 2019) und dass die Toleranz möglicher Risiken autonomer Fahrzeuge im Vergleich zu der Toleranz möglicher Risiken menschlicher Fahrender geringer sein könnte (z. B. Liu et al., 2020).

Insgesamt scheint menschlichen Fahrenden mehr Nachsicht für ihre Handlungen in Unfällen entgegengebracht zu werden als autonomen Fahrzeugen. Eine mögliche Erklärung dafür könnte sein, dass sich Menschen leichter in die Position menschlicher Fahrender hineinversetzen, einen möglichen Entscheidungskonflikt besser nachvollziehen und somit leichter rechtfertigen können als bei einem autonomen Fahrzeug (Scheutz & Malle, 2021). Zusätzlich könnte bei menschlichen Fahrenden berücksichtigt werden, dass sie ihre Entscheidungen in Sekundenbruchteilen treffen müssen (siehe auch z. B. Lin, 2016) und ihnen dadurch keine Absicht oder die systematische Benachteiligung bestimmter Verkehrsteilnehmender unterstellt wird. Autonome Fahrzeuge könnten hingegen die Erwartungshaltung hervorrufen, dass sie die Leistungen menschlicher Fahrender übertreffen, wie beispielsweise durch die Hoffnung auf eine erhöhte Verkehrssicherheit angedeutet wird. Um keine übersteigerte Erwartungshaltung etwa hinsichtlich der Unfallvermeidung durch autonome Fahrzeuge zu erzeugen und möglichen Erwartungsverletzungen entgegenzuwirken, scheint es daher sinnvoll, die Vorteile und Risiken von autonomen Fahrzeugen offen zu kommunizieren (siehe auch Shariff et al., 2017).

Betrachtet man die Ergebnisse der Experimente 1a bis 5 zusammen, lässt sich festhalten, dass utilitaristische Präferenzen zum Schutz der größtmöglichen Anzahl beteiligter Personen sowohl die präferierten Handlungen autonomer Fahrzeuge in unvermeidlichen Unfallsituationen beeinflussten als auch die moralische Bewertung verschiedener Handlungen autonomer Fahrzeuge und menschlicher Fahrender. Utilitaristische Tendenzen zeigten sich in der Fahrzeuginsassen-, Passanten- und Beobachter-Perspektive (Experimente 1a bis 2b), schienen nicht von sozialer Erwünschtheit verzerrt zu werden (Experiment 3) und führten zu moralisch positiveren Handlungsbewertungen für menschliche Fahrende und autonome Fahrzeuge in Unfallsituationen (Experimente 4 und 5). Da in der vorliegenden Arbeit abstrakte Szenarien in Form von Textvignetten und Bildern verwendet wurden, sind die hier identifizierten Präferenzen allerdings möglicherweise primär repräsentativ für Situationen oder Entscheidungen ohne immanente Bedrohung, wie beispielsweise Überlegungen zum Kauf eines autonomen Fahrzeugs. Inwiefern sich die beobachteten Präferenzen auf Extremsituationen mit tatsächlichen Entscheidungen über Leben und

Tod übertragen lassen, kann auf Grundlage der vorliegenden Ergebnisse nicht beurteilt werden. Allerdings beobachteten Kallioinen et al. (2019), die die Perspektiven in einer Unfallsituation in einer virtuellen Umgebung manipulierten, ähnliche Präferenzen sowie limitierte Selbstschutztendenzen. Auch in anderen Simulationsstudien finden sich Hinweise auf Präferenzen für utilitaristische Handlungen in Unfallsituationen (z. B. Bergmann et al., 2018; Faulhaber et al., 2019). Hinsichtlich der moralischen Handlungsbewertung in Unfallsituationen ähneln abstrakte Szenarien jedoch zu einem gewissen Grad Zeitungsberichten über Unfälle mit autonomen Fahrzeugen. Da Zeitungsberichte und Nachrichtenmeldungen Unfälle und deren Folgen häufig beschreiben, sind sie potentiell wenig immersiv. Um zu antizipieren, wie Menschen auf Unfälle mit autonomen Fahrzeugen reagieren, wenn sie darüber in der Zeitung lesen, erscheinen abstrakte Szenarien daher geeignet. Zusätzlich ist anzunehmen, dass die meisten Personen, vor allem in der Anfangsphase der Einführung autonomer Fahrzeuge, eher aus der Zeitung oder den Nachrichten von Unfällen mit autonomen Fahrzeugen erfahren als selbst als Zeuginnen und Zeugen oder Unfallbeteiligte in einen solchen Unfall involviert zu sein.

Neben dem übergreifenden Einfluss utilitaristischer Präferenzen deuten sich in den vorliegenden Experimenten auch Konflikte und Schwierigkeiten an, die bei der Einführung autonomer Fahrzeuge relevant sein könnten. Zum einen weisen die Experimente 1a bis 2b auf mögliche Interessenskonflikte zwischen verschiedenen Verkehrsteilnehmenden hinsichtlich der präferierten Handlungen autonomer Fahrzeuge im Sinne von Selbstschutztendenzen hin, zum anderen wurden in den Experimenten 4 und 5 die Handlungen autonomer Fahrzeuge moralisch kritischer bewertet als die Handlungen menschlicher Fahrender. Allerdings legen die vorliegenden Ergebnisse auch Annäherungs- und Konsensmöglichkeiten nahe. Utilitaristische Präferenzen scheinen den Selbstschutztendenzen involvierter Perspektiven entgegenwirken und eine Annäherung der perspektivenspezifischen Präferenzen unterstützen zu können (Experimente 1a bis 2b). Die Divergenz in der moralischen Bewertung von Handlungen autonomer Fahrzeuge und menschlicher Fahrender scheint sich über eine Anthropomorphisierung des autonomen Fahrzeugs verringern zu lassen (Experiment 5). Solche Prinzipien und Maßnahmen, die zu einer Annäherung von verschiedenen Interessen oder auch zu einer ähnlicheren Bewertung von kritischen Ereignissen mit autonomen Fahrzeugen und menschlichen Fahrenden beitragen können, haben das Potential, die öffentliche Akzeptanz autonomer Fahrzeuge sowie deren Nutzung

positiv zu beeinflussen. Das hier untersuchte Prinzip des Utilitarismus und der Gestaltungsansatz des Anthropomorphismus stellen zwei mögliche Ansatzpunkte dar. Perspektivisch wären weitere Prinzipien und Ansätze denkbar, wie beispielsweise eine perspektivenübergreifende Einigung auf die Befolgung von Verkehrsregeln, eine Verdeutlichung der stetig wechselnden eigenen Rollen im Straßenverkehr, wodurch im Vorfeld unklar ist, in welcher Rolle man persönlich in einen Unfall mit einem autonomen Fahrzeug involviert sein könnte (siehe auch Schleier des Nichtwissens; für empirische Befunde hierzu siehe z. B. Huang et al., 2019) oder auch die Transparenz von Entscheidungsregeln für autonome Fahrzeuge in Unfallsituationen. Darüber hinaus ist zu beachten, dass sich die Einstellungen gegenüber autonomen Fahrzeugen durchaus mit zunehmender Erfahrung mit und mehr Wissen über die zugrundeliegenden Technologien verändern könnten. So finden sich beispielsweise Hinweise darauf, dass praktische Erfahrung mit autonomen Fahrzeugen wie eine Testfahrt das Vertrauen in die Technologie erhöhen kann (z. B. Xu et al., 2018) und dass die Einstellung gegenüber autonomen Fahrzeugen positiver ausfällt, je mehr Wissen über diese Technologie vorhanden ist (König & Neumayr, 2017).

Zusammenfassend lässt sich festhalten, dass die Perspektive, aus der moralische Dilemmas mit autonomen Fahrzeugen beurteilt werden, die präferierten Handlungen autonomer Fahrzeuge beeinflusste. Teilnehmende mit den Perspektiven der in den Unfall involvierte Parteien zeigten Tendenzen sich selbst zu schützen, waren jedoch mit einer steigenden Anzahl potentieller Opfer zunehmend bereit, sich selbst zugunsten einer Gruppe anderer Verkehrsteilnehmender zu opfern. Des Weiteren wurde die moralische Bewertung von Handlungen in Unfallsituationen davon beeinflusst, ob ein Mensch oder ein autonomes Fahrzeug handelte. Die Handlungen menschlicher Fahrender wurden moralisch positiver bewertet als die Handlungen autonomer Fahrzeuge, auch wenn utilitaristische Überlegungen die Bewertung der Handlungen beider Handelnden beeinflussten. Ein Einfluss sozialer Erwünschtheit auf utilitaristische Tendenzen konnte hingegen nicht festgestellt werden.

Literatur

Anania, E. C., Rice, S., Walters, N. W., Pierce, M., Winter, S. R., & Milner, M. N. (2018). The effects of positive and negative information on consumers' willingness to ride in a driverless vehicle. *Transport Policy*, *72*, 218-224. https://doi.org/10.1016/j.tranpol.2018.04.002

Anderson, J. M., Nidhi, K., Stanley, K. D., Sorensen, P., Samaras, C., & Oluwatola, O. A. (2016). *Autonomous vehicle technology: A guide for policymakers* (Rev. ed.). RAND Corporation. https://doi.org/10.7249/RR443-2

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The Moral Machine experiment. *Nature*, *563*(7729), 59-64. https://doi.org/10.1038/s41586-018-0637-6

Bagloee, S. A., Tavana, M., Asadi, M., & Oliver, T. (2016). Autonomous vehicles: Challenges, opportunities, and future implications for transportation policies. *Journal of Modern Transportation*, *24*(4), 284-303. https://doi.org/10.1007/s40534-016-0117-3

Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics*, *1*(1), 71-81. https://doi.org/10.1007/s12369-008-0001-3

Becker, F., & Axhausen, K. W. (2017). Literature review on surveys investigating the acceptance of automated vehicles. *Transportation*, *44*(6), 1293-1306. https://doi.org/10.1007/s11116-017-9808-9

Bentham, J. (1789). *An introduction to the principles of morals and legislation*. T. Payne and son. Retrieved April 20, 2022, from http://galenet.galegroup.com/servlet/MOME?af=RN&ae=U102143420&srchtp=a&ste=14

Bergmann, L. T., Schlicht, L., Meixner, C., König, P., Pipa, G., Boshammer, S., & Stephan, A. (2018). Autonomous vehicles require socio-political acceptance—An empirical and philosophical perspective on the problem of moral decision making. *Frontiers in Behavioral Neuroscience*, *12*, Article 31. https://doi.org/10.3389/fnbeh.2018.00031

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition*, *181*, 21-34. https://doi.org/10.1016/j.cognition.2018.08.003

Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*, *352*(6293), 1573-1576. https://doi.org/10.1126/science.aaf2654

Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2019). The trolley, the bull bar, and why engineers should care about the ethics of autonomous cars [point of view]. *Proceedings of the IEEE*, *107*(3), 502-504. https://doi.org/10.1109/JPROC.2019.2897447

Bundesanstalt für Straßenwesen. (2021, March 11). *Selbstfahrende Autos – assistiert, automatisiert oder autonom?* [Press release]. Retrieved July 28, 2021, from https://www.bast.de/BASt_2017/DE/Presse/Mitteilungen/2021/06-2021.html

Bundesministerium für Verkehr und digitale Infrastruktur. (2017). *Ethik-Kommission Automatisiertes und Vernetztes Fahren – Bericht Juni 2017*. Retrieved January 22, 2018, from https://www.bmvi.de/SharedDocs/DE/Publikationen/DG/bericht-der-ethik-kommission.pdf?__blob=publicationFile

Cushman, F., & Greene, J. D. (2012). Finding faults: How moral dilemmas illuminate cognitive structure. *Social Neuroscience*, *7*(3), 269-279. https://doi.org/10.1080/17470919.2011.614000

Dawes, R. M. (1980). Social dilemmas. *Annual Review of Psychology*, *31*, 169-193. https://doi.org/10.1146/annurev.ps.31.020180.001125

Deutsche Bahn AG. (2019, April 15). *Mit der Bahn wie zu Hause fühlen. Neue Fernverkehrkampagne: DB als Gastgeber der Zukunft* [Press release]. Retrieved August 2, 2021, from https://web.archive.org/web/20210416141258/https://www.deutschebahn.com/de/presse/pressestart_zentrales_uebersicht/Mit-der-Bahn-wie-zu-Hause-fuehlen-4089996

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, *114*(4), 864-886. https://doi.org/10.1037/0033-295X.114.4.864

Faulhaber, A. K., Dittmer, A., Blind, F., Wächter, M. A., Timm, S., Sütfeld, L. R., Stephan, A., Pipa, G., & König, P. (2019). Human decisions in moral dilemmas are largely described by utilitarianism: Virtual car driving study provides guidelines for autonomous driving vehicles. *Science and Engineering Ethics*, *25*(2), 399-418. https://doi.org/10.1007/s11948-018-0020-x

Favarò, F. M., Nader, N., Eurich, S. O., Tripp, M., & Varadaraju, N. (2017). Examining accident reports involving autonomous vehicles in California. *PLoS ONE*, *12*(9), Article e0184952. https://doi.org/10.1371/journal.pone.0184952

Foot, P. (1967). The problem of abortion and the doctrine of double effect. *Oxford Reviews*, *5*, 5-15.

Frank, D.-A., Chrysochou, P., Mitkidis, P., & Ariely, D. (2019). Human decision-making biases in the moral dilemmas of autonomous vehicles. *Scientific Reports*, *9*, Article 13080. https://doi.org/10.1038/s41598-019-49411-7

Franklin, M., Awad, E., & Lagnado, D. (2021). Blaming automated vehicles in difficult situations. *Iscience*, *24*(4), Article 102252. https://doi.org/10.1016/j.isci.2021.102252

Gawronski, B., Armstrong, J., Conway, P., Friesdorf, R., & Hütter, M. (2017). Consequences, norms, and generalized inaction in moral dilemmas: The CNI model of moral decision-making. *Journal of Personality and Social Psychology*, *113*(3), 343-376. https://doi.org/10.1037/pspa0000086

Gill, T. (2020). Blame it on the self-driving car: How autonomous vehicles can alter consumer morality. *Journal of Consumer Research*, *47*(2), 272-291. https://doi.org/10.1093/jcr/ucaa018

Gill, T. (2021). Ethical dilemmas are really important to potential adopters of autonomous vehicles. *Ethics and Information Technology*, *23*(4), 657-673. https://doi.org/10.1007/s10676-021-09605-y

Gogoll, J., & Müller, J. F. (2017). Autonomous cars: In favor of a mandatory ethics setting. *Science and Engineering Ethics*, *23*(3), 681-700. https://doi.org/10.1007/s11948-016-9806-x

Gogoll, J., & Uhl, M. (2018). Rage against the machine: Automation in the moral domain. *Journal of Behavioral and Experimental Economics*, *74*, 97-103. https://doi.org/10.1016/j.socec.2018.04.003

Gong, L. (2008). How social is social responses to computers? The function of the degree of anthropomorphism in computer representations. *Computers in Human Behavior*, *24*(4), 1494-1509. https://doi.org/10.1016/j.chb.2007.05.007

Goodall, N. J. (2014a). Ethical decision making during automated vehicle crashes. *Transportation Research Record*, *2424*(1), 58-65. https://doi.org/10.3141/2424-07

Goodall, N. J. (2014b). Machine ethics and automated vehicles. In G. Meyer & S. Beiker (Eds.), *Road vehicle automation* (pp. 93-102). Springer. https://doi.org/10.1007/978-3-319-05990-7_9 (Corrected version of the book published 2018)

Goodall, N. J. (2016a). Away from trolley problems and toward risk management. *Applied Artificial Intelligence*, *30*(8), 810-821. https://doi.org/10.1080/08839514.2016.1229922

Goodall, N. J. (2016b). Can you program ethics into a self-driving car? *IEEE Spectrum*, *53*(6), 28-58. https://doi.org/10.1109/MSPEC.2016.7473149

Greene, J. D. (2016). Our driverless dilemma. *Science*, *352*(6293), 1514-1515. https://doi.org/10.1126/science.aaf9534

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, *44*(2), 389-400. https://doi.org/10.1016/j.neuron.2004.09.027

Hauser, M., Cushman, F., Young, L., Kang-Xing Jin, R., & Mikhail, J. (2007). A dissociation between moral judgments and justifications. *Mind & Language*, *22*(1), 1-21. https://doi.org/10.1111/j.1468-0017.2006.00297.x

Heck, D. W., Hoffmann, A., & Moshagen, M. (2018). Detecting nonadherence without loss in efficiency: A simple extension of the crosswise model. *Behavior Research Methods*, *50*(5), 1895-1905. https://doi.org/10.3758/s13428-017-0957-8

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*(2), 65-70. http://www.jstor.org/stable/4615733

Hong, J.-W., Wang, Y., & Lanz, P. (2020). Why is artificial intelligence blamed more? Analysis of faulting artificial intelligence for self-driving car accidents in experimental settings. *International Journal of Human–Computer Interaction*, *36*(18), 1768-1774. https://doi.org/10.1080/10447318.2020.1785693 Correction: *International Journal of Human-Computer Interaction*, *38*(1), 102. https://doi.org/10.1080/10447318.2021.2004139

Huang, K., Greene, J. D., & Bazerman, M. (2019). Veil-of-ignorance reasoning favors the greater good. *Proceedings of the National Academy of Sciences*, *116*(48), 23989-23995. https://doi.org/10.1073/pnas.1910125116

Jelinski, L., Etzrodt, K., & Engesser, S. (2021). Undifferentiated optimism and scandalized accidents: The media coverage of autonomous driving in Germany. *Journal of Science Communication*, *20*(4), Article A02. https://doi.org/10.22323/2.20040202

Kallioinen, N., Pershina, M., Zeiser, J., Nosrat Nezami, F., Pipa, G., Stephan, A., & König, P. (2019). Moral judgements on the actions of self-driving cars and human drivers in dilemma situations from different perspectives. *Frontiers in Psychology*, *10*, Article 2415. https://doi.org/10.3389/fpsyg.2019.02415

Kant, I. (2011). *Groundwork of the metaphysics of morals. A German-English edition.* (M. Gregor & J. Timmermann, Eds. & Trans.). Cambridge University Press. https://doi.org/10.1017/CBO9780511973741 (Original work published 1786)

Keeling, G. (2020). Why trolley problems matter for the ethics of automated vehicles. *Science and Engineering Ethics*, *26*(1), 293-307. https://doi.org/10.1007/s11948-019-00096-1

König, M., & Neumayr, L. (2017). Users' resistance towards radical innovations: The case of the self-driving car. *Transportation Research Part F: Traffic Psychology and Behaviour*, *44*, 42-52. https://doi.org/10.1016/j.trf.2016.10.013

Koopman, P., & Wagner, M. (2017). Autonomous vehicle safety: An interdisciplinary challenge. *IEEE Intelligent Transportation Systems Magazine*, *9*(1), 90-96. https://doi.org/10.1109/MITS.2016.2583491

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: A literature review. *Quality & Quantity*, *47*(4), 2025-2047. https://doi.org/10.1007/s11135-011-9640-9

Lee, J.-G., Kim, K. J., Lee, S., & Shin, D.-H. (2015). Can autonomous vehicles be safe and trustworthy? Effects of appearance and autonomy of unmanned driving systems. *International Journal of Human-Computer Interaction*, *31*(10), 682-691. https://doi.org/10.1080/10447318.2015.1070547

Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies*, *40*(1), 153-184. https://doi.org/10.1006/ijhc.1994.1007

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50-80. https://doi.org/10.1518/hfes.46.1.50_30392

Lensvelt-Mulders, G. J. L. M., Hox, J. J., Van der Heijden, P. G. M., & Maas, C. J. M. (2005). Meta-analysis of randomized response research: Thirty-five years of validation. *Sociological Methods & Research*, *33*(3), 319-348. https://doi.org/10.1177/0049124104268664

Li, J., Zhao, X., Cho, M.-J., Ju, W., & Malle, B. F. (2016). *From trolley to autonomous vehicle: Perceptions of responsibility and moral norms in traffic accidents with self-driving cars* (SAE Technical Paper 2016-01-0164). SAE International. https://doi.org/10.4271/2016-01-0164

Lin, P. (2016). Why ethics matters for autonomous cars. In M. Maurer, J. C. Gerdes, B. Lenz, & H. Winner (Eds. & Trans.), *Autonomous driving* (pp. 69-85). Springer. https://doi.org/10.1007/978-3-662-48847-8_4 (Original German version of the book published 2015)

Liu, P., Du, Y., & Xu, Z. (2019). Machines versus humans: People's biased responses to traffic accidents involving self-driving vehicles. *Accident Analysis & Prevention*, *125*, 232-240. https://doi.org/10.1016/j.aap.2019.02.012

Liu, P., & Liu, J. (2021). Selfish or utilitarian automated vehicles? Deontological evaluation and public acceptance. *International Journal of Human-Computer Interaction*, *37*(13), 1231-1242. https://doi.org/10.1080/10447318.2021.1876357

Liu, P., Wang, L., & Vincent, C. (2020). Self-driving vehicles against human drivers: Equal safety is far from enough. *Journal of Experimental Psychology: Applied*, *26*(4), 692-704. https://doi.org/10.1037/xap0000267

Malle, B. F., Magar, S. T., & Scheutz, M. (2019). AI in the sky: How people morally evaluate human and machine decisions in a lethal strike dilemma. In M. I. Aldinhas Ferreira, J. Silva Sequeira, G. Singh Virk, M. O. Tokhi, & E. E. Kader (Eds.), *Robotics and well-being* (pp. 111-133). Springer. https://doi.org/10.1007/978-3-030-12524-0_11

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015). Sacrifice one for the good of many? People apply different moral norms to human and robot agents. In *HRI'15 Proceedings of the 2015 ACM/IEEE international conference on human-robot interaction* (pp. 117-124). Association for Computing Machinery. https://doi.org/10.1145/2696454.2696458

Malle, B. F., Scheutz, M., Forlizzi, J., & Voiklis, J. (2016). Which robot am I thinking about? The impact of action and appearance on people's evaluations of a moral robot. In *HRI'16 The eleventh ACM/IEEE international conference on human robot interaction* (pp. 125-132). Institute of Electrical and Electronics Engineers, Inc. https://doi.org/10.1109/HRI.2016.7451743

Mayer, M. M., Bell, R., & Buchner, A. (2021). Self-protective and self-sacrificing preferences of pedestrians and passengers in moral dilemmas involving autonomous vehicles. *PLoS ONE*, *16*(12), Article e0261673. https://doi.org/10.1371/journal.pone.0261673

Meisters, J., Hoffmann, A., & Musch, J. (2020). Controlling social desirability bias: An experimental investigation of the extended crosswise model. *PLoS ONE*, *15*(12), Article e0243384. https://doi.org/10.1371/journal.pone.0243384

Messick, D. M., & Brewer, M. B. (1983). Solving social dilemmas: A review. In L. Wheeler & P. Shaver (Eds.), *Review of personality and social psychology* (Vol. 4, pp. 11-44). Sage Publications.

Mieth, L., Mayer, M. M., Hoffmann, A., Buchner, A., & Bell, R. (2021). Do they really wash their hands? Prevalence estimates for personal hygiene behaviour during the COVID-19 pandemic based on indirect questions. *BMC Public Health*, *21*, Article 12. https://doi.org/10.1186/s12889-020-10109-5

Mill, J. S. (2010). *Utilitarianism / Der Utilitarismus* (D. Birnbacher, Ed. & Trans.; Reprinted ed.). Reclam. (Original work published 1871)

Muir, B. M., & Moray, N. (1996). Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, *39*(3), 429-460. https://doi.org/10.1080/00140139608964474

National Highway Traffic Safety Administration. (2008). *National Motor Vehicle Crash Causation Survey—Report to Congress* (DOT HS 811 059). Retrieved August 12, 2020, from https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/811059

National Transportation Safety Board. (2017). *Collision between a car operating with automated vehicle control systems and a tractor-semitrailer truck near Williston, Florida, May 7, 2016* (NTSB/HAR-17/02, Product No. PB2017-102600). Retrieved October 8, 2021, from https://www.ntsb.gov/investigations/AccidentReports/Reports/HAR1702.pdf

National Transportation Safety Board. (2019). *Collision between vehicle controlled by developmental automated driving system and pedestrian, Tempe, Arizona, March 18, 2018* (NTSB/HAR-19/03, Product No. PB2019-101402). Retrieved October 15, 2021, from https://www.ntsb.gov/investigations/AccidentReports/Reports/HAR1903.pdf

Niu, D., Terken, J., & Eggen, B. (2018). Anthropomorphizing information to enhance trust in autonomous vehicles. *Human Factors and Ergonomics in Manufacturing & Service Industries*, *28*(6), 352-359. https://doi.org/10.1002/hfm.20745

Nobis, C., & Kuhnimhof, T. (2019). *Mobilität in Deutschland – MiD Ergebnisbericht. Studie von infas, DLR, IVT und infas 360 im Auftrag des Bundesministers für Verkehr und digitale Infrastruktur. Version 1.1* (FE No. 70.904/15, Project No. 5431). Retrieved June 11, 2019, from http://www.mobilitaet-in-deutschland.de/pdf/MiD2017_Ergebnisbericht.pdf

Nyholm, S. (2018). The ethics of crashes with self-driving cars: A roadmap, I. *Philosophy Compass*, *13*(7), Article e12507. https://doi.org/10.1111/phc3.12507

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), Article aac4716. https://doi.org/10.1126/science.aac4716

Pak, R., Fink, N., Price, M., Bass, B., & Sturre, L. (2012). Decision support aids with anthropomorphic characteristics influence trust and performance in younger and older adults. *Ergonomics*, *55*(9), 1059-1072. https://doi.org/10.1080/00140139.2012.691554

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*(2), 230-253. https://doi.org/10.1518/001872097778543886

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2008). Situation awareness, mental workload, and trust in automation: Viable, empirically supported cognitive engineering constructs. *Journal of Cognitive Engineering and Decision Making*, *2*(2), 140-160. https://doi.org/10.1518/155534308X284417

Sachdeva, S., Iliev, R., Ekhtiari, H., & Dehghani, M. (2015). The role of self-sacrifice in moral dilemmas. *PLoS ONE*, *10*(6), Article e0127409. https://doi.org/10.1371/journal.pone.0127409

SAE International. (2021). *Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles* (SAE Ground Vehicle Standard J3016_202104). Retrieved March 11, 2022, from https://www.sae.org/standards/content/j3016_202104/

Scheutz, M., & Malle, B. F. (2021). May machines take lives to save lives? Human perceptions of autonomous robots (with the capacity to kill). In J. Galliott, D. MacIntosh, & J. D. Ohlin (Eds.), *Lethal autonomous weapons: Re-examining the law and ethics of robotic warfare* (pp. 89-101). Oxford University Press. https://doi.org/10.1093/oso/9780197546048.003.0007

Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2017). Psychological roadblocks to the adoption of self-driving vehicles. *Nature Human Behaviour*, *1*(10), 694-696. https://doi.org/10.1038/s41562-017-0202-6

Smith, A., & Anderson, M. (2017). *Automation in everyday life*. Pew Research Center. Retrieved August 12, 2021, from https://www.pewresearch.org/internet/2017/10/04/automation-in-everyday-life/

Spieser, K., Treleaven, K., Zhang, R., Frazzoli, E., Morton, D., & Pavone, M. (2014). Toward a systematic approach to the design and evaluation of automated mobility-on-demand systems: A case study in Singapore. In G. Meyer & S. Beiker (Eds.), *Road vehicle automation* (pp. 229-245). Springer. https://doi.org/10.1007/978-3-319-05990-7_20 (Corrected version of the book published 2018)

Statistisches Bundesamt. (2020). *Verkehrsunfälle – Zeitreihen – 2019*. Retrieved March 31, 2021, from https://www.statistischebibliothek.de/mir/servlets/MCRFileNodeServlet/DEHeft_derivate_00061690/5462403197004_aktualisiert10032021.pdf

Sudman, S., & Bradburn, N. M. (1974). *Response effects in surveys: A review and synthesis*. ALDINE Publishing Company.

Sütfeld, L. R., Ehinger, B. V., König, P., & Pipa, G. (2019). How does the method change what we measure? Comparing virtual reality and text-based surveys for the assessment of moral decisions in traffic dilemmas. *PLoS ONE*, *14*(10), Article e0223108. https://doi.org/10.1371/journal.pone.0223108

Sütfeld, L. R., Gast, R., König, P., & Pipa, G. (2017). Using virtual reality to assess ethical decisions in road traffic scenarios: Applicability of value-of-life-based models and influences of time pressure. *Frontiers in Behavioral Neuroscience*, *11*, Article 122. https://doi.org/10.3389/fnbeh.2017.00122

Thomson, J. J. (1976). Killing, letting die, and the trolley problem. *The Monist*, *59*(2), 204-217. https://doi.org/10.5840/monist197659224

Thomson, J. J. (1985). The trolley problem. *The Yale Law Journal*, *94*(6), 1395-1415. https://doi.org/10.2307/796133

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, *133*(5), 859-883. https://doi.org/10.1037/0033-2909.133.5.859

Volz, L. J., Welborn, B. L., Gobel, M. S., Gazzaniga, M. S., & Grafton, S. T. (2017). Harm to self outweighs benefit to others in moral decision making. *Proceedings of the National Academy of Sciences*, *114*(30), 7963-7968. https://doi.org/10.1073/pnas.1706693114

Waldrop, M. M. (2015). Autonomous vehicles: No drivers required. *Nature*, *518*(7537), 20-23. https://doi.org/10.1038/518020a

Wang, J., Zhang, L., Huang, Y., & Zhao, J. (2020). Safety of autonomous vehicles. *Journal of Advanced Transportation*, *2020*, Article 8867757. https://doi.org/10.1155/2020/8867757

Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, *60*(309), 63-69. https://doi.org/10.2307/2283137

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, *52*, 113-117. https://doi.org/10.1016/j.jesp.2014.01.005

Winkler, M., Mehl, R., Erander, H., Sule, S., Buvat, J., KVJ, S., Sengupta, A., & Khemka, Y. (2019). *The autonomous car: A consumer perspective*. Capgemini Research Institute. Retrieved October 19, 2021, from https://www.capgemini.com/wp-content/uploads/2019/05/30min-–-Report.pdf

Wolf, I. (2016). The interaction between humans and autonomous agents. In M. Maurer, J. C. Gerdes, B. Lenz, & H. Winner (Eds. & Trans.), *Autonomous driving* (pp. 103-124). Springer. https://doi.org/10.1007/978-3-662-48847-8_6 (Original German version of the book published 2015)

Wolkenstein, A. (2018). What has the Trolley Dilemma ever done for us (and what will it do in the future)? On some recent debates about the ethics of self-driving cars. *Ethics and Information Technology*, *20*(3), 163-173. https://doi.org/10.1007/s10676-018-9456-6

Xu, Z., Zhang, K., Min, H., Wang, Z., Zhao, X., & Liu, P. (2018). What drives people to accept automated vehicles? Findings from a field experiment. *Transportation Research Part C: Emerging Technologies*, *95*, 320-334. https://doi.org/10.1016/j.trc.2018.07.024

Young, A. D., & Monroe, A. E. (2019). Autonomous morals: Inferences of mind predict acceptance of AI behavior in sacrificial moral dilemmas. *Journal of Experimental Social Psychology*, *85*, Article 103870. https://doi.org/10.1016/j.jesp.2019.103870

Yu, J.-W., Tian, G.-L., & Tang, M.-L. (2008). Two new models for survey sampling with sensitive characteristic: Design and analysis. *Metrika*, *67*(3), 251-263. https://doi.org/10.1007/s00184-007-0131-x

Einzelarbeiten

Experimente 1a, 1b, 2a, 2b und 3:

Mayer, M. M., Bell, R., & Buchner, A. (2021). Self-protective and self-sacrificing preferences of pedestrians and passengers in moral dilemmas involving autonomous vehicles. *PLoS ONE, 16*(12), Article e0261673. https://doi.org/10.1371/journal.pone.0261673

Experimente 4 und 5:

Mayer, M. M., Buchner, A., & Bell, R. (2022). Men, machines, and double standards? The moral evaluation of the actions of autonomous vehicles and human drivers in road-accident scenarios. *Manuscript submitted for publication.*

PLOS ONE

RESEARCH ARTICLE

Self-protective and self-sacrificing preferences of pedestrians and passengers in moral dilemmas involving autonomous vehicles

Maike M. Mayer*, Raoul Bell, Axel Buchner

Department of Experimental Psychology, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

* maike.mayer@hhu.de

OPEN ACCESS

Citation: Mayer MM, Bell R, Buchner A (2021) Self-protective and self-sacrificing preferences of pedestrians and passengers in moral dilemmas involving autonomous vehicles. PLoS ONE 16(12): e0261673. https://doi.org/10.1371/journal.pone.0261673

Editor: Quan Yuan, Tsinghua University, CHINA

Received: May 17, 2021

Accepted: December 8, 2021

Published: December 23, 2021

Data Availability Statement: The data of all experiments reported here as well as more detailed information about the samples of Experiments 1a to 2b together with a description of the respective questionnaires as supplementary material are available at https://osf.io/4xhz7/.

Funding: The publication fee was paid by the open access fund of Heinrich Heine University Düsseldorf. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Upon the introduction of autonomous vehicles into daily traffic, it becomes increasingly likely that autonomous vehicles become involved in accident scenarios in which decisions have to be made about how to distribute harm among involved parties. In four experiments, participants made moral decisions from the perspective of a passenger, a pedestrian, or an observer. The results show that the preferred action of an autonomous vehicle strongly depends on perspective. Participants' judgments reflect self-protective tendencies even when utilitarian motives clearly favor one of the available options. However, with an increasing number of lives at stake, utilitarian preferences increased. In a fifth experiment, we tested whether these results were tainted by social desirability but this was not the case. Overall, the results confirm that strong differences exist among passengers, pedestrians, and observers about the preferred course of action in critical incidents. It is therefore important that the actions of autonomous vehicles are not only oriented towards the needs of their passengers, but also take the interests of other road users into account. Even though utilitarian motives cannot fully reconcile the conflicting interests of passengers and pedestrians, there seem to be some moral preferences that a majority of the participants agree upon regardless of their perspective, including the utilitarian preference to save several other lives over one's own.

Introduction

As autonomous driving technologies constantly improve, the introduction of automated and eventually fully autonomous vehicles into daily traffic for private and commercial uses is in the progress of being realized [1]. Many governments around the world are aware of the economic importance of automated driving and support the development and introduction of autonomous driving technologies [cf. 2–4]. The expected improvements to safety, accessibility of transportation, and traffic flow [cf. 1, 5] spur the interest in these technologies. Given the high number of annual traffic fatalities [globally about 1,35 million in 2016, 6, United States: 36,560 in 2018, 7, European Union: about 22,800 [estimated] in 2019, 8] and human error as a major cause of accidents [9], the prospect of increased traffic safety [e.g., 1, 5] is one of the most

Competing interests: The authors have declared that no competing interests exist.

salient advantages of automated driving [e.g., 10]. Nonetheless, autonomous vehicles cannot avoid all accidents—regardless of the system's reliability or the frequency of such incidents—as they share the roads with other road users such as pedestrians, human drivers, and animals whose behaviors are difficult to predict [11–14]. Thus, in order to enable autonomous vehicles to participate in public traffic, it is necessary to program them for how to handle accidents [12, 14]. While human drivers have to make split-second decisions in critical traffic situations, autonomous vehicles provide the unique opportunity for considering in advance how critical situations should be handled [10, 15–18]. However, there are two sides to every coin. Designing an algorithm to handle accidents and implementing it in numerous vehicles implies the danger that any intended or unintended bias introduced to the system may determine decisions about life and death [10, 19]. This is a particularly delicate matter as autonomous vehicles may face difficult moral decisions [12, 13, 15, 20] such as whom to harm or even sacrifice in an inevitable accident.

Various partially conflicting norms and principles—among which deontology and utilitarianism are probably the most prominent—affect moral decision making [e.g., 21, 22] including decisions about how autonomous vehicles should handle unavoidable collisions comprising moral aspects [e.g., 11, 14, 23–25]. While deontology focuses on moral rules [such as obligations and prohibitions, e.g., 26], utilitarianism is concerned with the outcome of particular actions. Following a deontological approach, an action is morally acceptable and permissible if it is consistent with moral norms (e.g., "You shall not kill") whereas from a utilitarian point-of-view an action is permissible and acceptable if it maximizes utility [e.g., 27, 28] by minimizing negative consequences such as overall damage or harm.

A popular approach to studying moral decision making is the Trolley Problem [29–31]. In the original version of this moral dilemma, a runaway trolley is speeding down the tracks. On the tracks, there are five people who are unable to move out of the way in time. It is, however, possible to lead the trolley to a side track where it will kill only one person. Is it morally acceptable to kill this person to save five other lives? It is easy to envision similar scenarios with autonomous vehicles: Imagine that five pedestrians suddenly step into a road upon which an autonomous vehicle is driving. The autonomous vehicle cannot come to a stop in time; it only has the option to either crash into the group of pedestrians or swerve to the side into an obstacle, killing the passenger. Should the autonomous vehicle sacrifice the passenger to save the lives of the pedestrians or should it sacrifice the pedestrians, leaving the passenger unharmed?

The question of how to program autonomous vehicles for handling accidents that require moral decisions has sparked interdisciplinary research and considerable debate. Scenarios modeled after the Trolley Problem have become standard tools to investigate moral dilemmas involving autonomous vehicles [e.g., 10, 14, 32]. Most prominently, in the Moral Machine experiment [13] different scenarios were tested against each other, involving millions of people from more than 200 countries. Among the strongest moral preferences identified in this study was the utilitarian preference to spare more lives, but there was considerable variation in preferences. Scenarios modeled after the Trolley Problem cannot serve as a blueprint for how to program autonomous vehicles [e.g., 32, 33], but they can serve to identify morally relevant properties of accident scenarios [e.g., 34], to test ethical theories [e.g., 16], and to examine moral intuitions and moral decision making [e.g., 32–34]. This is particularly relevant as public acceptance is a prerequisite for the success of autonomous vehicles [12, 13, 17, 18, 25, 32, 34, 35]. The programming of autonomous vehicles for handling moral decisions in accident scenarios requires careful consideration of what decisions people are willing to accept.

While from a societal perspective it may seem desirable that the actions of autonomous vehicles are guided by moral norms and aim at saving a maximum number of lives, research suggests that people's preferences are not only guided by moral and utilitarian considerations

but also by self-protective tendencies. From an evolutionary perspective, it seems possible that self-protective tendencies are ingrained in cognitive decision making [36] which implies that people show a preference for actions that protect their own life. In line with this prediction, Bonnefon et al.'s [35] participants indicated that they were unwilling to buy utilitarian autonomous vehicles for themselves despite agreeing that utilitarian programming represents a good —or morally superior—approach. Similarly, Liu and Liu [37] observed that participants showed a higher intention to use autonomous vehicles programmed to protect their passengers and were overall more willing to pay extra money for this type of self-driving technology compared to utilitarian autonomous vehicles. This pattern of results suggests that determining the actions of autonomous vehicles in critical accidents may represent a social dilemma [e.g., 18, 35, 38]. While it may be desirable for society to minimize the number of people harmed in accidents, customers might display a selfish interest to protect their own lives. In consequence, automated vehicles that value the lives of the passengers higher than that of other road users may prevail in the market. However, there are also results suggesting that there may be a limit to people's selfishness. Faulhaber et al. [18] observed an increasing willingness to self-sacrifice with an increasing number of potential victims that could be saved by a self-sacrifice. Taken together, people's preferences may be characterized as utilitarianism biased by self-protective tendencies. Specifically, most people agree that an autonomous vehicle should sacrifice their own life if this action saves the lives of many other people, but this utilitarian preference to reduce harm and save lives is limited by a tendency to value one's own physical safety more than that of another person. In consequence, a number of other people's lives have to be at stake before self-sacrifice is considered the preferred option.

So far, many governments have refrained from touching upon the moral dilemmas that may arise from unavoidable accidents involving the deaths of passengers and other road users while they have realized the importance of autonomous driving and have discussed concerns of traffic safety [2–4]. A notable exception is the German Federal Ministry of Transport and Digital Infrastructure whose Ethics Commission has published official guidelines for how autonomous vehicles should be programmed to handle morally relevant situations [39]. This is all the more interesting as these guidelines do not always align with the laypeople's preferences found in experimental studies. For example, the guidelines neither prescribe nor prohibit sacrificing few to protect many although several studies demonstrate that participants tolerate or even prefer a utilitarian approach for autonomous vehicles [e.g., 13, 18, 25, 35, 40, 41]. The guidelines also state that parties who do not generate a mobility risk (e.g., pedestrians) must not be sacrificed to save those generating that risk (the passengers of the autonomous vehicles). This suggestion is especially noticeable because it explicitly distinguishes between the safety concerns of different road users. However, most research has focused only on the perspectives of passengers and observers [e.g., 13, 18, 35, 40]. This is a narrowed perspective as other road users are also directly affected by the actions of autonomous vehicles and may well differ in their preferences for certain outcomes of moral dilemmas from passengers of autonomous vehicles. The perspective of the pedestrian seems particularly important because pedestrians represent the largest group of non-motorized road users [42].

To date, there are only few studies investigating to what degree moral preferences of non-motorized road users differ from those of passengers regarding the programming of autonomous vehicles. In the study of Kallioinen et al. [43], participants experienced the perspectives of passengers and pedestrians from the first-person perspective in an immersive virtual environment. The results lent support to the hypothesis that pedestrians have a self-protective preference for the passenger to be sacrificed. The study also hints at the possibility that there are moral principles that transcend these self-protective biases as both passengers and pedestrians agreed upon the utilitarian principle that the option that preserves most lives is to be preferred.

However, when interpreting these findings it is important to consider that Kallioinen et al. [43] tested the influence of perspective in an immersive environment in which the pedestrians saw the approaching car from the first-person perspective. It is thus possible to speculate that the saliency of the imminent threat for survival may have amplified self-protective tendencies in this study.

Relevant decisions about purchasing a car or about determining algorithms for dealing with accidents are often made in the absence of imminent threat. It is thus interesting to test whether the same effects can be found when people reason about abstract scenarios in which the threat to survival is less salient. Here it is relevant that Frank et al. [44] cued participants into the perspective of passengers, pedestrians, and observers when judging abstract scenarios of moral dilemma situations with autonomous vehicles. They observed self-protective biases in the sense that participants who were cued into the perspective of the passenger were more willing to sacrifice the pedestrian than participants who were cued into the perspective of the pedestrian. However, these self-protective tendencies were less pronounced than one might think. First, the majority of the participants favored sacrificing the passenger to save the pedestrian even when evaluating the scenarios from the passenger perspective, which suggests that there are limitations to the degree to which moral judgments are biased by self-protective tendencies. When the numbers of passengers and pedestrians were manipulated, the participants expressed preferences in line with the utilitarian principle that it is preferable to sacrifice one life to save many others.

Here, we revisited this issue by testing, across four experiments (Experiments 1a to 2b), people's decisions in moral dilemmas with autonomous vehicles in which people's self-protective tendencies are put against the utilitarian preference of saving the maximum number of lives. This was done by systematically manipulating the number of pedestrians on the road (Experiments 1a and 1b) and the number of passengers inside the autonomous vehicle (Experiments 2a and 2b). In each experiment, participants were randomly assigned to one of three perspectives (passenger, pedestrian, observer) and asked to indicate their preferred course of action for different accident scenarios with autonomous vehicles. To anticipate, we observed a strong and robust influence of perspective on the preferred action of autonomous vehicles in moral dilemma situations. However, even though differences among perspectives persisted, self-sacrificing tendencies dominated over self-protective tendencies when many lives could be saved by a self-sacrifice. In Experiment 3, we tested whether these self-sacrificing preferences are due to a social desirability bias by employing an indirect questioning technique [45]. The hypothesis that people's self-sacrificing preferences are due to a social desirability bias had to be rejected, which supports the validity of people's stated preference to self-sacrifice when the utilitarian principle strongly favors this option.

Experiment 1a

Method

The experiment was conducted online. It was programmed with *SoSci Survey* [46] and was made available for participation at www.soscisurvey.de. Completing the experiment took about 15 minutes. This experiment and its subsequent replications were approved by the ethics committee of the Faculty of Mathematics and Natural Sciences at Heinrich Heine University Düsseldorf and all reported studies were conducted in accordance with the Declaration of Helsinki and its later amendments. Written informed consent was obtained from all participants prior to participation in each study.

Participants. Participants were recruited on campus at Heinrich Heine University Düsseldorf and via online advertisements. As a compensation for participating, all participants

could enter a lottery to win one of three € 20 gift cards for a popular online store. Psychology students received course credit for participation. Of the participants who started the study, 62 did not complete the experiment, four were not of legal age (a requirement for being able to consent to the processing of one's data in Germany), and 26 did not respond to all items. The final sample included the data of 325 participants (248 female, 76 male, one diverse) aged between 18 and 61 years ($M = 24$, $SD = 7$). A sensitivity analysis performed with *G*Power* [47] showed that, with a total sample size of $N = 325$ participants and 15 observations per participant in the experiment, small effects of size $w = .06$ [48] could be detected at an α level of .05 with a statistical power of $1—\beta = .95$ in the model-based statistical tests (see Results section) for the overall comparison among perspectives ($df = 4$). Participants were randomly assigned to one of three perspectives—pedestrian ($n = 109$), observer ($n = 111$), or passenger ($n = 105$)—from which they were asked to evaluate the moral dilemma scenarios. More detailed information about the sample—including information about the participants' trait empathy [German version of the Interpersonal Reactivity Index; 49], affinity for technology [usage of, and opinion on, electronic devices; TA-EG; 50], and acceptance of autonomous vehicles [based on the questionnaire of 51, 52]—are available at the Open Science Framework (OSF) project page (https://osf.io/4xhz7/).

Material and procedure. Participants were first provided with a definition of autonomous vehicles. Autonomous vehicles were defined as self-driving cars capable of participating in traffic on their own without the need of human intervention or back-up. Furthermore, participants were asked to adopt the perspective of a pedestrian, an observer, or a passenger (between-subjects factor). Two example dilemmas were described in detail. In the experiment proper, different moral dilemma scenarios were presented in random order, one at a time. Each scenario comprised an autonomous vehicle driving down a single-lane road with one or more pedestrians and an obstacle (such as a boulder) on the road ahead (see Fig 1 for an example). Participants were instructed that the vehicle could not come to a stop in time and an accident was inevitable. Only two options remained: The autonomous vehicle would collide either with the obstacle—killing the passenger—or with the pedestrian/s—killing them in the

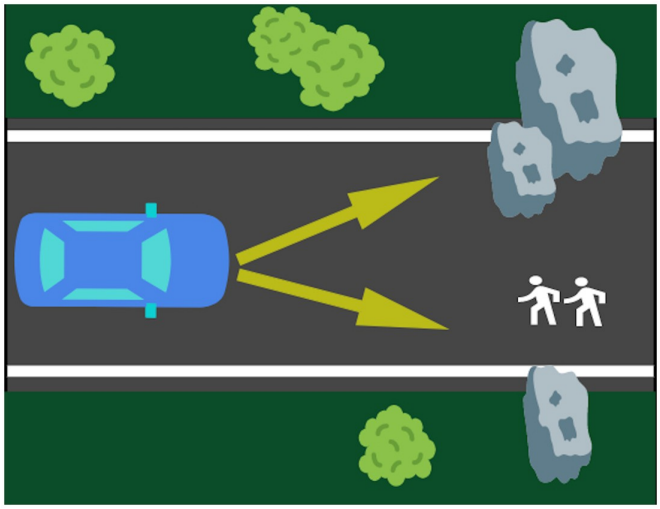


Fig 1. Example of an illustration of an accident scenario. In this example, the passenger-to-pedestrian ratio is 1:2, which means that the life of one passenger is weighed against that of two pedestrians. The visual illustrations of the scenarios were created using Microsoft PowerPoint.

https://doi.org/10.1371/journal.pone.0261673.g001

process. The scenarios were depicted as abstract sketches from a bird's eye view and showed the vehicle as well as the pedestrians and obstacles in its path. The two options available to the autonomous vehicle were illustrated with arrows. In each scenario either one, two, five, or ten pedestrians were on the road. The different numbers of pedestrians were presented in four different environments, yielding 16 different scenarios in total. The position of the vehicle (right or left side of the image) and of the pedestrians (upper or lower half of the road) was counterbalanced for each combination of number of pedestrians and environment. The experiment thus employed a 3 (perspective: pedestrian, observer, passenger; between-subjects factor) × 4 (passenger-to-pedestrian ratio: 1:1, 1:2, 1:5, 1:10; within-subjects factor) design.

Immediately below the image of the scenario, a short reminder of the respective perspective was given ("You are the/a pedestrian/observer/passenger."). Then, participants were asked: "How should the autonomous vehicle act in your opinion?" Participants had to choose whether it should "sacrifice the pedestrian/s" or "sacrifice the passenger".

The scenario and the question were presented for a maximum of 15 seconds. If participants failed to answer the question in that time span, the next scenario was automatically presented. Data sets of participants failing to evaluate all scenarios were marked as incomplete and were excluded from analysis.

Results

We used *multiTree* [53] to estimate the preferences for sacrificing the passenger for each passenger-to-pedestrian ratio and each perspective based on the observed answer frequencies. To maintain consistency in the analysis with Experiment 3, we used the simple model depicted in Fig 2 to estimate the participants' preference—in terms of a probability between 0 and 1—to sacrifice the passenger as a function of the perspective (pedestrian, observer, passenger) and the passenger-to-pedestrian ratio (1:1, 1:2, 1:5, 1:10). Participants' preferences are shown in Fig 3. Due to technical difficulties with the display of one scenario, three (instead of four) responses were analyzed for the passenger-to-pedestrian ratio of 1:5.

Fig 3 suggests that the preference for sacrificing the passenger increases with an increasing number of pedestrians that can be saved by this action. The results also suggest that the preference to sacrifice the passenger differs as a function of perspective. Participants who had adopted the perspective of a pedestrian showed the strongest preference for sacrificing the passenger while participants who had adopted the perspective of a passenger showed the lowest preference for sacrificing the passenger at all levels of the passenger-to-pedestrian-ratio variable. We used *multiTree* [53] to compare the preferences among conditions. The α level for these analyses was set to .05 and Bonferroni-Holm adjusted [54]. Confirming the visual

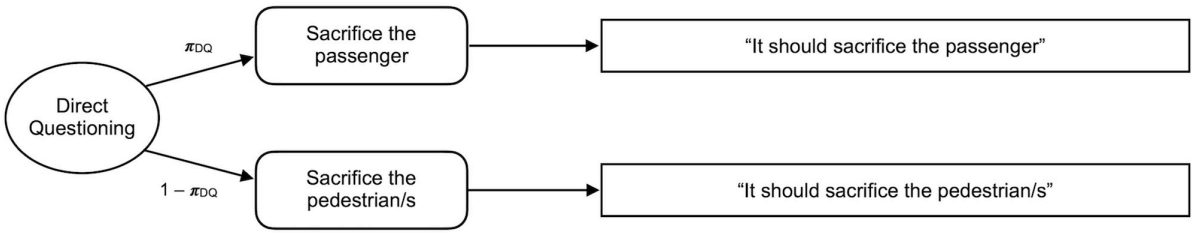


Fig 2. The multinomial processing tree model used in Experiments 1a and 1b. The rectangles on the right represent the answer categories available in each condition. Parameter π_{DQ} represents the parameter estimate for the preference that the autonomous vehicle should sacrifice the passenger instead of the pedestrian/s. Separate model trees were necessary for each combination of the 3 (perspective: pedestrian, observer, passenger) × 4 (passenger-to-pedestrian ratio: 1:1, 1:2, 1:5, 1:10) design. Note that the model corresponds to the model representing the direct questioning approach in Experiment 3.

https://doi.org/10.1371/journal.pone.0261673.g002

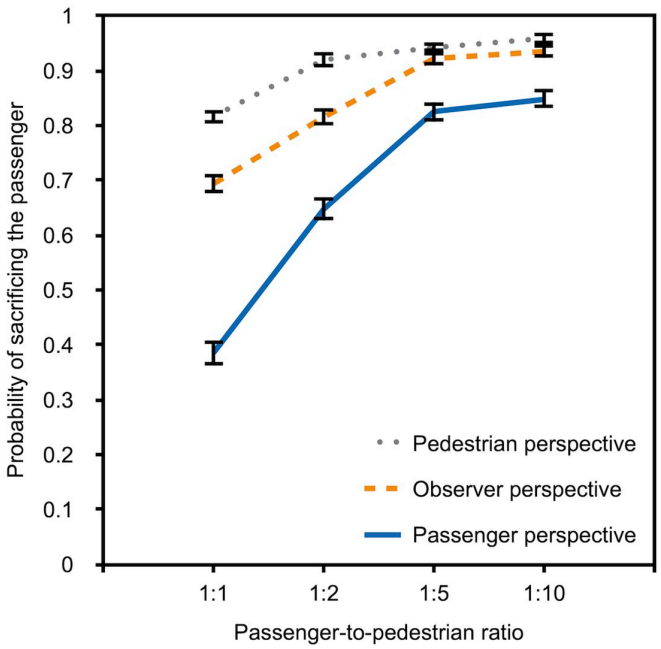


Fig 3. Descriptive data for Experiment 1a. The probability of sacrificing the passenger rather than the pedestrian/s is depicted as a function of passenger-to-pedestrian ratio (1:1, 1:2, 1:5, and 1:10) and perspective (pedestrian, observer, and passenger). The error bars represent bootstrapped standard errors.

https://doi.org/10.1371/journal.pone.0261673.g003

impression from Fig 3, the multinomial analysis confirmed that the preferences of pedestrians differed significantly from those of the passengers, $G^2(4) = 326.60$, $p < .001$, $w = .26$. The preferences of pedestrians, $G^2(4) = 43.01$, $p < .001$, $w = .09$, and passengers, $G^2(4) = 146.58$, $p < .001$, $w = .17$, differed from those of observers.

Next, we compared the preferences for sacrificing the passenger among the three perspectives at each level of the passenger-to-pedestrian-ratio variable. Again, the α level for these analyses was set to .05 and Bonferroni-Holm adjusted [54]. The test statistics are reported in Table 1. All pairwise comparisons are significant with the exception of the comparisons between pedestrians and observers at the passenger-to-pedestrian ratios of 1:5 and 1:10, which

Table 1. Comparisons among perspectives separately for each passenger-to-pedestrian ratio in Experiment 1a.

	1:1	1:2	1:5	1:10
Pedestrian vs. passenger	$G^2(1) = 172.80$	$G^2(1) = 99.65$	$G^2(1) = 22.12$	$G^2(1) = 32.04$
	$p < .001^*$	$p < .001^*$	$p < .001^*$	$p < .001^*$
	$w = .19$	$w = .14$	$w = .07$	$w = .08$
Pedestrian vs. observer	$G^2(1) = 18.09$	$G^2(1) = 21.35$	$G^2(1) = 1.04$	$G^2(1) = 2.54$
	$p < .001^*$	$p < .001^*$	$p = .308$	$p = .111$
	$w = .06$	$w = .07$	$w = .01$	$w = .02$
Observer vs. passenger	$G^2(1) = 83.85$	$G^2(1) = 31.36$	$G^2(1) = 14.02$	$G^2(1) = 17.35$
	$p < .001^*$	$p < .001^*$	$p < .001^*$	$p < .001^*$
	$w = .13$	$w = .08$	$w = .05$	$w = .06$

The α level was set to .05 and Bonferroni-Holm adjusted [54]. Significant comparisons are indicated by an asterisk.

https://doi.org/10.1371/journal.pone.0261673.t001

is probably simply due to the fact that an overwhelming majority of the pedestrians and observers preferred the utilitarian option of sacrificing the passenger in order to save the lives of five or more pedestrians.

Discussion

Experiment 1a confirms that preferences about the actions of autonomous vehicles in moral dilemmas strongly depend on perspective. Participants who evaluated the scenarios from the perspective of the pedestrian consistently displayed the highest preference for sacrificing the passenger while participants who were cued into the perspective of the passenger displayed the lowest preference for sacrificing the passenger, confirming the existence of self-protective tendencies in both pedestrians and passengers. With an increasing number of pedestrians who could be saved by sacrificing the passenger, the preference for sacrificing the passenger increased in all groups, suggesting a utilitarian preference for sacrificing one life to save several others. However, it seems noticeable that differences among the perspectives were not completely eliminated even at the most extreme passenger-to-pedestrian ratios (with the exception of pedestrians and observers who agreed that five and more pedestrians should be saved at the sacrifice of one passenger), suggesting that the utilitarian preference for saving a maximum number of lives does not completely eliminate the self-protective bias.

Given the current discussion about the robustness of psychological findings [55], we deemed it necessary to replicate the findings before drawing firm conclusions. To test the robustness of the findings, Experiment 1b served as a close replication of Experiment 1a, with the main difference to Experiment 1a being that participants were recruited from an online research panel.

Experiment 1b

Method

Participants. Participants were recruited from the online research panel of respondi AG based in Cologne, Germany. Participants received a small monetary compensation for participating in the study. Of the participants who started the study, 30 did not complete the experiment, four indicated that they had insufficient German language skills or were unable to properly read the text on the screen, 42 did not respond to all items, and five were excluded because they gave identical answers to all items of the three questionnaires at the end of the study and thus seemed to have "clicked through" the experiment. The final sample included the data of 365 participants (172 female, 193 male) aged between 18 and 69 years (M = 49, SD = 14). With this sample size, effects of size w = .06 could be detected at an α level of .05 with a statistical power of $1—\beta$ = .95 in the overall comparison among perspectives (df = 4). As in Experiment 1a, participants were randomly assigned to one of three perspectives—pedestrian (n = 118), observer (n = 124), or passenger (n = 123)—from which they were asked to evaluate the moral dilemma scenarios. Additional information about the sample is reported at the OSF project page (https://osf.io/4xhz7/).

Material and procedure. Material and procedure were identical to those of Experiment 1a.

Results

The results were analyzed as in Experiment 1a. The participants' preferences are shown in Fig 4. Due to technical difficulties with the display of one scenario, three (instead of four) responses were analyzed for the passenger-to-pedestrian ratio of 1:5.

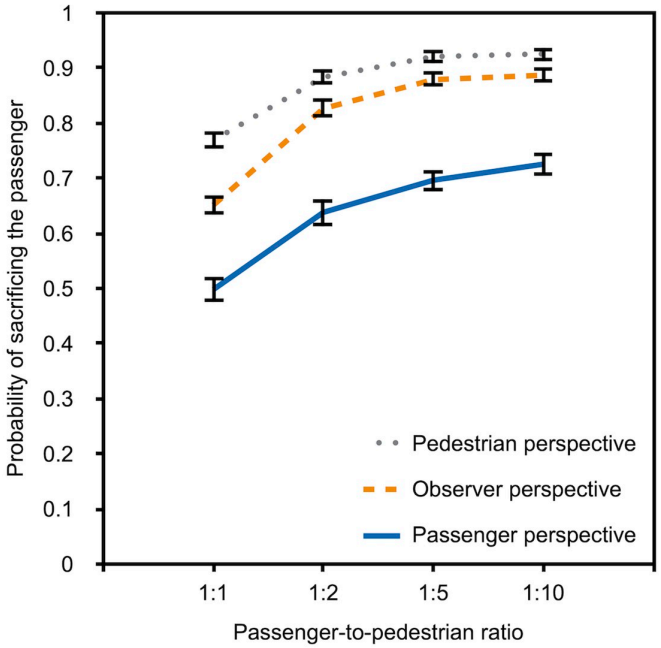


Fig 4. Descriptive data for Experiment 1b. The probability of sacrificing the passenger rather than the pedestrian/s is depicted as a function of passenger-to-pedestrian ratio (1:1, 1:2, 1:5, and 1:10) and perspective (pedestrian, observer, and passenger). The error bars represent bootstrapped standard errors.

https://doi.org/10.1371/journal.pone.0261673.g004

The results displayed in Fig 4 suggest that the preference for sacrificing the passenger increases with an increasing number of pedestrians that can be saved by this action. The results also suggest that the preference to sacrifice the passenger differs as a function of perspective. Participants who had adopted the perspective of a pedestrian showed the strongest preference for sacrificing the passenger while participants who had adopted the perspective of a passenger showed the lowest preference for sacrificing the passenger at all levels of the passenger-to-pedestrian ratio variable. Confirming the visual impression from Fig 4, the multinomial analysis confirmed that the preferences of pedestrians differed significantly from those of passengers, $G^2(4) = 292.34$, $p < .001$, $w = .23$. The preferences of pedestrians, $G^2(4) = 30.58$, $p < .001$, $w = .07$, and passengers, $G^2(4) = 149.58$, $p < .001$, $w = .17$, differed from those of observers.

Next, we compared the preferences for sacrificing the passenger among the perspectives at each level of the passenger-to-pedestrian ratio variable (Table 2). All pairwise comparisons are significant with the exception of the comparisons between pedestrians and observers at the passenger-to-pedestrian ratios of 1:5 and 1:10, which is probably due to the fact that an overwhelming majority of the pedestrians and observers preferred the utilitarian option of sacrificing the passenger in order to save the lives of five or more pedestrians. The passengers' preferences differed from those of pedestrians and observers even at these extreme passenger-to-pedestrian ratios. These findings replicate those obtained in Experiment 1a.

Discussion

The results of Experiment 1b replicate the main findings of Experiment 1a, suggesting that these findings are robust. Most importantly, the participants' preferences for the action of autonomous vehicles in moral dilemmas is determined by their perspective. Participants who

Table 2. Comparisons among perspectives separately for each passenger-to-pedestrian ratio in Experiment 1b.

	1:1	1:2	1:5	1:10
Pedestrian vs. passenger	$G^2(1) = 77.55$	$G^2(1) = 82.49$	$G^2(1) = 61.84$	$G^2(1) = 70.46$
	$p < .001^*$	$p < .001^*$	$p < .001^*$	$p < .001^*$
	$w = .12$	$w = .12$	$w = .11$	$w = .11$
Pedestrian vs. observer	$G^2(1) = 16.39$	$G^2(1) = 6.34$	$G^2(1) = 3.55$	$G^2(1) = 4.31$
	$p < .001^*$	$p = .012^*$	$p = .060$	$p = .038$
	$w = .05$	$w = .03$	$w = .03$	$w = .03$
Observer vs. passenger	$G^2(1) = 23.84$	$G^2(1) = 45.49$	$G^2(1) = 37.93$	$G^2(1) = 42.32$
	$p < .001^*$	$p < .001^*$	$p < .001^*$	$p < .001^*$
	$w = .07$	$w = .09$	$w = .08$	$w = .09$

The α level was set to .05 and Bonferroni-Holm adjusted [54]. Significant comparisons are indicated by an asterisk.

https://doi.org/10.1371/journal.pone.0261673.t002

evaluated the scenarios from the perspective of a pedestrian had a stronger preference for sacrificing the passenger to save the pedestrian/s than participants who were cued into the perspective of the passenger. Even though differences between the pedestrian perspective and the passenger perspective were obtained at the most extreme passenger-to-pedestrian ratios, the results hint at the possibility that some degree of consensus can be reached about the preferred action of the autonomous vehicle as a majority of the participants agreed that the passenger should be sacrificed to save the lives of two or more pedestrians. Even a majority of the participants who were cued into the perspective of a passenger showed this utilitarian preference to save a maximum number of lives.

However, based on the results of Experiments 1a and 1b, we do not yet know whether pedestrians show a complementary preference for sacrificing a pedestrian to save the lives of several passengers of an autonomous vehicle. It is worth repeating here that the Ethics Commission of the German Federal Ministry of Transport and Digital Infrastructure [39] has granted a special status to road users outside of autonomous vehicles as they have argued that those who do not generate a mobility risk such as pedestrians must never be sacrificed to save those generating that risk such as passengers of an autonomous vehicle. If laypeople share the same moral believes, it is not certain that participants who are cued into the perspective of pedestrians will show an increasing preference for sacrificing a pedestrian to save the lives of two or more passengers of an autonomous vehicle. Instead, they may show a persistent preference to spare the pedestrian regardless of the number of passengers that could be saved by taking a different course of action. To test this hypothesis, we manipulated the number of passengers who could be saved by sacrificing a pedestrian in Experiments 2a and 2b.

Experiment 2a

Method

Participants. Participants were recruited and compensated as in Experiment 1a. Only participants who did not participate in Experiment 1a were allowed to participate. Of the participants who started the study, 50 did not complete the experiment, six were not of legal age, and 34 were excluded because they did not respond to all items. The final sample included the data of 312 participants (232 female, 80 male), aged between 18 and 63 years ($M = 25$, $SD = 8$). With this sample size and 16 evaluations, effects of size $w = .06$ could be detected at an α level of .05 with a statistical power of $1—\beta = .95$ in the overall comparison among perspectives ($df = 4$). As in the previous experiments, participants were randomly assigned to one of three

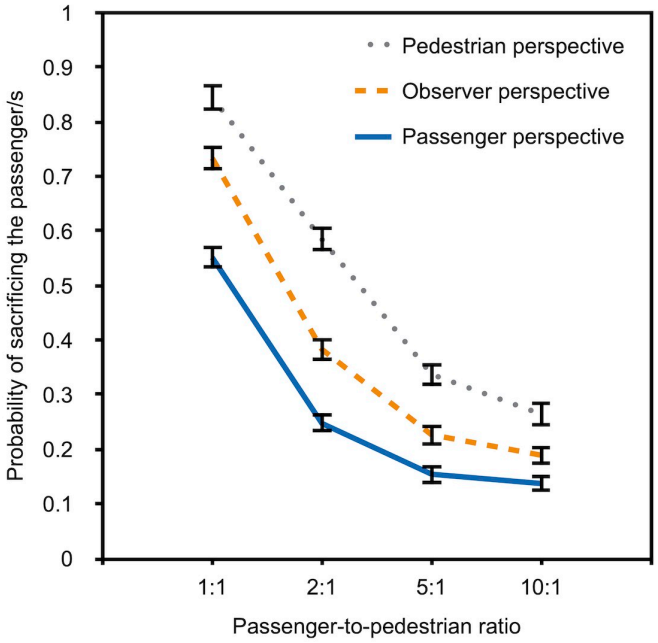


Fig 5. Descriptive data for Experiment 2a. The probability of sacrificing the passenger/s rather than the pedestrian is depicted as a function of passenger-to-pedestrian ratio (1:1, 2:1, 5:1, and 10:1) and perspective (pedestrian, observer, and passenger). The error bars represent bootstrapped standard errors.

https://doi.org/10.1371/journal.pone.0261673.g005

perspectives—pedestrian (n = 103), observer (n = 107), or passenger (n = 102)—from which they were asked to evaluate the moral dilemma scenarios. Additional information about the sample is reported at the OSF project page (https://osf.io/4xhz7/).

Material and procedure. Material and procedure were identical to those of Experiment 1a with the following exception. Instead of varying the number of pedestrians on the road, we now manipulated the number of passengers of the autonomous vehicle (within-subjects factor). Accordingly, there were one, two, five, or ten passengers inside the vehicle, but there was only one pedestrian. Thus, the experiment employed a 3 (perspective: pedestrian, observer, passenger; between-subjects factor) × 4 (passenger-to-pedestrian ratio: 1:1, 2:1, 5:1, 10:1; within-subjects factor) design.

Results

Results were analyzed in the same way as in the previous experiments. The participants' preferences are shown in Fig 5.

The first thing that seems noticeable is that, just as in the previous experiments, most participants prefer to sacrifice the passenger rather than the pedestrian when the passenger-to-pedestrian ratio is 1:1. However, there is an increasingly strong preference reversal with an increasing number of passengers whose lives can be saved by crashing into the pedestrian. As in the previous experiments, there were strong self-protective biases in the participants' preferences. Participants who had adopted the perspective of a pedestrian showed the strongest preference for sacrificing the passenger while participants who had adopted the perspective of the passenger showed the lowest preference for sacrificing the passenger. Confirming the visual impression from Fig 5, the multinomial analysis confirmed that the preferences of pedestrians

differed significantly from those of the passengers, $G^2(4) = 243.30$, $p < .001$, $w = .22$. The preferences of pedestrians, $G^2(4) = 69.74$, $p < .001$, $w = .12$, and passengers, $G^2(4) = 59.58$, $p < .001$, $w = .11$, differed from those of observers.

Next, we compared the preferences for sacrificing the passenger among the perspectives at each level of the passenger-to-pedestrian ratio variable (Table 3). All pairwise comparisons are significant.

Discussion

As in the previous experiments, most participants preferred to sacrifice the passenger rather than the pedestrian when the life of a passenger had to be weighed against the life of a pedestrian (passenger-to-pedestrian ratio 1:1). This mirrors the conviction expressed by the official guidelines of the Ethics Commission of the German Federal Ministry of Transport and Digital Infrastructure [39] that the life of a pedestrian should not be sacrificed to save the passenger of the autonomous vehicle. However, laypeople's moral intuition assessed in Experiment 2a are much less rigid than the recommendations of the Ethics Commission. With an increasing number of passengers whose lives can be saved by sacrificing the pedestrian, preferences shift towards sacrificing the pedestrian. Even those participants who were cued into the perspective of a pedestrian show a utilitarian preference for sacrificing the pedestrian to save that of several passengers. Nevertheless, the results also confirm that differences among the perspectives are not completely eliminated even at the most extreme passenger-to-pedestrian ratios, showing a strong influence of self-protective biases on moral decision making. Again, we thought it desirable to test the robustness of these findings by performing a close replication with participants who were recruited from an online research panel.

Experiment 2b

Method

Participants. Participants were recruited and compensated as in Experiment 1b. None of the participants had participated in Experiment 1b. Of the participants who started the study, 36 did not complete the experiment, two were excluded because they indicated that they had insufficient German language skills to understand the instructions, 43 did not respond to all items, and 10 gave identical answers to all items of the three questionnaires. The final sample included the data of 388 participants (180 female, 208 male), aged between 19 and 69 years ($M = 48$, $SD = 13$). With this sample size and 16 evaluations, effects of size $w = .05$ could be

Table 3. Comparisons among perspectives separately for each passenger-to-pedestrian ratio in Experiment 2a.

	1:1	2:1	5:1	10:1
Pedestrian vs. passenger	$G^2(1) = 86.38$	$G^2(1) = 98.24$	$G^2(1) = 37.70$	$G^2(1) = 20.98$
	$p < .001^*$	$p < .001^*$	$p < .001^*$	$p < .001^*$
	$w = .13$	$w = .14$	$w = .09$	$w = .06$
Pedestrian vs. observer	$G^2(1) = 15.68$	$G^2(1) = 34.46$	$G^2(1) = 12.79$	$G^2(1) = 6.82$
	$p < .001^*$	$p < .001^*$	$p < .001^*$	$p = .009^*$
	$w = .06$	$w = .08$	$w = .05$	$w = .04$
Observer vs. passenger	$G^2(1) = 30.46$	$G^2(1) = 17.88$	$G^2(1) = 7.09$	$G^2(1) = 4.14$
	$p < .001^*$	$p < .001^*$	$p = .008^*$	$p = .042^*$
	$w = .08$	$w = .06$	$w = .04$	$w = .03$

The α level was set to .05 and Bonferroni-Holm adjusted [54]. Significant comparisons are indicated by an asterisk.

https://doi.org/10.1371/journal.pone.0261673.t003

detected at an α level of .05 with a statistical power of 1—β = .95 in the overall comparison among perspectives ($df = 4$). As in the previous experiments, participants were randomly assigned to one of three perspectives—pedestrian ($n = 133$), observer ($n = 123$), or passenger ($n = 132$)—from which they were asked to evaluate the moral dilemma scenarios. Additional information about the sample is reported at the OSF project page (https://osf.io/4xhz7/).

Material and procedure. Material and procedure were identical to those of Experiment 2a.

Results

The results were analyzed in the same way as in the previous experiments. The participants' preferences are shown in Fig 6.

The results show that the majority of the participants (with the exception of those who were cued into the perspective of a passenger) had a preference for sacrificing the passenger when the passenger-to-pedestrian ratio is 1:1. However, the preference for sacrificing the passenger/s decreases with an increasing number of passengers whose lives can be saved by crashing into the pedestrian. Just as in the previous experiments, the preference to sacrifice the passenger differed as a function of perspective. Participants who had adopted the perspective of the pedestrian showed a much stronger preference for sacrificing the passenger to save the pedestrians than those who were cued into the perspective of a passenger, $G^2(4) = 422.06$, $p < .001$, $w = .26$. The preferences of pedestrians, $G^2(4) = 372.76$, $p < .001$, $w = .25$, and passengers, $G^2(4) = 13.89$, $p = .008$, $w = .05$, differed from that of observers.

Next, we compared the preferences for sacrificing the passenger among the perspectives at each level of the passenger-to-pedestrian ratio (Table 4). All pairwise comparisons are significant with the exception of the comparisons between observers and passengers at the passenger-to-pedestrian ratios of 2:1, 5:1, and 10:1. This could be attributed to the utilitarian preference of observers to minimize harm and save a maximum number of lives.

Discussion

Overall, there is a high degree of consistency across all four experiments. Preferences about the action of an autonomous vehicle in a moral dilemma scenario strongly depends on perspective. Participants who evaluated the scenario from the perspective of the pedestrian consistently displayed the highest preference for sacrificing the passenger while participants who were cued into the perspective of the passenger showed the lowest preference for sacrificing the passenger. This suggests that these preferences are strongly affected by self-protective biases. However, there is some degree of agreement among all of the perspectives. With an increasing number of lives that can be saved through sacrificing either the passenger or the pedestrian, preferences for the utilitarian option of saving a greater number of lives increases. This implies that a considerable proportion of people are willing to self-sacrifice when a large number of people can be saved by such a selfless act.

However, a possible caveat is that these people may have chosen to self-sacrifice based on a social desirability bias [cf. 56]. That is, participants may openly indicate to favor the utilitarian option of sacrificing themselves to save the lives of many others because they want to avoid the embarrassment of being perceived as selfish by choosing the self-protective option. There is evidence that sacrificing someone else for one's own good is seen as morally less acceptable than self-sacrificing [57]. It thus seems possible to speculate that some subset of participants may have chosen the self-sacrificing option only to present themselves in a favorable light. In other words, people may respond in line with what they perceive to be a moral norm instead of admitting to their self-protective preferences. If this were the case, then the results of

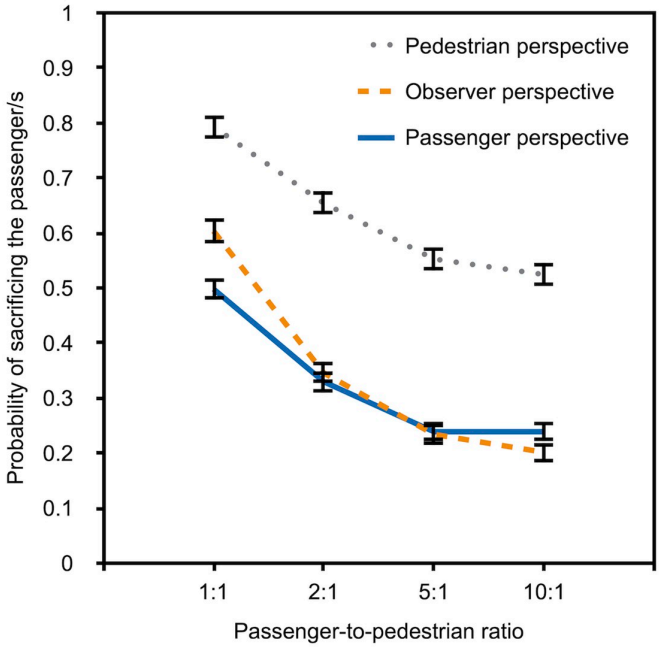


Fig 6. Descriptive data for Experiment 2b. The probability of sacrificing the passenger/s rather than the pedestrian is depicted as a function of passenger-to-pedestrian ratio (1:1, 2:1, 5:1, and 10:1) and perspective (pedestrian, observer, and passenger). The error bars represent bootstrapped standard errors.

https://doi.org/10.1371/journal.pone.0261673.g006

Experiments 1a to 2b would have overestimated the preference for the socially desirable utilitarian option and underestimated the role of socially undesirable self-protective preferences.

Fortunately, there are questioning techniques that address the issue of social desirability. Indirect questioning techniques, such as the Randomized Response Technique [58], guarantee respondents confidentiality to counteract social desirability bias [for a more detailed introduction to indirect questioning techniques see e.g., 59]. The underlying idea is to add obvious random noise to the data so that it is not possible to determine, at an individual level, what answer the respondent gave to the sensitive question which assesses the attribute potentially affected by social desirability. In consequence, the influence of social desirability on responding is

Table 4. Comparisons among perspectives separately for each passenger-to-pedestrian ratio in Experiment 2b.

	1:1	2:1	5:1	10:1
Pedestrian vs. passenger	$G^2(1) = 103.26$	$G^2(1) = 113.76$	$G^2(1) = 111.60$	$G^2(1) = 93.44$
	$p < .001^*$	$p < .001^*$	$p < .001^*$	$p < .001^*$
	$w = .13$	$w = .14$	$w = .13$	$w = .12$
Pedestrian vs. observer	$G^2(1) = 44.28$	$G^2(1) = 98.99$	$G^2(1) = 111.18$	$G^2(1) = 118.31$
	$p < .001^*$	$p < .001^*$	$p < .001^*$	$p < .001^*$
	$w = .08$	$w = .13$	$w = .13$	$w = .14$
Observer vs. passenger	$G^2(1) = 11.49$	$G^2(1) = 0.29$	$G^2(1) = 0.03$	$G^2(1) = 2.08$
	$p = .001^*$	$p = .590$	$p = .854$	$p = .149$
	$w = .04$	$w = .01$	$w < .01$	$w = .02$

The α level was set to .05 and Bonferroni-Holm adjusted [54]. Significant comparisons are indicated by an asterisk.

https://doi.org/10.1371/journal.pone.0261673.t004

reduced and the corresponding prevalence estimates are considered more valid compared to conventional direct questioning approaches [60]. To illustrate, participants may be presented with the statements "I have never driven under the influence of alcohol" and "I have driven under the influence of alcohol". Unobserved by the experimenter, they then roll a dice to determine whether they respond to the first or the second statement with "yes" or "no". Given that the interviewer does not know to which statement the answer belongs, one can assume that participants are more willing to answer truthfully. However, provided that the randomization probability is known, the prevalence of the sensitive attribute can be determined at the group level. Since the Randomized Response Technique has been proposed, indirect questioning techniques have been improved to address limitations of the original method such as relying on an external randomization device. The Crosswise Model [61] requires participants answer two questions or evaluate two statements at once. One of these refers to the sensitive attribute (e.g., "I have driven under the influence of alcohol") that is to be assessed while the other refers to an attribute with known prevalence. For example, the second statement may be "I was born in November or December". The probability of a "yes" response to the second statement can be estimated from official birth statistics. Participants then have to choose between the options "I agree with both statements or with neither statement" and "I agree with only one of the statements (irrespective of which one)". The Crosswise Model is mathematically identical to the Randomized Response Technique but it has the advantage that it does not require an external randomization device (as the non-sensitive statement is used for adding random noise to the data). Another advantage of this procedure is that it does not offer participants a "safe" response option such as "no". It is also easier to understand than other indirect questioning techniques [62]. In line with the assumption that the increased confidentiality of responding reduces the influence of socially desirability, the Crosswise Model leads to higher estimates of socially undesirable attitudes, preferences, and behaviors such as tax evasion [63], plagiarism in student papers [64], distrust [65], prejudice against women leaders [66], xenophobia, and islamophobia [67]. What is more, the Crosswise Model leads to more accurate estimates of cheating behavior whose prevalence is known [68]. In the present study, we will rely on the Extended Crosswise Model [45]. This extension of the Crosswise Model [61] has the additional advantage that one can detect whether participants systematically deviate from the instructions (e.g., by misunderstanding the instructions or by responding carelessly) and thus allows to test the validity of the data without a loss in efficiency. This model has been successfully validated [45] and was favorably evaluated in a recent experimental application [69].

If the participants' answers in response to moral dilemmas with autonomous vehicles that involve self- and other-sacrifices were biased by socially desirable responding, the indirect questioning approach should yield higher approval for the sacrificing of several other people to save one's own life than the direct questioning approach. In consequence, the approval for the socially desirable option to self-sacrifice should decrease. To illustrate, in a study conducted in an early phase of the COVID-19 pandemic on the compliance with the precautionary measures against infections with the SARS-CoV-2 virus [70], 94.5% of the participants claimed to wash their hands regularly and sufficiently long with soap and water in response to a direct question but the indirect questioning approach yielded a significantly smaller prevalence estimate of 78.1%. By comparing estimates that are based on the Extended Crosswise Model [45] and a direct question, it is possible to test whether, and to what degree, direct self-reports are contaminated by social desirability. To simplify the analysis, participants were asked to evaluate only one scenario in Experiment 3. A passenger-to-pedestrian ratio of 5:1 was selected because previous evidence suggests that a group size of five represents a switching point. In a study of Faulhaber et al. [18], the participants' willingness to self-sacrifice in order to save others increased when the number of lives that could be saved by the selfless act

increased from one to five but it did not increase further beyond this point. The data of Experiments 1a to 2b reported here also indicate only small changes in the willingness to sacrifice the passenger between a group of five and a group of 10 (cf. Figs 3–6). There thus seems to be a comparatively strong utilitarian norm to self-sacrifice in order to save five other lives. If the preference for this utilitarian norm to save the lives of others is partly or fully caused by social desirability bias, the preference to self-sacrifice should be decreased in the indirect questioning condition in comparison to the direct self-reports.

Experiment 3

Method

Participants. Participants were recruited and compensated as in Experiments 1b and 2b. Of the participants who started the study, 79 did not complete the experiment, 13 were not of legal age, indicated that they were unable to properly read the text presented on the screen or that they had insufficient German language skills, and another two data sets were excluded because of double participation. The final sample consisted of $N = 1{,}380$ participants (621 female, 756 male, 3 diverse) aged between 18 and 99 years ($M = 55$, $SD = 13$). With this sample size, effects of $w = .10$ [that is, a small effect according to 48] could be detected at an α level of .05 with a statistical power of $1-\beta = .95$ in the comparison of the preference estimates between the direct questioning approach and the indirect questioning approach ($df = 1$). Participants were randomly assigned to one of three experimental groups (between-subjects factor; see explanation below): Direct Questioning Group ($n = 459$), Indirect Questioning Group 1 ($n = 461$), and Indirect Questioning Group 2 ($n = 460$).

Materials and procedure. All participants were asked to adopt the perspective of the pedestrian. They saw only one scenario in which the life of one pedestrian had to be weighed against the lives of five passengers inside the autonomous vehicle. The image showed an autonomous vehicle on a single-lane road heading towards a roadblock and a single pedestrian on the road from a bird's eye view (Fig 7). In a written account of the incident, participants were informed that the accident would inevitably result in the death of either the passengers or the pedestrian.

Below the image and the vignette, participants were asked to evaluate the depicted scenario from the perspective of the pedestrian. Participants in the Direct Questioning Group were asked to evaluate the sensitive statement "In the depicted situation, the autonomous vehicle should sacrifice the five passengers to save me as a pedestrian" with "True" or "False".

Participants in the indirect questioning groups received instructions on the indirect questioning technique explaining to them that it was impossible to infer, at an individual level, their true opinion on the matter from the answers that were given. As in previous studies [e.g., 65, 66, 71], the month of birth was used as the non-sensitive attribute in the two indirect questioning groups. Participants in the indirect questioning groups also knew that we did not know their month of birth and that they would not be asked about it. Accordingly, when presented with the scenario following the instructions, participants in Indirect Questioning Group 1 were asked to evaluate the sensitive statement together with the non-sensitive statement "I was born in November or December" by choosing from the answer categories "I either agree with both statements or with neither statement" and "I agree with only one statement (irrespective of which one)". Participants in Indirect Questioning Group 2 were provided with the same answer categories, but the non-sensitive statement was replaced by the complementary non-sensitive statement "I was born between January and October".

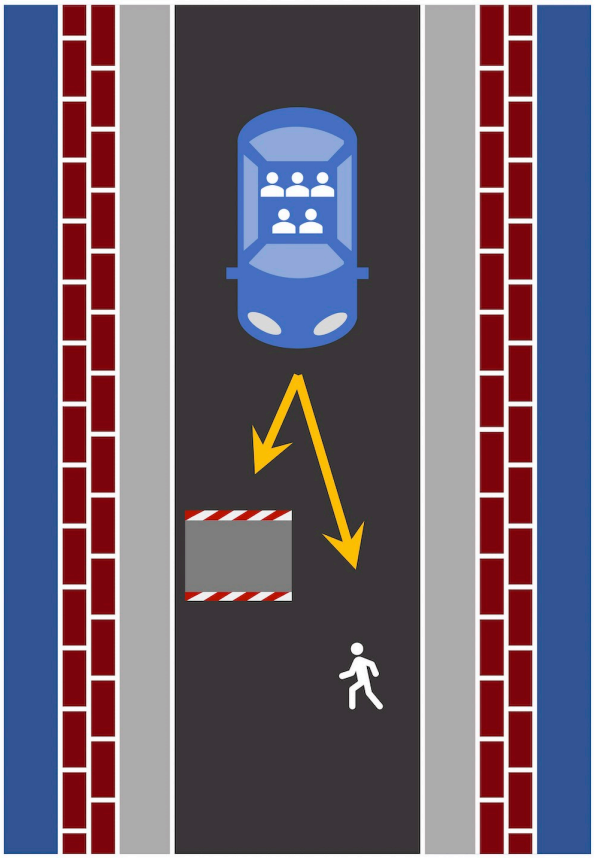


Fig 7. The illustration that was used for the moral dilemma scenario. The passenger-to-pedestrian ratio was 5:1, which implies that the life of five passengers was weighed against that of one pedestrian. The visual illustration of the scenario were created using Microsoft PowerPoint.

https://doi.org/10.1371/journal.pone.0261673.g007

The experiment thus employed a group design with three experimental groups (Direct Questioning Group, Indirect Questioning Groups 1 and 2). In total, participation in the experiment took about 5 minutes.

Results

As in the previous experiments, we used *multiTree* [53] to estimate the preference for sacrificing the passengers based on the observed answer frequencies and to compare these preferences among the groups. The Extended Crosswise Model [45] as used here is shown in Fig 8.

In the Direct Questioning Group (see upper tree in Fig 8), the prevalence π_{DQ} of the sensitive attribute (preference for sacrificing the five passengers to save the pedestrian) corresponds directly to the probability that the answer category "True" was obtained. Note that the upper tree corresponds to the way in which the parameters were obtained in the previous experiments (Fig 2). Obtaining the prevalence estimates for the sensitive attribute in the indirect questioning groups (lower two trees in Fig 8) is somewhat more complex as participants' true status on the assessed attributes cannot be directly inferred from the provided answers. Parameters π_{IQ1} and π_{IQ2} represent prevalence estimates of the sensitive attribute. Parameter $p_{Nov\text{-}Dec}$

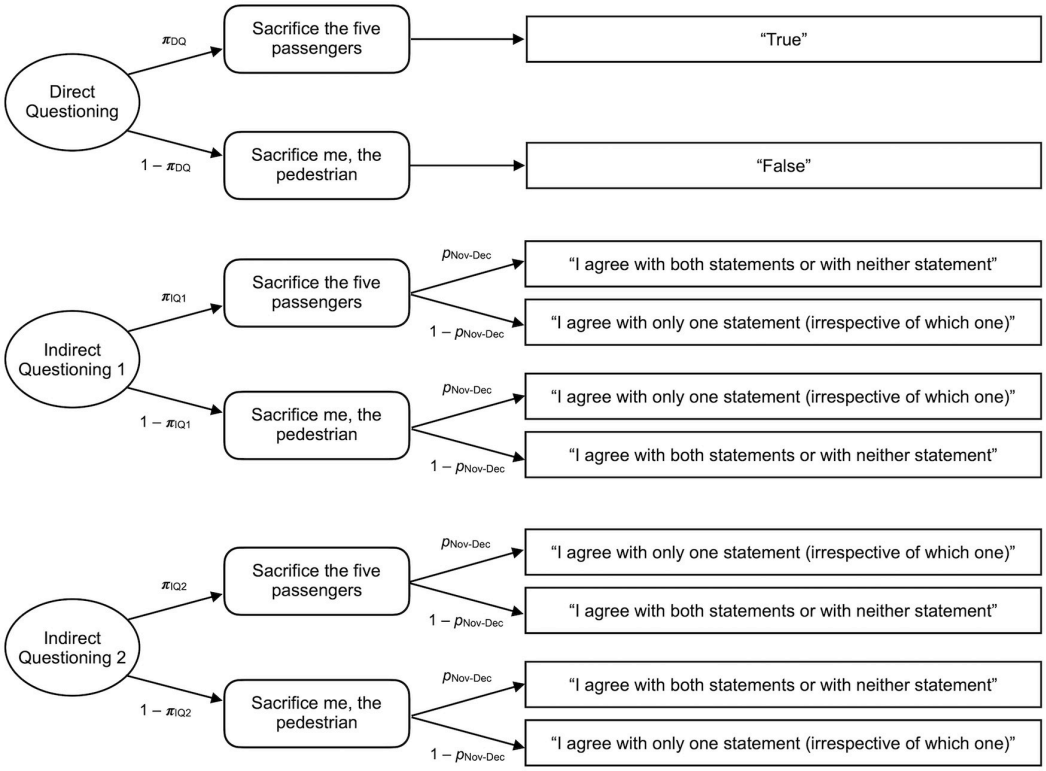


Fig 8. Multinomial processing tree model. The combined multinomial processing tree model for the Direct Questioning Group—represented by the upper tree—and for Indirect Questioning Groups 1 and 2—represented by the lower two trees—for the Extended Crosswise Model [45] adapted to the present experiment. The rectangles on the right contain the answer categories available in each condition. Parameter π represents the prevalence estimates for the preferences that the autonomous vehicle should sacrifice five passengers of the autonomous vehicle in order to save the pedestrian, depending on the condition. Parameter $p_{\text{Nov-Dec}}$ represents the known prevalence of the non-sensitive attribute, in this case, the participant being born in November or December.

https://doi.org/10.1371/journal.pone.0261673.g008

represents the known prevalence of being born in November or December. The respective prevalence can be derived from official birth statistics. According to the German birth statistics, the probability of being born in November or December is approximately 15.8% [72]. Hence, we set parameter $p_{\text{Nov-Dec}}$ to .158 in the following analyses. To obtain a prevalence estimate of the month of birth in the present sample, participants in the Direct Questioning Group were asked to indicate whether they were born between January and October. The statistical conclusions do not change when the prevalence estimate is based on the sample prevalence estimate of $p_{\text{Nov-Dec}} = .176$. Participants in the two indirect questioning groups evaluated the sensitive and the non-sensitive statement simultaneously. The only difference between the two indirect questioning groups was the non-sensitive statement. In Indirect Questioning Group 1 the non-sensitive statement was "I was born in November or December" ($p_{\text{Nov-Dec}}$), in the Indirect Questioning Group 2 it was "I was born between January and October" ($1 - p_{\text{Nov-Dec}}$). The answer categories depicted in Fig 8 are therefore swapped for Indirect Questioning Group 2 ("Indirect Questioning 2" in Fig 8) in comparison to Indirect Questioning Group 1 ("Indirect Questioning 1" in Fig 8).

As noted earlier, the Extended Crosswise Model allows to test whether participants follow the instructions. Specifically, the prevalence estimates for the sensitive attribute must not differ

between the two indirect questioning groups when participants follow the instructions. An analysis of the Extended Crosswise Model thus starts with equating the prevalence estimates of the two indirect questioning groups. All subsequent model-based results can only be trusted when it is possible to combine the two parameters π_{IQ1} and π_{IQ2} into a single parameter π_{IQ} representing the prevalence estimate based on indirect questioning. The model assuming that the prevalence estimates do not differ between the two groups ($\pi_{IQ1} = \pi_{IQ2}$) fitted the data well, $G^2(1) = 1.37$, $p = .243$, $w = .03$. According to Heck et al. [45], this indicates that participants adhered to the instructions and "the prevalence estimate can be considered trustworthy" (p. 1897). Therefore, the two indirect questioning groups were pooled for further analysis.

Next, we tested whether the prevalence estimates differed between the direct questioning approach and the indirect questioning approach. In the Direct Questioning Group ($n = 459$), 41.0% ($SE = 2.3$) indicated that the autonomous vehicle should sacrifice five passengers to save them while the prevalence estimate for the sensitive attribute in the Combined Indirect Questioning Group ($n = 921$) was 40.1% ($SE = 2.4$). The assumption that the prevalence estimates did not differ between the Direct Questioning Group and the Combined Indirect Questioning Group ($\pi_{DQ} = \pi_{IQ}$) was compatible with the data, $\Delta G^2(1) = 0.07$, $p = .790$, $w = .01$. This indicates that the prevalence estimates did not differ between the direct and the indirect questioning approach. In other words, the hypothesis that the prevalence estimates based on the direct questioning approach are compromised by social desirability must be rejected.

Discussion

Experiment 3 served to test whether direct self-reports of a (utilitarian) self-sacrificing preference in a moral dilemma with autonomous vehicles are compromised by social desirability. To this end, we used the Extended Crosswise Model [45] to test whether increased confidentiality of responding would decrease the approval of the self-sacrificing option. Disconfirming the hypothesis that the utilitarian preference for self-sacrifice is only due to social desirability, preference estimates did not differ between the direct and the indirect questioning. Participants expressed the preference to sacrifice themselves to save the lives of five others even when a high degree of confidentiality was guaranteed. This is all the more interesting given that the indirect questioning technique used here has been shown to reliably reveal effects of social desirability on answers to questions about sensitive topics such as prejudice against Muslims and hand hygiene [69, 70].

This indicates that people's preference for the utilitarian option of sacrificing themselves to save the lives of five other people was not, or at least not to an appreciable degree, affected by social desirability [see 33, for further evidence that the influence of social desirability on people's preferences in moral dilemmas is limited]. It also seems noticeable that the results of Experiment 3 are well aligned with the results of the previous experiments. There is an overall preference for the autonomous vehicle to save a maximum number of lives even if this means sacrificing oneself. Nevertheless, a substantial proportion of participants (about 40%) prefer the self-protective option even if this means to kill five other people.

General discussion

Automated vehicle technologies promise benefits such as improved accessibility of transportation, and increased traffic safety [e.g., 1, 5]. Yet, before autonomous vehicles can be implemented on a large scale, several challenges need to be addressed—besides the technical implementation—for example issues regarding ensuring the safety of road users and passengers as well as software security, developing the legal requirements, and creating the necessary infrastructure [e.g., 2]. A hotly debated topic is how autonomous vehicles should handle

accident situations [e.g., 10, 12–14, 18, 35, 73] and whether, and to what degree, people prefer actions of autonomous vehicles that are biased to save their own lives (as passengers or pedestrians) at the cost of those of others [e.g., 18, 35, 38]. These self-protective biases may clash with those of other road users, leading to potential conflicts that may slow or complicate the introduction of autonomous vehicles. To investigate to what extend preferences of non-motorized road users regarding moral dilemmas involving autonomous vehicles may differ from those of passengers, we compared the preferred action of an autonomous vehicle from the perspectives of a passenger, a pedestrian, and an observer in moral dilemma scenarios involving a varying number of potential victims. Our results suggest that perspective strongly determines the preferred course of action. Specifically, people cued into the perspective of passengers consistently expressed the least preference for sacrificing the passenger/s of the autonomous vehicle while pedestrians consistently expressed the highest preference for sacrificing the passenger/s.

Scenarios commonly employed to investigate moral dilemmas with autonomous vehicles often feature passengers and pedestrians [e.g., 18, 35, 40, 73] but, as yet, only few studies [43, 44] have required participants to evaluate the scenarios from the perspective of the pedestrians. The results strongly indicate that evaluations from the pedestrian perspective are important as pedestrians and passengers evaluate moral dilemmas with autonomous vehicles very differently. This is because the pedestrians, just like the passengers, display clear self-protective tendencies.

Given that passengers and pedestrians differ in their preferences for how autonomous vehicles should handle accident situations, the question arises of how the conflicting positions can be reconciled. Even though the present results show pervasive self-protective biases across all experiments, the results also suggest that it might be possible to reach some degree of agreement among the perspectives. A majority of those participants who were cued into the perspective of an uninvolved observer preferred protecting the pedestrian when the passenger-to-pedestrian ratio was 1:1 but preferred the utilitarian option of sparing maximum lives in all other conditions. With an increasing number of lives at stake, more and more participants preferred the utilitarian option of sparing a maximum number of lives even at the cost of sacrificing their own lives. This seems to imply that not all people want to save their own life at all cost. This was true both for passengers (Experiments 1a and 1b) and pedestrians (Experiments 2a and 2b). This suggests that, contrary to official guidelines [39], pedestrians may be willing to accept some degree of risk caused by autonomous vehicles.

At first sight it seemed possible to assume that this self-sacrificing tendency could be attributed to social desirability bias. However, this hypothesis has to be rejected given the results of Experiment 3. Even when an indirect questioning technique [45] guaranteed confidentiality of responding, the majority of the participants (about 60%) expressed the preference for sacrificing themselves to save the five passengers inside the autonomous vehicle and this majority was equally large when participants were questioned directly. The results suggest that the participants' preference for a self-sacrifice to save the lives of several others is not only due to social desirability bias. Instead, it seems that they were privately convinced that the utilitarian option is the right course of action. This suggests that the preferences of passengers of autonomous vehicles and other road users can, to some degree, be reconciled with each other despite the persistent self-protective tendencies. More knowledge may be gained about how the differences between perspectives can be reconciled by examining the degree to which people's preferences in moral dilemmas change depending on the degree to which it is emphasized that the same person might take different roles in traffic. This approach resembles the so-called veil-of-ignorance reasoning employed, for example, by Huang et al. [74]. In their study, participants were asked which option they would prefer in a moral dilemma if they did not know who

among the affected parties they would be. Participants who engaged in this type of veil-of-ignorance reasoning displayed a higher preference for the utilitarian option in response to a subsequently presented dilemma than participants in a control condition. Thus, encouraging participants to consider that the role one takes in traffic varies may provide a means to reduce self-protective tendencies.

A limitation of the present research is that participants were asked to evaluate abstract scenarios. The decisions may thus be representative of situations such as when contemplating to purchase an autonomous vehicle in which people are able to make judgments about moral dilemmas without the imminent threat or stress of a real-life accident. It is unclear whether the preferences that were identified here generalize to decisions that are made in more extreme situations when life and death are a matter of seconds. Here it seems relevant that the present results are largely consistent with those of Kallioinen et al. [43] who manipulated the perspective from which an imminent accident was observed in an immersive virtual environment. Participants who experienced the scenario from a passenger perspective were less willing to put themselves at risk by guiding the autonomous vehicle off a cliff than participants who viewed the scene from the perspective of a pedestrian. It seems noticeable that the self-protection bias was limited in the study of Kallioinen et al. even though they used an immersive methodology in which the accident was experienced first-hand. In their first study, a conflict between the passenger and the pedestrian emerged only in a specific scenario in which serious harm to the passenger seemed likely. Possibly, scenarios that imply a clear self-sacrifice provide a higher potential for strong disagreement between the involved parties [25] than scenarios with more ambiguous consequences. Together, the results suggest that the self-protection bias is a pervasive cognitive bias that affects moral decision making both when being immersed in a critical traffic situation and when reasoning about abstract moral dilemmas.

Another limitation of the present study is that there is some culture-specific variation in moral preferences [13] so that it cannot be taken for granted that the findings reported here generalize across different samples. As a first step for testing the robustness of the present findings, we tested whether the results of Experiments 1a and 2a that were obtained with student samples (mostly young adults with little driving experience) could be replicated in Experiments 1b and 2b with samples from online research panels (adults with higher driving experience and more heterogeneous age and education). The fact that most of the results of the student samples could be replicated in the online samples is encouraging, as is the fact that the present results are largely consistent with those obtained in other labs in Denmark and Germany as well as international and US online samples [43, 44]. Nevertheless, most of the studies focused on well-educated Western samples so that examining the degree to which the self-protective and self-sacrificing preferences generalize to other samples is an interesting avenue for further research. Larger and more diverse samples than those used in the present study would be necessary to test how the self-protection bias is affected by potentially moderating factors such as gender, age, and personality.

In conclusion, the studies presented here aim at contributing to the discussion surrounding moral dilemmas involving autonomous vehicles. The perspective from which participants evaluated moral dilemma scenarios strongly affected the preferred action of the autonomous vehicle in the respective scenario. Specifically, passengers and pedestrians differed in their preferences from each other, but also from uninvolved observers, which suggests that self-protective biases have a strong influence on the evaluation of moral dilemmas involving autonomous vehicles. As a consequence of these conflicting interests, focusing on only one perspective may be problematic for the acceptance of autonomous vehicles in the long run. To guarantee widespread social acceptance, which is necessary for the success of autonomous vehicles [e.g., 13, 25, 32, 35], a careful balancing of the conflicting interests of the involved

perspectives might be required. The present results suggest that some degree of consensus can be reached among the different perspectives. Regardless of the perspective, many participants preferred the utilitarian option of saving a maximum number of lives, even when the utilitarian option implied a self-sacrifice. Although differences among the perspectives did not completely vanish even when utilitarian principles clearly favored one of the available options, the majority of the participants who were cued into the perspective of the passenger agreed that the passenger should be sacrificed to save the lives of a group of pedestrians. Similarly, a majority of the participants who were cued into the role of the pedestrian agreed that the pedestrian should be sacrificed to save the lives of several passengers inside the autonomous vehicle. There is no evidence that the utilitarian preference for a self-sacrifice is caused by social desirability as participants expressed this preference even in an indirect questioning format that is known to reveal effects of social desirability. The results therefore suggest that, despite prevailing self-protective tendencies, there are some moral principles that all road users can agree upon.

Author Contributions

Conceptualization: Maike M. Mayer, Raoul Bell, Axel Buchner.

Formal analysis: Maike M. Mayer, Raoul Bell, Axel Buchner.

Investigation: Maike M. Mayer.

Methodology: Maike M. Mayer, Raoul Bell, Axel Buchner.

Supervision: Raoul Bell, Axel Buchner.

Writing – original draft: Maike M. Mayer.

Writing – review & editing: Raoul Bell, Axel Buchner.

References

1. Bagloee SA, Tavana M, Asadi M, Oliver T. Autonomous vehicles: Challenges, opportunities, and future implications for transportation policies. J Mod Transport. 2016; 24(4):284–303. https://doi.org/10.1007/s40534-016-0117-3
2. European Commission. GEAR 2030: High Level Group on the competitiveness and sustainable growth of the automotive industry in the European Union–Final report 2017. European Union, European Commission; 2017 Oct. https://www.europarl.europa.eu/cmsdata/141562/GEAR%202030%20Final%20Report.pdf.
3. European Commission. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee, the Committee of Regions–On the road to automated mobility: An EU strategy for mobility of the future. Brussels: European Commission; 2018 May. https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A52018DC0283
4. National Science & Technology Council, & United States Department of Transportation. Ensuring American leadership in automated vehicle technologies: Automated vehicles 4.0. Washington, DC: United States Department of Transportation; 2020 Jan. https://www.transportation.gov/sites/dot.gov/files/2020-02/EnsuringAmericanLeadershipAVTech4.pdf
5. Anderson JM, Nidhi K, Stanley KD, Sorensen P, Samaras C, Oluwatola OA. Autonomous vehicle technology: A guide for policymakers. Revised ed. Santa Monica (CA): RAND Corporation; 2016.
6. World Health Organization. Global status report on road safety 2018. Geneva: World Health Organization; 2018 Jun. https://www.who.int/publications/i/item/global-status-report-on-road-safety-2018
7. National Center for Statistics and Analysis. 2018 fatal motor vehicle crashes: Overview [Traffic Safety Facts Research Note]. Washington, DC: United States Department of Transportation, National Highway Traffic Safety Administration; 2019 Oct. Report No.: DOT HS 812 826. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812826.

8. European Commission [Internet]. Road safety: Europe's roads are getting safer but progress remains too slow. Brussels: European Commission; 2020 [cited 2021 May 3]. https://ec.europa.eu/commission/presscorner/detail/en/IP_20_1003.
9. National Highway Traffic Safety Administration. National Motor Vehicle Crash Causation Survey—Report to Congress. Washington, DC: United States Department of Transportation, National Highway Traffic Safety Administration; 2008 Jul. Report No.: DOT HS 811 059. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/811059.
10. Bonnefon J-F, Shariff A, Rahwan I. The trolley, the bull bar, and why engineers should care about the ethics of autonomous cars [point of view]. Proceedings of the IEEE. 2019; 107(3):502–4. https://doi.org/10.1109/JPROC.2019.2897447
11. Goodall NJ. Ethical decision making during automated vehicle crashes. Transp Res Rec. 2014; 2424 (1):58–65. https://doi.org/10.3141/2424-07
12. Lin P. Why ethics matters for autonomous cars. In: Maurer M, Gerdes JC, Lenz B, Winner H, editors. Autonomous driving. Berlin Heidelberg: Springer; 2016. pp. 69–85.
13. Awad E, Dsouza S, Kim R, Schulz J, Henrich J, Shariff A, et al. The Moral Machine experiment. Nature. 2018; 563(7729):59–64. https://doi.org/10.1038/s41586-018-0637-6 PMID: 30356211
14. Nyholm S. The ethics of crashes with self-driving cars: A roadmap, I. Philos Compass. 2018; 13(7): e12507. https://doi.org/10.1111/phc3.12507
15. Goodall NJ. Can you program ethics into a self-driving car? IEEE Spectr. 2016; 53(6):28–58. https://doi.org/10.1109/MSPEC.2016.7473149
16. Goodall NJ. Away from trolley problems and toward risk management. Appl Artif Intell. 2016; 30 (8):810–21. https://doi.org/10.1080/08839514.2016.1229922
17. Shariff A, Bonnefon J-F, Rahwan I. Psychological roadblocks to the adoption of self-driving vehicles. Nat Hum Behav. 2017; 1(10):694–6. https://doi.org/10.1038/s41562-017-0202-6 PMID: 31024097
18. Faulhaber AK, Dittmer A, Blind F, Wächter MA, Timm S, Sütfeld LR, et al. Human decisions in moral dilemmas are largely described by utilitarianism: Virtual car driving study provides guidelines for autonomous driving vehicles. Sci Eng Ethics. 2019; 25:399–418. https://doi.org/10.1007/s11948-018-0020-x PMID: 29357047
19. Sütfeld LR, Gast R, König P, Pipa G. Using virtual reality to assess ethical decisions in road traffic scenarios: Applicability of value-of-life-based models and influences of time pressure. Front Behav Neurosci. 2017; 11:122. https://doi.org/10.3389/fnbeh.2017.00122 PMID: 28725188
20. Greene JD. Our driverless dilemma. Science. 2016; 352(6293):1514–5. https://doi.org/10.1126/science.aaf9534 PMID: 27339966
21. Greene JD, Nystrom LE, Engell AD, Darley JM, Cohen JD. The neural bases of cognitive conflict and control in moral judgment. Neuron. 2004; 44(2):389–400. https://doi.org/10.1016/j.neuron.2004.09.027 PMID: 15473975
22. Gawronski B, Armstrong J, Conway P, Friesdorf R, Hütter M. Consequences, norms, and generalized inaction in moral dilemmas: The CNI model of moral decision-making. J Pers Soc Psychol. 2017; 113 (3):343–76. https://doi.org/10.1037/pspa0000086 PMID: 28816493
23. Goodall NJ. Machine ethics and automated vehicles. In: Meyer G, Beiker S, editors. Road vehicle automation. Lecture Notes in Mobility. 1st ed. Cham: Springer; 2014. pp. 93–102.
24. Thornton SM, Pan S, Erlien SM, Gerdes JC. Incorporating ethical considerations into automated vehicle control. IEEE trans Intell Transp Syst. 2017; 18(6):1429–39. https://doi.org/10.1109/TITS.2016.2609339
25. Bergmann LT, Schlicht L, Meixner C, König P, Pipa G, Boshammer S, et al. Autonomous vehicles require socio-political acceptance—An empirical and philosophical perspective on the problem of moral decision making. Front Behav Neurosci. 2018; 12:31. https://doi.org/10.3389/fnbeh.2018.00031 PMID: 29541023
26. Kant I. Grundlegung zur Metaphysik der Sitten Riga: bey Johann Friedrich Hartknoch; 1785. http://db.saur.de/DLO/saveUrl.jsf?type=document&documentId=BDL01887_0001&volumeId=BDL01887_0001. Document No.: BDL01887_0001. Acess required. German.
27. Bentham J. An introduction to the principles of morals and legislation. London: T. Payne and son; 1789. http://galenet.galegroup.com/servlet/MOME?af=RN&ae=U102143420&srchtp=a&ste=14.
28. Mill JS. Utilitarianism / Der Utilitarismus. Birnbacher D, editor. Stuttgart: Reclam; 1871/2010.
29. Foot P. The problem of abortion and the doctrine of double effect. Oxford Reviews. 1967; 5:5–15.
30. Thomson JJ. Killing, letting die, and the trolley problem. Monist. 1976; 59(2):204–17. https://doi.org/10.5840/monist197659224 PMID: 11662247
31. Thomson JJ. The trolley problem. Yale Law J. 1985; 94(6):1395–415. https://doi.org/10.2307/796133

32. Wolkenstein AJE. What has the Trolley Dilemma ever done for us (and what will it do in the future)? On some recent debates about the ethics of self-driving cars. Ethics Inf Technol. 2018; 20(3):163–73. https://doi.org/10.1007/s10676-018-9456-6

33. Sütfeld LR, Ehinger BV, König P, Pipa G. How does the method change what we measure? Comparing virtual reality and text-based surveys for the assessment of moral decisions in traffic dilemmas. PLoS One. 2019; 14(10):e022310. https://doi.org/10.1371/journal.pone.0223108 PMID: 31596864

34. Keeling G. Why trolley problems matter for the ethics of automated vehicles. Sci Eng Ethics. 2020; 26:293–307. https://doi.org/10.1007/s11948-019-00096-1 PMID: 30830593

35. Bonnefon J-F, Shariff A, Rahwan I. The social dilemma of autonomous vehicles. Science. 2016; 352 (6293):1573–6. https://doi.org/10.1126/science.aaf2654 PMID: 27339987

36. Volz LJ, Welborn BL, Gobel MS, Gazzaniga MS, Grafton ST. Harm to self outweighs benefit to others in moral decision making. Proc Natl Acad Sci U S A. 2017; 114(30):7963–8. https://doi.org/10.1073/pnas.1706693114 PMID: 28696302

37. Liu P, Liu J. Selfish or utilitarian automated vehicles? Deontological evaluation and public acceptance. Int J Hum Comput Interact. 2021; 37(13):1231–42. https://doi.org/10.1080/10447318.2021.1876357

38. Gogoll J, Müller JF. Autonomous cars: In favor of a mandatory ethics setting. Sci Eng Ethics. 2017; 23 (3):681–700. https://doi.org/10.1007/s11948-016-9806-x PMID: 27417644

39. Federal Ministry of Transport and Digital Infrastructure. Ethics Commission Automated and Connected Driving. Report (extract). Federal Ministry of Transport and Digital Infrastructure; 2017 Jun. https://www.bmvi.de/SharedDocs/EN/publications/report-ethics-commission-automated-and-connected-driving.pdf?__blob=publicationFile

40. Li J, Zhao X, Cho M-J, Ju W, Malle BF. From trolley to autonomous vehicle: Perceptions of responsibility and moral norms in traffic accidents with self-driving cars. Society of Automotive Engineers World Congress, 2016 Apr 12–14; Detroit, MI, USA: SAE Technical Paper 2016-01-0164; 2016.

41. Pugnetti C, Schläpfer R. Customer preferences and implicit tradeoffs in accident scenarios for self-driving vehicle algorithms. J Risk Financial Manag. 2018; 11(2):28. https://doi.org/10.3390/jrfm11020028

42. Nobis C, Kuhnimhof T. Mobilität in Deutschland–MiD Ergebnisbericht. Studie von infas, DLR, IVT und infas 360 im Auftrag des Bundesministers für Verkehr und digitale Infrastruktur [cited 2021 May 3]. Version 1.1 Bonn: infas Institut für angewandte Sozialwissenschaft GmbH; 2019 Feb. FE-Project No.: 70.904/15. Commissioned by the Federal Ministry of Transport and Digital Infrastructure. http://www.mobilitaet-in-deutschland.de/pdf/MiD2017_Ergebnisbericht.pdf. German.

43. Kallioinen N, Pershina M, Zeiser J, Nosrat Nezami F, Stephan A, Pipa G, et al. Moral judgements on the actions of self-driving cars and human drivers in dilemma situations from different perspectives. Front Psychol. 2019; 10:2415. https://doi.org/10.3389/fpsyg.2019.02415 PMID: 31749736

44. Frank D-A, Chrysochou P, Mitkidis P, Ariely D. Human decision-making biases in the moral dilemmas of autonomous vehicles. Sci Rep. 2019; 9:13080. https://doi.org/10.1038/s41598-019-49411-7 PMID: 31511560

45. Heck DW, Hoffmann A, Moshagen M. Detecting nonadherence without loss in efficiency: A simple extension of the crosswise model. Behav Res Methods. 2018; 50(5):1895–905. https://doi.org/10.3758/s13428-017-0957-8 PMID: 28916924

46. Leiner DJ. SoSci Survey [software]. SoSci Survey GmbH; 2019. https://www.soscisurvey.de

47. Faul F, Erdfelder E, Lang A-G, Buchner A. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods. 2007; 39(2):175–91. https://doi.org/10.3758/bf03193146 PMID: 17695343

48. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hilsdale (NJ): Lawrence Earlbaum Associates; 1988.

49. Paulus C. Der Saarbrücker Persönlichkeitsfragebogen SPF (IRI) zur Messung von Empathie: Psychometrische Evaluation der deutschen Version des Interpersonal Reactivity Index. 2009. http://bildungswissenschaften.uni-saarland.de/personal/paulus/empathy/SPF_Artikel.pdf. German.

50. Karrer K, Glaser C, Clemens C, Bruder C. Technikaffinität erfassen–Der Fragebogen TA-EG In: Lichtenstein A, Stößel C, Clemens C, editors. Der Mensch im Mittelpunkt technischer Systemes. 8. Berliner Werkstatt Mensch-Maschine-Systeme (Vol. 29). Düsseldorf: VDI Verlag GmbH; 2009. pp. 196–201. German.

51. Schoettle B, Sivak M. A survey of public opinion about autonomous and self-driving vehicles in the US, the UK, and Australia. Ann Arbor (MI): The University of Michigan, Transportation Reserach Institute; 2014 Jul. Report No.: UMTRI-2014-21. https://deepblue.lib.umich.edu/bitstream/handle/2027.42/108384/103024.pdf?sequence=1&isAllowed=y.

52. Schoettle B, Sivak M. Public opinion about self-driving vehicles in China, India, Japan, the US, the UK, and Australia. Ann Arbor (MI): The University of Michigan, Transportation Reserach Institute; 2014

Oct. Report No.: UMTRI-2014-30. https://deepblue.lib.umich.edu/bitstream/handle/2027.42/109433/103139.pdf?sequence=1.

53. Moshagen M. multiTree: A computer program for the analysis of multinomial processing tree models. Behav Res Methods. 2010; 42(1):42–54. https://doi.org/10.3758/BRM.42.1.42 PMID: 20160285

54. Holm S. A simple sequentially rejective multiple test procedure. Scand J Statist. 1979; 6(2):65–70.

55. Open Science Collaboration. Estimating the reproducibility of psychological science. Science. 2015; 349(6251):aac4716. https://doi.org/10.1126/science.aac4716 PMID: 26315443

56. Tourangeau R, Yan T. Sensitive questions in surveys. Psychol Bull. 2007; 133(5):859–83. https://doi.org/10.1037/0033-2909.133.5.859 PMID: 17723033

57. Sachdeva S, Iliev R, Ekhtiari H, Dehghani M. The role of self-sacrifice in moral dilemmas. PLoS One. 2015; 10(6):e0127409. https://doi.org/10.1371/journal.pone.0127409 PMID: 26075881

58. Warner SL. Randomized response: A survey technique for eliminating evasive answer bias. J Am Stat Assoc. 1965; 60(309):63–9. PMID: 12261830

59. Chaudhuri A, Christofides TC. Indirect questioning in sample surveys. 1st ed. Berlin Heidelberg: Springer; 2013.

60. Lensvelt-Mulders GJ, Hox JJ, Van der Heijden PG, Maas CJ. Meta-analysis of randomized response research: Thirty-five years of validation. Sociol Methods Res. 2005; 33(3):319–48. https://doi.org/10.1177/0049124104268664

61. Yu J-W, Tian G-L, Tang M-L. Two new models for survey sampling with sensitive characteristic: Design and analysis. Metrika. 2008; 67(3):251–63. https://doi.org/10.1007/s00184-007-0131-x

62. Hoffmann A, de Puiseau BW, Schmidt AF, Musch J. On the comprehensibility and perceived privacy protection of indirect questioning techniques. Behav Res Methods. 2017; 49(4):1470–83. https://doi.org/10.3758/s13428-016-0804-3 PMID: 27631988

63. Korndörfer M, Krumpal I, Schmukle SC. Measuring and explaining tax evasion: Improving self-reports using the crosswise model. J Econ Psychol. 2014; 45:18–32. https://doi.org/10.1016/j.joep.2014.08.001

64. Jann B, Jerke J, Krumpal I. Asking sensitive questions using the crosswise model: An experimental survey measuring plagiarism. Public Opin Q. 2012; 76(1):32–49. https://doi.org/10.1093/poq/nfr036

65. Thielmann I, Heck DW, Hilbig BE. Anonymity and incentives: An investigation of techniques to reduce socially desirable responding in the Trust Game. Judgm Decis Mak. 2016; 11(5):527–36.

66. Hoffmann A, Musch J. Prejudice against women leaders: Insights from an indirect questioning approach. Sex Roles. 2019; 80(11–12):681–92. https://doi.org/10.1007/s11199-018-0969-6

67. Hoffmann A, Musch J. Assessing the validity of two indirect questioning techniques: A Stochastic Lie Detector versus the Crosswise Model. Behav Res Methods. 2016; 48(3):1032–46. https://doi.org/10.3758/s13428-015-0628-6 PMID: 26182857

68. Hoffmann A, Diedenhofen B, Verschuere B, Musch J. A strong validation of the Crosswise Model using experimentally-induced cheating behavior. Exp Psychol. 2015; 62(2):403–14. https://doi.org/10.1027/1618-3169/a000304 PMID: 27120562

69. Meisters J, Hoffmann A, Musch J. Controlling social desirability bias: An experimental investigation of the extended crosswise model. PLoS One. 2020; 15(12):e0243384. https://doi.org/10.1371/journal.pone.0243384 PMID: 33284820

70. Mieth L, Mayer MM, Hoffmann A, Buchner A, Bell R. Do they really wash their hands? Prevalence estimates for personal hygiene behaviour during the COVID-19 pandemic based on indirect questions. BMC Public Health. 2021; 21:12. https://doi.org/10.1186/s12889-020-10109-5 PMID: 33397344

71. Waubert de Puiseau B, Hoffmann A, Musch J. How indirect questioning techniques may promote democracy: A preelection polling experiment. Basic Appl Soc Psych. 2017; 39(4):209–17. https://doi.org/10.1080/01973533.2017.1331351

72. Pötzsch O. Geburten in Deutschland. Wiesbaden: Statistisches Bundesamt; 2012 Jan. https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bevoelkerung/Geburten/Publikationen/Downloads-Geburten/broschuere-geburten-deutschland-0120007129004.pdf?__blob=publicationFile. German.

73. Wintersberger P, Prison A-K, Riener A, Hasirlioglu S. The experience of ethics: Evaluation of self harm risks in automated vehicles. IEEE Intelligent Vehicles Symposium (IV), 2017 Jun 11–14; Los Angeles, CA, USA: IEEE; 2017. pp. 385–91.

74. Huang K, Greene JD, Bazerman M. Veil-of-ignorance reasoning favors the greater good. Proc Natl Acad Sci U S A. 2019; 116(48):23989–95. https://doi.org/10.1073/pnas.1910125116 PMID: 31719198

Men, machines, and double standards? The moral evaluation of the actions of autonomous vehicles and human drivers in road-accident dilemmas

Maike M. Mayer, Axel Buchner, and Raoul Bell

Department of Experimental Psychology, Heinrich Heine University Düsseldorf, 40225 Düsseldorf, Germany

Manuscript Type: Extended Multi-Phase Study

Word count: 6733 (including Keywords, Précis, and Key Points)

Running head: Action evaluation of different drivers

Correspondence concerning this article should be addressed to Maike M. Mayer, Department of Experimental Psychology, Heinrich Heine University Düsseldorf, Universitätsstraße 1, 40225 Düsseldorf, Germany. E-Mail: maike.mayer@hhu.de

Abstract

Objective: In two experiments we investigated how actions of autonomous vehicles, anthropomorphized autonomous vehicles, and human drivers are morally evaluated in accident scenarios.

Background: Autonomous vehicles have to interact with other road users whose behavior is difficult to predict. It is unlikely that all accidents can be avoided. A more critical evaluation of the actions of autonomous vehicles in accidents compared to those of human drivers may complicate the introduction of autonomous vehicles.

Method: Participants evaluated, from a moral perspective, the actions (sacrificing the passenger/pedestrian/s) of autonomous vehicles (Experiments 1-2), anthropomorphized autonomous vehicles (Experiment 2), and human drivers (Experiments 1-2) in otherwise identical accident scenarios with varying numbers of pedestrians.

Results: The actions of human drivers were always evaluated more morally justifiable than those of autonomous vehicles. This evaluation difference between humans and machines was reduced but not completely eliminated by anthropomorphizing the autonomous vehicle. The more lives were spared, the better the action was evaluated irrespective of the agent.

Conclusion: An evaluation bias exists in favor of the actions of human drivers which may be reduced by anthropomorphizing the autonomous vehicle. Utilitarian considerations are involved in the moral evaluation of the actions of human drivers, anthropomorphized autonomous vehicles, and autonomous vehicles.

Application: The observed differences could cause a more critical public response to accidents involving autonomous vehicles compared to those involving only human drivers and might negatively affect public acceptance of autonomous vehicles. Anthropomorphism might help to align the moral evaluation of autonomous vehicles with that of human drivers.

Keywords

autonomous agents, autonomous driving, vehicle automation, system design, driver behavior

Précis

The actions of human drivers of cars are evaluated as more morally justifiable than the corresponding actions of autonomous vehicles. Anthropomorphizing the autonomous vehicle reduces this evaluation difference to some degree. Utilitarian considerations are involved in all evaluations.

Autonomous driving promises to bring many benefits such as increased traffic accessibility for people without a driver's license and less traffic congestion, potentially resulting in less pollution and reduced energy consumption (Bagloee et al., 2016). Autonomous vehicles may also represent a new chapter in mobility-on-demand and car-sharing services that might reduce the individual and societal costs of mobility (Spieser et al., 2014). Once autonomous driving technologies will have reached an automation level that does not require humans to intervene, these technologies could increase the comfort of daily driving: Being freed of the driving task, passengers of autonomous vehicles could use the driving time for other activities (Anderson et al., 2016; Bagloee et al., 2016). Given that human error is a major cause of road accidents (National Highway Traffic Safety Administration, 2008), autonomous vehicles are also expected to increase traffic safety in the future (Anderson et al., 2016; Bagloee et al., 2016; Waldrop, 2015). However, accidents cannot be completely avoided. Apart from the fact that no technology will ever function without flaws (e.g., Gogoll & Müller, 2017; Lin, 2016), there is another reason why autonomous vehicles cannot avoid all accidents regardless of their driving performance: they share the roads with human road users whose behaviors are hard to predict (e.g., Koopman & Wagner, 2017; Lin, 2016; Nyholm, 2018).

Fatal accidents with autonomous vehicles can be expected to attract strong media attention during the first years of introducing automated driving technologies into daily traffic (Bonnefon et al., 2016; Jelinski et al., 2021; Shariff et al., 2017), for example due to the novelty of automated driving technologies. Two of the best-known examples of accidents involving vehicles with automated driving technologies are the Tesla accident in 2016 and the Uber accident in 2018. In 2016, a Tesla Model S collided with a semitrailer, resulting in the death of its driver (National Transportation Safety Board, 2017). The Tesla accident likely represents the first fatal crash involving a vehicle with automated driving technologies (Yadron & Tynan, 2016). The Uber accident in 2018—in which an Uber vehicle struck and killed a

pedestrian (National Transportation Safety Board, 2019)—might be the first fatal crash of a vehicle with automated driving technologies involving a non-motorized road user (Levin & Wong, 2018; Wakabayashi, 2018). The critical coverage of accidents in the media can negatively affect the public perception and acceptance of autonomous vehicles (Anania et al., 2018; Shariff et al., 2017). Currently, the public's opinion on automated vehicles is mixed (Becker & Axhausen, 2017). Some studies indicate prevailing positive anticipation (Winkler et al., 2019) but others show more negative than positive emotions (Hassol et al., 2019; Tennant et al., 2019). People who are skeptical about using automated driving technologies often cite an unwillingness to yield control over the driving task to the autonomous vehicle as a reason for their skeptical attitude (Smith & Anderson, 2017; Winkler et al., 2019). The prospect of machines making decisions that might harm or kill humans might contribute to the discomfort of handing over the control of driving to autonomous vehicles (Bigman & Gray, 2018; Li et al., 2016; Malle et al., 2016). This widespread discomfort with autonomous vehicles making life-and-death decisions may delay the adoption of automated driving technologies (Li et al., 2016; Malle et al., 2016).

Therefore, it is interesting to understand how people morally evaluate the actions of autonomous vehicles in comparison to actions of human drivers in fatal accident scenarios. What is considered the moral choice does not need to be identical for a human and an autonomous vehicle. However, the results of several studies suggest that people want humans and machines to make similar choices in road-accident dilemmas (e.g., Bonnefon et al., 2016; Kallioinen et al., 2019; Li et al., 2016; Young & Monroe, 2019). Most of these studies are typically modeled after the so-called Trolley dilemma (Foot, 1967; Thomson, 1976, 1985) which can be used to assess moral preferences. In the Trolley dilemma, a trolley is racing towards five people on the tracks. It is possible to divert the trolley to a side track which will, however, result in the death of an unsuspecting track worker. Is it morally permissible to

sacrifice one person to safe five? Or should the trolley continue on its path and kill five people? According to utilitarianism, sacrificing one life to save many is morally justifiable based on the principle that decisions should reduce harm and death (e.g., Bentham, 1789; Mill, 1871/2010) while deontology, which focuses on moral rights and duties (e.g., Kant, 1786/2011), may declare the same action as impermissible as it violates the duty not to kill otherwise uninvolved people as means to an end. A road-accident scenario with an autonomous vehicle fashioned after the trolley dilemma could be the following: An autonomous vehicle is about to crash into one or more pedestrian/s on the road. The only other option being left is to crash the vehicle into a road block which results in the death of the passenger of the autonomous vehicle. Even though there is some degree of variability in people's preferences for the action of the autonomous vehicle in such a moral dilemma (e.g., Awad et al., 2018), one of the most pervasive preferences that have been identified is the utilitarian preference to minimize the number of deaths that result from the accident with the autonomous vehicle (e.g. Awad et al., 2018; Kallioinen et al., 2019; Li et al., 2016; Mayer et al., 2021).

Here we are not concerned with the action that people think autonomous vehicles and human drivers *ought to choose* but in the moral *evaluation* of the actions of autonomous vehicles and human drivers in accidents that have already occurred. Whether people evaluate the actions of autonomous vehicles and human drivers differently is a two-part question: First, are the same moral principles applied to human drivers and autonomous vehicles to morally evaluate their actions? Second, is there a general bias towards evaluating the actions of human drivers less critically than those of autonomous vehicles? In several studies, Malle and colleagues used different versions of trolley-type moral dilemmas to test whether people evaluate the actions of robots, artificial intelligence decision agents, and drones differently than those of humans (Malle et al., 2019; Malle et al., 2015; Malle et al., 2016). Interestingly, the results suggest that there might be some differences in the evaluation of moral behaviors of

humans and machines. Specifically, the results of Malle et al. (2015) suggest that "robots are expected—and possibly obligated—to make utilitarian choices" (p. 122) and thus people "regarded the act of sacrificing one person in order to save four (a 'utilitarian' choice) as more permissible for a robot than for a human" (p. 122). There is also evidence indicating that people have a general tendency for blaming autonomous vehicles more harshly for their actions in road-accident scenarios than human drivers (Young & Monroe, 2019). If the latter result turns out to be a robust finding and people are more critical of the actions of autonomous vehicles than of those of human drivers, then the question arises as to whether anthropomorphizing automated vehicles (that is, assigining humanlike characteristics or properties to them; Bartneck et al., 2009; Epley et al., 2007) will help to shift the moral evaluation of such vehicles closer to the moral evaluation of human drivers. Young and Monroe (2019) found evidence suggesting that describing the decision-making process of the autonomous vehicle in mentalistic terms (i.e., ascribing thoughts and feelings to the autonomous vehicle) may reduce the differences in blame between the autonomous vehicle and a human driver, and made people's responses to the autonomous vehicle's decisions less negative. In a similar way, Malle et al. (2016) found that presenting a robot with a more human-like appearance reduced the differences in blame for the decisions of robots and humans in comparison to presenting a robot with a mechanical appearance.

The present two experiments serve to test whether there are differences in the moral evaluation of actions of autonomous vehicles and human drivers in road-accident scenarios. As a first step, we compared the moral evaluation of the actions taken by a human driver and by an autonomous vehicle. Participants were randomly assigned to either the human-driver condition or the autonomous-vehicle condition and were asked to evaluate the actions of the respective agent from a moral perspective. Participants were presented with road-accident scenarios in which the life of a passenger is weighted against the lives of one, two, or five

pedestrians. Specifically, we wanted to test the hypothesis that people have an aversion against machines making moral decisions (Bigman & Gray, 2018; Gogoll & Uhl, 2018), which translates into the hypothesis that the actions of autonomous vehicles should be evaluated as more reprehensible and less morally justifiable than those of human drivers. Furthermore, we aimed at testing the hypothesis that people evaluate utilitarian actions of machines more favorably than those of humans (Malle et al., 2015), which translates into the hypothesis that the moral evaluations of the decisions to sacrifice the passenger or the pedestrian/s should depend more on the passenger-to-pedestrian ratio when the agent is the autonomous vehicle than when the agent is the human driver. To anticipate, the results lend clear support to the hypothesis that the actions of the autonomous vehicle are evaluated as less morally justifiable and more reprehensible than those of the human driver. In Experiment 2 we tested whether this negative evaluation bias can be reduced by anthropomorphizing the autonomous vehicle by assigning a first name ("Alina") to it (e.g., Hong et al., 2020; Waytz et al., 2014). The actions of the anthropomorphized autonomous vehicle were indeed evaluated more positively than the actions of the non-anthropomorphized autonomous vehicle which provides further support of the hypothesis that the difference in the moral evaluation of the actions of human drivers and autonomous vehicles can be reduced by assigning humanlike characteristics or properties to the autonomous vehicle (Young & Monroe, 2019).

Experiment 1

Method

The experiment was conducted online using *SoSci Survey* (Leiner, 2019). In total, participation took about 10 minutes. Experiments 1 and 2 complied with the American Psychological Association Code of Ethics and the Declaration of Helsinki. Informed consent was obtained from each participant before the experiment.

Participants

The sample was recruited via online advertisements. Undergraduate Psychology students received course credit for participating; other participants could enter a lottery to win a € 20 voucher for an online store. Of the 444 participants who started the study, 79 did not complete the experiment. In addition, five participants did not meet the a-priori defined inclusion criteria (being of legal age, having sufficient German language skills, and being able to read the text on screen according to self-reports). Valid data sets of 360 participants (266 women, 94 men), aged between 18 and 80 years (M = 27, SD = 11) were included in the analyses. Participants were randomly assigned to the human-driver condition (n = 187) or the autonomous-vehicle condition (n = 173).

We conducted a sensitivity power analysis with *G*Power* (Faul et al., 2007) in which we focused on the agent variable (human driver, autonomous vehicle; between-subjects) and on the action variable (sacrifice the pedestrian/s, sacrifice the passenger; within-subject) while ignoring the number of pedestrians variable (one pedestrian, two pedestrians, five pedestrians; within-subject). Given a total sample size of N = 360, $\alpha = \beta = .05$ and assuming a correlation of $\rho = .20$ between the levels of the action variable (estimated based on related results), small effects of about $f = .15$ (Cohen, 1988) could be detected for the agent variable.

Material and Procedure

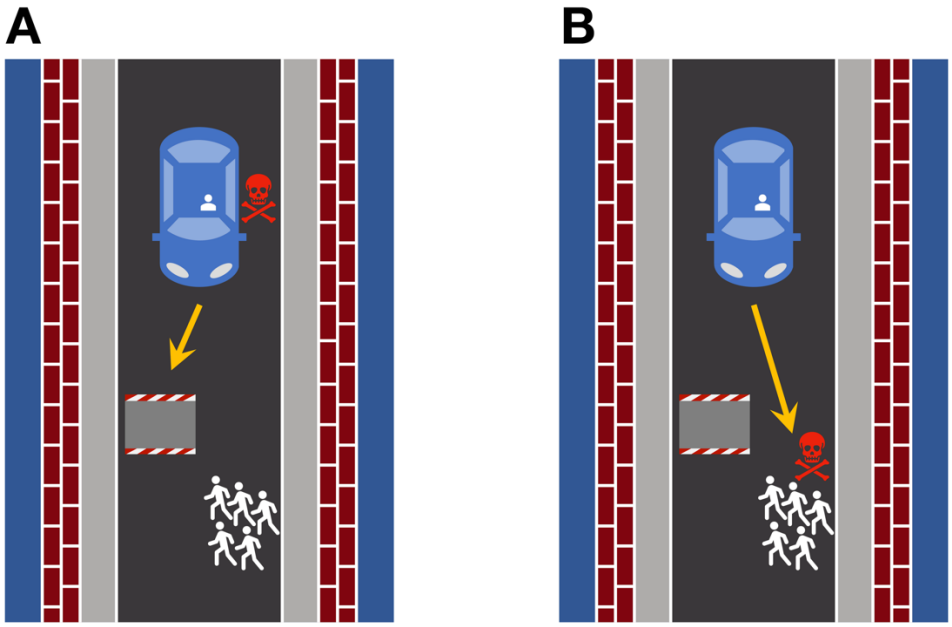
First, participants read an introductory text including the following information. Depending on the assigned condition, the text stated that human drivers or autonomous vehicles have to handle different traffic situations, including inevitable accidents. In the autonomous-vehicle condition, autonomous vehicles were defined as fully self-driving cars capable of participating in traffic without the need of human intervention. Participants were then provided

with an exemplary description of the accident scenarios they were later asked to evaluate in the experiment.

In each of the scenarios (see Figure 1 for an example), the agent (either a human driver or an autonomous vehicle) drove on a single-lane road and was suddenly confronted with an obstacle and at least one pedestrian on the road. As the agent could neither brake nor swerve, only two actions remained: The agent could either sacrifice the passenger inside the vehicle to save the pedestrian/s by crashing into the obstacle or sacrifice the pedestrian/s to save the passenger. The scenarios were depicted as abstract sketches from a bird's eye view. There were either one, two, or five pedestrians on the road. The agent had already taken one of the two available actions, represented by a yellow arrow. In each scenario, the agent either sacrificed the passenger inside the autonomous vehicle (Figure 1A) or the pedestrian/s (Figure 1B) who died because of the accident. The fatal consequence of the decision was illustrated by a red skull that was presented next to the passenger or the pedestrian/s, depending on who was sacrificed. The visual depiction of the scenario was accompanied by a text vignette describing the situation, the action taken, and the action's consequences. For example, if the autonomous vehicle sacrificed five pedestrians to save the passenger, the text stated: "The autonomous vehicle drives into the persons on the street. The person inside the vehicle remains unharmed. The five persons on the street are killed". Six different scenarios were obtained by combining two actions and three different numbers of pedestrians. The positions of the obstacle and the pedestrian/s (left or right side of the road) were counterbalanced. Altogether, four presentations of each of the six scenarios were presented, yielding 24 evaluations in total. The scenarios were presented in random order.

Figure 1

Two examples of the illustrations of the road-accident scenarios used in the experiment



Note. The images depict the two available actions for a passenger-to-pedestrian ratio of 1:5. A The passenger is sacrificed to save the five pedestrians. B The five pedestrians are sacrificed to save the passenger. The scenarios were created using Microsoft PowerPoint® and Apple Keynote®.

Below each image and the corresponding text vignette, participants were asked to evaluate the action (sacrifice the passenger vs. sacrifice the pedestrian/s) of the agent (human driver vs. autonomous vehicle) from a moral perspective. The question repeated the agent, the action, and the action's consequences for the two involved parties. For example, if the autonomous vehicle decided to sacrifice five pedestrians to save the passenger, the question was: "How do you evaluate, from a moral point-of-view, the action of the autonomous vehicle to save the person inside the vehicle and to sacrifice the five persons on the street?". Participants were asked to complete the sentence "From a moral point-of-view, I perceive the action as…" by choosing a category on a scale ranging from "very reprehensible" (1) to "very justifiable" (6). These labels were chosen based on a pilot study ($N = 16$) in which we determined

the labels that were most commonly associated with the moral evaluations of actions in the road-accident dilemmas.

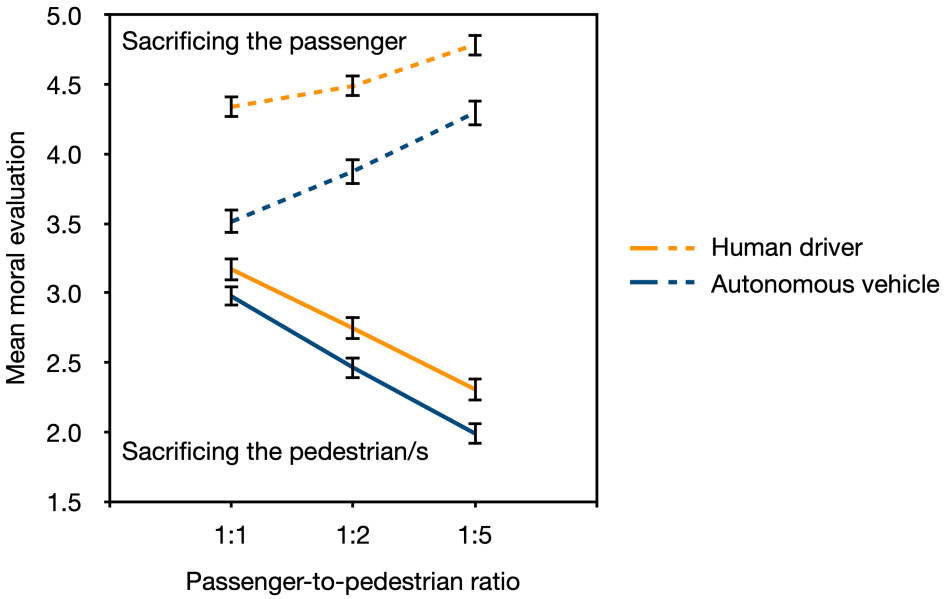
Following the evaluation of all scenarios, participants were asked to indicate which agent had acted in the previously presented scenarios ("A human driver", "An autonomous vehicle", "I do not know"). As the statistical conclusions did not change in both Experiment 1 or Experiment 2 if participants who failed the attention check were included in the statistical analysis, we decided against the exclusion of data.

Results

In our analyses, we used the MANOVA approach to repeated-measures analyses (O'Brien & Kaiser, 1985). Instead of multivariate test criteria we report the exact F statistic corresponding to Wilk's Lambda. The α level was set to .05 and all post-hoc comparisons were Bonferroni-Holm adjusted (Holm, 1979). The partial eta squared is used as an effect size measure. The descriptive data are depicted in Figure 2.

Figure 2

Descriptive data for Experiment 1



Note. The mean moral evaluation of the actions (sacrificing the passenger [dashed lines], sacrificing the pedestrian/s [solid lines]) is displayed as a function of the passenger-to-pedestrian ratio (1:1, 1:2, and 1:5) and the agent (human driver, autonomous vehicle). The moral-evaluation scale ranged from "very reprehensible" (1) to "very justifiable" (6). The error bars represent standard errors of the means.

A 2 (agent: human driver, autonomous vehicle; between-subjects) × 2 (action: sacrifice the pedestrian/s, sacrifice the passenger; within-subject) × 3 (number of pedestrians: one pedestrian, two pedestrians, five pedestrians; within-subject) MANOVA showed that the actions of the human driver were evaluated as more morally justifiable than the actions of the autonomous vehicle, $F(1,358) = 40.51$, $p < .001$, $\eta_p^2 = .10$. Sacrificing the passenger was evaluated more favorably than sacrificing the pedestrian/s, $F(1,358) = 340.82$, $p < .001$, $\eta_p^2 = .49$. The interaction between action and agent was statistically significant as well, $F(1,358) = 4.72$, $p = .030$, $\eta_p^2 = .01$. Simple main effect analyses revealed that the human

driver's actions were always evaluated more favorably than those of the autonomous vehicle, but the difference between agents was more pronounced for the decision to sacrifice the passenger ($\eta_p^2 = .08$, $p < .001$) than for the decision to sacrifice the pedestrian/s ($\eta_p^2 = .02$, $p = .016$).

In addition, there was a significant main effect of the number of pedestrians on the road, $F(2,357) = 23.94$, $p < .001$, $\eta_p^2 = .12$. The direction of this effect, however, depended on the action that was taken, $F(2,357) = 187.20$, $p < .001$, $\eta_p^2 = .51$. An increase in the number of pedestrians led to a significant increase in the moral evaluation of sacrificing the passenger (simple main effect analyses; all comparisons of the different passenger-to-pedestrian ratios: $p < .001$) while it led to a significant decrease in the moral evaluation of sacrificing the pedestrian/s (simple main effect analyses; all comparisons: $p < .001$). The effect of the number of pedestrians did not differ between agents, $F(2,357) = 2.73$, $p = .067$, $\eta_p^2 = .02$.

Finally, there was a significant three-way interaction, $F(2,357) = 4.54$, $p = .011$, $\eta_p^2 = .02$. We conducted a 2 (action: sacrifice the pedestrian/s, sacrifice the passenger; within-subject) × 3 (number of pedestrians: one pedestrian, two pedestrians, five pedestrians; within-subject) repeated-measures MANOVA for each of the two agents separately. The action of sacrificing the passenger was evaluated as significantly more morally justifiable than the action of sacrificing the pedestrian/s [human driver: $F(1,186) = 241.01$, $p < .001$, $\eta_p^2 = .56$; autonomous vehicle: $F(1,172) = 117.40$, $p < .001$, $\eta_p^2 = .41$]. There was again a significant main effect of the number of pedestrians on the road [human driver: $F(2,185) = 18.70$, $p < .001$, $\eta_p^2 = .17$; autonomous vehicle: $F(2,171) = 6.79$, $p = .001$, $\eta_p^2 = .07$] and a significant interaction between the taken action and the number of pedestrians on the road [human driver: $F(2,185) = 90.23$, $p < .001$, $\eta_p^2 = .49$; autonomous vehicle: $F(2,171) = 96.98$, $p < .001$, $\eta_p^2 = .53$]. When the passenger was sacrificed, the action was evaluated as significantly more morally justifiable with an increasing number of pedestrians on the road (simple main effect

analyses; for both agents: all comparisons $p < .001$) while the reverse pattern emerged when the pedestrian/s were sacrificed (simple main effect analyses; for both agents: all comparisons $p < .001$). The three-way interaction thus does not indicate that fundamentally different moral principles were applied to the evaluation of the actions of human drivers and autonomous vehicles, but the effect of the number of pedestrians on how sacrificing the passenger was evaluated was simply somewhat less pronounced for the human driver than for the autonomous vehicle.

Discussion

The present study served to test whether there are differences in the moral evaluation of the actions of human drivers and autonomous vehicles. Overall, the moral evaluations of the actions of human drivers and autonomous vehicles depended on both the type of action that was evaluated (sacrificing the passenger or the pedestrian/s) and the passenger-to-pedestrian ratio. Regardless of whether the actions of human drivers or autonomous vehicles were evaluated, participants regarded actions that spared the maximum number of lives as more morally justifiable. The more favorable evaluations of utilitarian actions are in line with demonstrations of overall preferences for utilitarian actions of human and machine agents in other studies (e.g., Kallioinen et al., 2019; Li et al., 2016). There was an interaction between agent and action, indicating that the decision of the human driver to self-sacrifice was evaluated more favorably than the action of the autonomous vehicle to sacrifice the passenger. Furthermore, there was a three-way interaction between agent, action, and number of pedestrians, suggesting that the evaluation of the human drivers' decisions to sacrifice themselves was less dependent on the number of pedestrians on the road than the evaluation of the autonomous vehicles' decisions to sacrifice the passenger. This finding can easily be explained by the fact that the decision of the human driver to self-sacrifice received favorable moral evaluations already when this meant spearing the live of only one pedestrian, and this favorable

evaluation was hard to boost when more pedestrians were saved at the expense of the driver. It is also worth pointing out that the sample effect sizes of these interactions are quite small (the sample effect sizes of the two-way interaction between agent and action and the three-way interaction between agent, action, and number of pedestrians were $\eta_p^2 = .01$ and $\eta_p^2 = .02$, respectively). Therefore, it seems questionable whether interactions of such small magnitude can be robustly replicated in future experiments (see Discussion of Experiment 2).

The dominant difference in the evaluation of human drivers and autonomous vehicles was that, regardless of the type of action and the number of pedestrians on the road, the actions of the human driver were always evaluated more favorably than those of the autonomous vehicle. This is in line with the hypothesis that people have an aversion against machines making moral decisions (Bigman & Gray, 2018; Gogoll & Uhl, 2018). If the actions of autonomous vehicles are evaluated more critically than those of human drivers, this difference in the moral evaluations of the actions of human drivers and autonomous vehicles could negatively affect the acceptance of autonomous driving. It thus may be desirable to search for interventions that reduce the more negative evaluations of the actions of autonomous vehicles compared to those of human drivers in fatal accidents. Anthropomorphizing the autonomous vehicle could be such an intervention. By making the autonomous vehicle more similar to humans by assigning human characteristics or properties to it (Bartneck et al., 2009; Epley et al., 2007), the moral evaluation of the actions of the autonomous vehicle might become more similar to those of the human driver. There is, in fact, some evidence to suggest that anthropomorphizing a non-human agent leads to a more positive perception or evaluation, resulting in higher trustworthiness of or trust in the respective technology compared to a more mechanistic one (e.g., Gong, 2008; Lee et al., 2015; Niu et al., 2018; Pak et al., 2012). Specifically, Young and Monroe (2019) found evidence suggesting that describing the driving algorithm of an autonomous vehicle in mentalistic terms (thoughts and feelings) may reduce the blame

assigned for an accident in comparison to using a mechanistic language. Similarly, Waytz et al. (2014) demonstrated that assigning a human name, a gender, and a voice to an autonomous vehicle positively influenced the perception of the vehicle. Participants trusted the anthropomorphized vehicle to a higher degree and blamed it less for an accident caused by another party than an autonomous vehicle without name, gender, or voice.

Experiment 2 had two main aims. The first aim was to test whether the differences in the moral evaluations of the actions of human drivers and autonomous vehicles found in Experiment 1 could be replicated. Based on the sample effect sizes observed in Experiment 1, we expected that the main effect of agent—reflecting a more critical evaluation of the actions of the autonomous vehicle in comparison to those of the human driver—should also be obtained in Experiment 2 whereas it was questionable whether the two-way interaction between agent and action or the three-way interaction between agent, action, and number of pedestrians—that were both associated with small sample effect sizes—could be replicated. The second aim of Experiment 2 was to test whether anthropomorphizing the autonomous vehicle may help to narrow the gap between the moral evaluation of actions taken by an autonomous vehicle and a human driver in inevitable accidents with human fatalities. Therefore, we expected that the difference in the evaluation of the actions of human drivers and autonomous vehicles should be reduced when the autonomous vehicle was anthropomorphized by assigning a first name to it.

Experiment 2

Method

Participants

Participants were recruited from the research panels of GapFish GmbH (Berlin, Germany). Of the 892 participants who started the study, 80 did not complete the experiment, 10

did not meet the a-priori defined inclusion criteria (being of legal age, having sufficient German language skills, and being able to read the text on screen according to self-reports), and 37 either withdrew their consent to the processing of their data or reported that not all pictures had been displayed during the study. Additionally, 10 participants were excluded due to double participation. The final sample consisted of 755 participants (317 women, 437 men, 1 diverse), aged between 18 and 87 years (M = 46, SD = 15). Participants were randomly assigned to the human-driver condition (n = 248), the anthropomorphized-autonomous-vehicle condition (n = 250), or the autonomous-vehicle condition (n = 257).

We conducted a sensitivity power analysis parallel to that conducted for Experiment 1. Given a total sample size of N = 755 and otherwise identical assumptions, small effects of about f = .11 (Cohen, 1988) could be detected for the agent variable (human driver, anthropomorphized autonomous vehicle, autonomous vehicle).

Material and Procedure

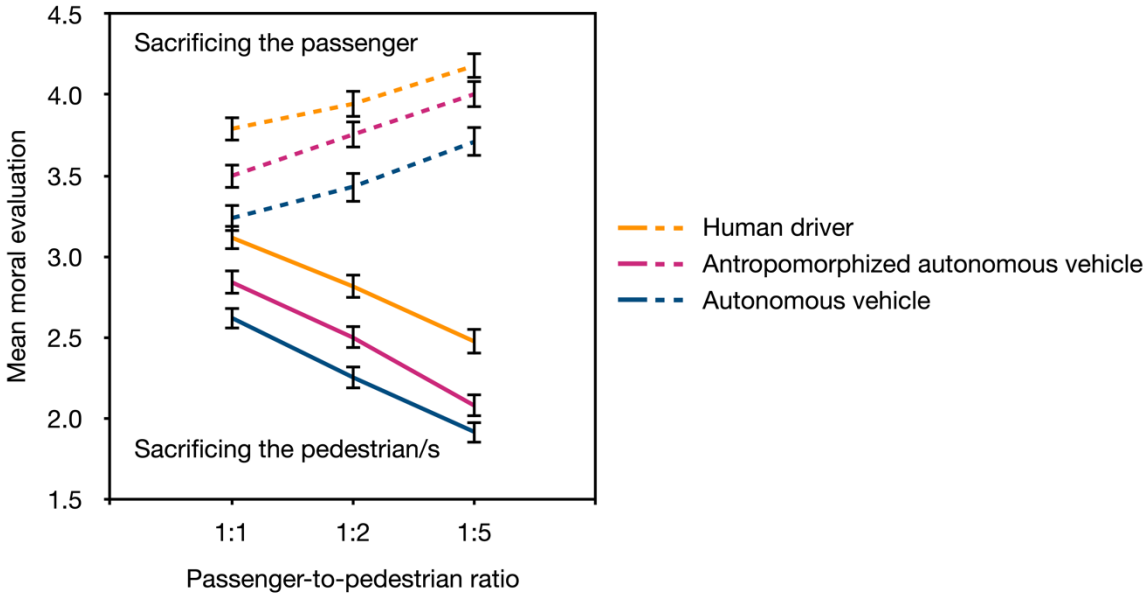
Material and procedure were identical to those of Experiment 1 with one exception. In addition to the two experimental conditions used in the first experiment (human driver and autonomous vehicle), we included a third condition with an anthropomorphized autonomous vehicle. This vehicle was introduced as a self-driving vehicle controlled by an intelligent driving system called “Alina”. Subsequently, the vehicle was only referred to by its name.

Results

The data were analyzed in the same way as in Experiment 1. The descriptive data are depicted in Figure 3.

Figure 3

Descriptive data for Experiment 2



Note. The mean moral evaluation of the actions (sacrificing the passenger [dashed lines], sacrificing the pedestrian/s [solid lines]) is displayed as a function of the passenger-to-pedestrian ratio (1:1, 1:2, and 1:5) and the agent (human driver, anthropomorphized autonomous vehicle, autonomous vehicle). The moral-evaluation scale ranged from "very reprehensible" (1) to "very justifiable" (6). The error bars represent standard errors of the mean.

As in Experiment 1, there was a significant main effect of the agent, $F(2,752) = 24.72$, $p < .001$, $\eta_p^2 = .06$. Orthogonal (Helmert) contrasts showed that the actions of the human driver were evaluated more favorably from a moral perspective than the actions of both vehicle types together, $F(1,752) = 37.76$, $p < .001$, $\eta_p^2 = .05$, and that the actions of the anthropomorphized autonomous vehicle were evaluated more favorably than the actions of the autonomous vehicle, $F(1,752) = 11.35$, $p = .001$, $\eta_p^2 = .01$. Sacrificing the passenger was evaluated as more morally justifiable than sacrificing the pedestrian/s, $F(1,752) = 399.57$, $p < .001$,

$\eta_p^2 = .35$. The interaction between these two variables was not significant, $F(2,752) = 0.30$, $p = .742$, $\eta_p^2 < .01$.

Furthermore, the analysis revealed a significant main effect of the number of pedestrians on the road, $F(2,751) = 28.18$, $p < .001$, $\eta_p^2 = .07$. As in Experiment 1, the direction of this effect depended on the action that was taken, $F(2,751) = 219.12$, $p < .001$, $\eta_p^2 = .37$. An increase in the number of pedestrians led to a significant increase in the moral evaluation of the act of sacrificing the passenger (simple main effect analyses; all comparisons of the different passenger-to-pedestrian ratios: $p < .001$) while it led to a significant decrease in the moral evaluation of sacrificing the pedestrian/s (simple main effect analyses; all comparisons of the different passenger-to-pedestrian ratios: $p < .001$). The effect of the number of pedestrians did not differ among agents, $F(4,1502) = 1.10$, $p = .353$, $\eta_p^2 < .01$. The three-way interaction was also not significant, $F(4,1502) = 0.86$, $p = .485$, $\eta_p^2 < .01$.

Discussion

The global difference in the moral evaluation of the actions of human drivers and autonomous vehicles observed in Experiment 1 was clearly replicated in Experiment 2. Regardless of the type of action or the number of pedestrians on the road, the actions of the human driver were evaluated most favorably while the actions of the non-anthropomorphized autonomous vehicle were evaluated least favorably. This finding suggests that there is a clear bias to evaluate the actions of the autonomous vehicle particularly critically. We were interested in whether it would be possible to narrow this evaluation gap by anthropomorphizing the autonomous vehicle and refereeing to it using a first name. Whereas this rather simple measure did not fully eliminate the difference between the action evaluation of the human driver and the anthropomorphized autonomous vehicle, it reduced the evaluation gap between human drivers and autonomous vehicles.

In addition, the results of Experiment 2 add to the evidence supporting the assumption that utilitarian considerations are involved in the moral evaluation of both human drivers and autonomous vehicles. Specifically, participants evaluated the actions of both human drivers and autonomous vehicles more favorably if they were compatible with the utilitarian principle of saving more lives. Furthermore, if the life of one passenger had to be weighed against the life of one pedestrian, participants evaluated the action that spared the life of the pedestrian more favorably than the action that spared the life of the passenger. We found no evidence that either of these effects differed as a function of whether the agent was a human driver or an autonomous vehicle. In terms of statistical tests, a statistically significant two-way interaction between agent and action and a statistically significant three-way interaction between agent, action, and number of pedestrians could have been interpreted as evidence of qualitative differences in the moral evaluation of the actions between human drivers and autonomous vehicles. These interactions were statistically significant but rather small in Experiment 1 and clearly failed to replicate in Experiment 2 despite an increase in sample size which, in turn, resulted in an increased sensitivity to detect such effects. We thus concluded that these interactions are negligeable.

General Discussion

The prospect of increased traffic safety is among the most salient advantages linked to the introduction of autonomous vehicles (e.g., Anderson et al., 2016; Bagloee et al., 2016; Waldrop, 2015) as the potentially error-prone human driver is freed from the driving task (National Highway Traffic Safety Administration, 2008). However, autonomous vehicles cannot avoid all accidents. No technical system will ever work perfectly which is why autonomous vehicles can be expected to be involved in accidents, as a consequence of which they will harm and even kill people. In critical accident situations, human drivers and autonomous vehicles face difficult decisions that can be morally evaluated. It is important to investigate

how the public might react to such accidents and how the actions of autonomous vehicles in critical situations are perceived and evaluated to anticipate how such accidents might affect the acceptance of autonomous vehicles (e.g., Awad et al., 2020). In two experiments, we compared the moral evaluation of actions taken by human drivers and autonomous vehicles in identical road-accident situations fashioned after moral dilemmas. The results of the two experiments indicate that utilitarian considerations are involved in the moral evaluation of the actions of both human drivers and autonomous vehicles. The more lives were spared, the more morally favorable the action was evaluated and the more people were killed, the less morally favorable the action was evaluated—irrespective of the acting agent. This is in accordance with findings indicating that utilitarian actions are preferred in moral-dilemma situations for humans as well as machines (e.g., Kallioinen et al., 2019; Li et al., 2016).

Both experiments consistently showed a bias towards evaluating the actions of autonomous vehicles less favorably than those of human drivers. This is in line with the finding of Young and Monroe (2019) that people blame autonomous vehicles more harshly than human drivers for their decisions in accident scenarios. In addition, these findings are in line with the general observation that humans seem averse to machines making moral decisions. Gogoll and Uhl (2018) observed that their participants were reluctant to delegate a moral task to a machine. Similarly, Bigman and Gray (2018) concluded that their participants preferred humans over machines making life-and-death decisions in driving, law, medicine, and the military. This critical view of machines making moral decisions may stem from the fact that it may be easier for participants to put themselves in the human drivers' shoes and to imagine their decision conflict which might be harder in case of a machine agent (Scheutz & Malle, 2021). Consequently, people might more easily justify (and potentially condone) the actions of a human agent compared to the actions of an autonomous vehicle. Here it seems relevant that manipulations that make machines more human-like, for example by ascribing mental

properties such as thoughts and feelings to them (Young & Monroe, 2019), reduce the evaluation gap between human drivers and autonomous vehicles. In line with this interpretation, Experiment 2 showed that a manipulation of anthropomorphism as simple as assigning a first name to the autonomous vehicle shifted the critical moral evaluation of the actions of the autonomous vehicle in the road-accident scenarios towards the more positive moral evaluations of the same actions performed by a human driver. This finding is in line with the observation that anthropomorphism can positively affect the perception of a machine agent (e.g., Gong, 2008; Lee et al., 2015; Waytz et al., 2014). However, anthropomorphizing the autonomous vehicle did not completely eliminate the evaluation difference between the autonomous vehicle and the human driver. Here it seems relevant that the present manipulation of anthropomorphism made only use of a single attribute (the name). The effect of anthropomorphizing autonomous vehicles on moral evaluations might be increased by assigning more human attributes (e.g., gender, voice, or mental capacities) to the autonomous vehicles (cf. Waytz et al., 2014; Young & Monroe, 2019). When increasing the level of anthropomorphism, however, care should be taken that the human characteristics are in accordance with the actual capabilities of the machine agent and the users' expectations. Otherwise, expectations might be violated which could impair the interaction with the technological system and reduce its acceptance (Malle et al., 2016). Young and Monroe (2019) suggested that people might not only require autonomous vehicles to make correct decisions but might also expect the right motivation for these decisions. Therefore, the moral evaluation of the actions of autonomous vehicles might be further improved by a providing moral justifications of the systems' decision rules.

When interpreting the present results, it should be considered that participants were asked to evaluate abstract road-accident scenarios fashioned after moral dilemmas. Moral dilemmas are useful to identify factors of a scenario that are relevant for its evaluation (e.g.,

Hauser et al., 2007; Keeling, 2020), to probe different ethical principles or theories (e.g., Goodall, 2016; Hauser et al., 2007), and to investigate moral intuitions and moral decision making (e.g., Cushman & Greene, 2012; Goodall, 2016; Wolkenstein, 2018). Abstract scenarios obviously fall short of real-life accidents experienced first-hand but they bear resemblance to newspaper reports on accidents. Newspaper reports probably are associated with low levels of immersion as they primarily describe the accident itself and perhaps the accident's causes and consequences. In that sense abstract scenarios seem suitable for investigating how the public will react to accidents with autonomous vehicles they read about in the newspaper. This seems quite relevant given that it is more likely for the majority of people to learn about accidents from newspaper reports than by witnessing or by being directly involved in an accident.

The presents study's aim was neither to develop guidelines for programming autonomous vehicles (e.g., Wolkenstein, 2018) nor to determine whether autonomous vehicles or other machines can be regarded as moral agents (for some points of view see e.g., Bonnefon et al., 2019; Gogoll & Müller, 2017; Li et al., 2016; Scheutz & Malle, 2021) and in how far concepts such as responsibility, liability, or blame can or should be assigned to machines. We focused on the moral evaluation of actions in critical traffic situations as this might represent a first step in understanding the public's reaction to accidents with autonomous vehicles. The evaluation of the agent itself or questions of blame and responsibility are a separate issue. Investigating how actions of different agents are perceived in critical traffic situations is important in order to anticipate potential problems regarding the acceptance of autonomous vehicles and to address the corresponding issues properly in public discourse (see also Awad et al., 2020). In this respect, the perception and opinion of ordinary people is especially relevant as they have to accept the technology (Malle et al., 2019). Gogoll and Uhl (2018) argued that a disliking of autonomous vehicles making moral decisions (Bigman & Gray, 2018; Gogoll &

Uhl, 2018) might represent a societal phenomenon with the potential to slow down automation processes. Considering and openly addressing concerns and evaluation differences could thus be beneficial for the introduction and the success of autonomous driving technologies.

In conclusion, the present study aims at contributing to research focusing on how moral norms are applied to machine agents. The results suggest that utilitarian considerations are involved in the moral evaluations of human drivers and autonomous vehicles alike. Overall, the actions of a human driver were evaluated more morally justifiable than identical actions taken by an autonomous vehicle although the consequences were the same. This suggests that people have a bias to evaluate the actions of autonomous vehicles more critically than those of human drivers. Accidents resembling moral dilemmas might be rare but they are emotionally salient (Bonnefon et al., 2016) and there is evidence to suggest that moral dilemmas are regarded as an important challenge for autonomous vehicles (Gill, 2021). Thus, moral decisions—which include decisions about how to distribute harm in accident situations—have the potential to affect the perception of autonomous vehicles, for example via media coverage of accidents (e.g., Anania et al., 2018). At least during the early introduction phases, a strong media attention to accidents involving autonomous vehicles seems likely (Shariff et al., 2017). A more negative moral evaluation of the actions of autonomous vehicles in comparison to those of human drivers may have negative effects on the acceptance of autonomous driving technologies (Gogoll & Uhl, 2018). Therefore, it seems relevant to search for interventions that may decrease the differential moral evaluations of human drivers and autonomous vehicles. The results of Experiment 2 suggest that anthropomorphizing autonomous vehicles can reduce the action evaluation gap between autonomous vehicles and human drivers. Therefore, assigning human characteristics might represent a promising intervention for transferring some of the leniency people display towards human drivers to autonomous vehicles.

Key Points

- Utilitarian considerations are involved in the moral evaluation of the actions of both human drivers and autonomous vehicles.
- There is an evaluation bias in favor of the actions of a human driver: The actions of a human driver are evaluated more morally justifiable than the actions of an autonomous vehicle.
- The more critical evaluation of the actions of autonomous vehicles could lead to a more critical public response to accidents with autonomous vehicles compared to those with only human drivers which might negatively affect the public acceptance of autonomous vehicles.
- The evaluation gap can be reduced by anthropomorphizing the autonomous vehicle.

References

Anania, E. C., Rice, S., Walters, N. W., Pierce, M., Winter, S. R., & Milner, M. N. (2018). The effects of positive and negative information on consumers' willingness to ride in a driverless vehicle. *Transport Policy*, *72*, 218-224. https://doi.org/10.1016/j.tranpol.2018.04.002

Anderson, J. M., Nidhi, K., Stanley, K. D., Sorensen, P., Samaras, C., & Oluwatola, O. A. (2016). *Autonomous vehicle technology: A guide for policymakers* (Rev. ed.). RAND Corporation. https://doi.org/10.7249/RR443-2

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The Moral Machine experiment. *Nature*, *563*(7729), 59-64. https://doi.org/10.1038/s41586-018-0637-6

Awad, E., Levine, S., Kleiman-Weiner, M., Dsouza, S., Tenenbaum, J. B., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2020). Drivers are blamed more than their automated cars when both make mistakes. *Nature Human Behaviour*, *4*(2), 134-143. https://doi.org/10.1038/s41562-019-0762-8

Bagloee, S. A., Tavana, M., Asadi, M., & Oliver, T. (2016). Autonomous vehicles: Challenges, opportunities, and future implications for transportation policies. *Journal of Modern Transportation*, *24*(4), 284-303. https://doi.org/10.1007/s40534-016-0117-3

Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics*, *1*(1), 71-81. https://doi.org/10.1007/s12369-008-0001-3

Becker, F., & Axhausen, K. W. (2017). Literature review on surveys investigating the acceptance of automated vehicles. *Transportation*, *44*(6), 1293-1306. https://doi.org/10.1007/s11116-017-9808-9

Bentham, J. (1789). *An introduction to the principles of morals and legislation*. T. Payne and son. Retrieved April 20, 2022, from http://galenet.galegroup.com/servlet/MOME?af=RN&ae=U102143420&srchtp=a&ste=14

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition*, *181*, 21-34. https://doi.org/10.1016/j.cognition.2018.08.003

Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*, *352*(6293), 1573-1576. https://doi.org/10.1126/science.aaf2654

Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2019). The trolley, the bull bar, and why engineers should care about the ethics of autonomous cars [point of view]. *Proceedings of the IEEE*, *107*(3), 502-504. https://doi.org/10.1109/JPROC.2019.2897447

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Earlbaum Associates.

Cushman, F., & Greene, J. D. (2012). Finding faults: How moral dilemmas illuminate cognitive structure. *Social Neuroscience*, *7*(3), 269-279. https://doi.org/10.1080/17470919.2011.614000

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, *114*(4), 864-886. https://doi.org/10.1037/0033-295X.114.4.864

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175-191. https://doi.org/10.3758/BF03193146

Foot, P. (1967). The problem of abortion and the doctrine of double effect. *Oxford Reviews*, *5*, 5-15.

Gill, T. (2021). Ethical dilemmas are really important to potential adopters of autonomous vehicles. *Ethics and Information Technology*, *23*(4), 657-673. https://doi.org/10.1007/s10676-021-09605-y

Gogoll, J., & Müller, J. F. (2017). Autonomous cars: In favor of a mandatory ethics setting. *Science and Engineering Ethics*, *23*(3), 681-700. https://doi.org/10.1007/s11948-016-9806-x

Gogoll, J., & Uhl, M. (2018). Rage against the machine: Automation in the moral domain. *Journal of Behavioral and Experimental Economics*, *74*, 97-103. https://doi.org/10.1016/j.socec.2018.04.003

Gong, L. (2008). How social is social responses to computers? The function of the degree of anthropomorphism in computer representations. *Computers in Human Behavior*, *24*(4), 1494-1509. https://doi.org/10.1016/j.chb.2007.05.007

Goodall, N. J. (2016). Away from trolley problems and toward risk management. *Applied Artificial Intelligence*, *30*(8), 810-821. https://doi.org/10.1080/08839514.2016.1229922

Hassol, J., Perlman, D., Chajka-Cadin, L., & Shaw, J. (2019). *Understanding surveys of public sentiment regarding automated vehicles: Summary of results to date and implications of past research on the dynamics of consumer adoption* (DOT-VNTSC-FHWA-20-03/FHWA-JPO-19-764). U. S. Department of Transportation & Volpe National Transportation Systems Center. https://rosap.ntl.bts.gov/view/dot/43628

Hauser, M., Cushman, F., Young, L., Kang-Xing Jin, R., & Mikhail, J. (2007). A dissociation between moral judgments and justifications. *Mind & Language*, *22*(1), 1-21. https://doi.org/10.1111/j.1468-0017.2006.00297.x

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*(2), 65-70. http://www.jstor.org/stable/4615733

Hong, J.-W., Wang, Y., & Lanz, P. (2020). Why is artificial intelligence blamed more? Analysis of faulting artificial intelligence for self-driving car accidents in experimental settings. *International Journal of Human–Computer Interaction*, *36*(18), 1768-1774. https://doi.org/10.1080/10447318.2020.1785693 Correction: *International Journal of Human-Computer Interaction, 38*(1), 102. https://doi.org/10.1080/10447318.2021.2004139

Jelinski, L., Etzrodt, K., & Engesser, S. (2021). Undifferentiated optimism and scandalized accidents: The media coverage of autonomous driving in Germany. *Journal of Science Communication*, *20*(4), Article A02. https://doi.org/10.22323/2.20040202

Kallioinen, N., Pershina, M., Zeiser, J., Nosrat Nezami, F., Pipa, G., Stephan, A., & König, P. (2019). Moral judgements on the actions of self-driving cars and human drivers in dilemma situations from different perspectives. *Frontiers in Psychology*, *10*, Article 2415. https://doi.org/10.3389/fpsyg.2019.02415

Kant, I. (2011). *Groundwork of the metaphysics of morals. A German-English edition.* (M. Gregor & J. Timmermann, Eds. & Trans.). Cambridge University Press. https://doi.org/10.1017/CBO9780511973741 (Original work published 1786)

Keeling, G. (2020). Why trolley problems matter for the ethics of automated vehicles. *Science and Engineering Ethics*, *26*(1), 293-307. https://doi.org/10.1007/s11948-019-00096-1

Koopman, P., & Wagner, M. (2017). Autonomous vehicle safety: An interdisciplinary challenge. *IEEE Intelligent Transportation Systems Magazine*, *9*(1), 90-96. https://doi.org/10.1109/MITS.2016.2583491

Lee, J.-G., Kim, K. J., Lee, S., & Shin, D.-H. (2015). Can autonomous vehicles be safe and trustworthy? Effects of appearance and autonomy of unmanned driving systems. *International Journal of Human-Computer Interaction*, *31*(10), 682-691. https://doi.org/10.1080/10447318.2015.1070547

Leiner, D. J. (2019). *SoSci Survey* [Computer software]. SoSci Survey GmbH. https://www.soscisurvey.de

Levin, S., & Wong, J. C. (2018). Self-driving Uber kills Arizona woman in first fatal crash involving pedestrian. *The Guardian*. Retrieved October 15, 2021, from https://www.theguardian.com/technology/2018/mar/19/uber-self-driving-car-kills-woman-arizona-tempe

Li, J., Zhao, X., Cho, M.-J., Ju, W., & Malle, B. F. (2016). *From trolley to autonomous vehicle: Perceptions of responsibility and moral norms in traffic accidents with self-driving cars* (SAE Technical Paper 2016-01-0164). SAE International. https://doi.org/10.4271/2016-01-0164

Lin, P. (2016). Why ethics matters for autonomous cars. In M. Maurer, J. C. Gerdes, B. Lenz, & H. Winner (Eds. & Trans.), *Autonomous driving* (pp. 69-85). Springer. https://doi.org/10.1007/978-3-662-48847-8_4 (Original German version of the book published 2015)

Malle, B. F., Magar, S. T., & Scheutz, M. (2019). AI in the sky: How people morally evaluate human and machine decisions in a lethal strike dilemma. In M. I. Aldinhas Ferreira, J. Silva Sequeira, G. Singh Virk, M. O. Tokhi, & E. E. Kader (Eds.), *Robotics and well-being* (pp. 111-133). Springer. https://doi.org/10.1007/978-3-030-12524-0_11

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015). Sacrifice one for the good of many? People apply different moral norms to human and robot agents. In *HRI'15 Proceedings of the 2015 ACM/IEEE international conference on human-robot interaction* (pp. 117-124). Association for Computing Machinery. https://doi.org/10.1145/2696454.2696458

Malle, B. F., Scheutz, M., Forlizzi, J., & Voiklis, J. (2016). Which robot am I thinking about? The impact of action and appearance on people's evaluations of a moral robot. In *HRI'16 The eleventh ACM/IEEE international conference on human robot interaction* (pp. 125-132). Institute of Electrical and Electronics Engineers, Inc. https://doi.org/10.1109/HRI.2016.7451743

Mayer, M. M., Bell, R., & Buchner, A. (2021). Self-protective and self-sacrificing preferences of pedestrians and passengers in moral dilemmas involving autonomous vehicles. *PLoS ONE*, *16*(12), Article e0261673. https://doi.org/10.1371/journal.pone.0261673

Mill, J. S. (2010). *Utilitarianism / Der Utilitarismus* (D. Birnbacher, Ed. & Trans.; Reprinted ed.). Reclam. (Original work published 1871)

National Highway Traffic Safety Administration. (2008). *National Motor Vehicle Crash Causation Survey—Report to Congress* (DOT HS 811 059). Retrieved August 12, 2020, from https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/811059

National Transportation Safety Board. (2017). *Collision between a car operating with automated vehicle control systems and a tractor-semitrailer truck near Williston, Florida, May 7, 2016* (NTSB/HAR-17/02, Product No. PB2017-102600). Retrieved October 8, 2021, from https://www.ntsb.gov/investigations/AccidentReports/Reports/HAR1702.pdf

National Transportation Safety Board. (2019). *Collision between vehicle controlled by developmental automated driving system and pedestrian, Tempe, Arizona, March 18, 2018* (NTSB/HAR-19/03, Product No. PB2019-101402). Retrieved October 15, 2021, from https://www.ntsb.gov/investigations/AccidentReports/Reports/HAR1903.pdf

Niu, D., Terken, J., & Eggen, B. (2018). Anthropomorphizing information to enhance trust in autonomous vehicles. *Human Factors and Ergonomics in Manufacturing & Service Industries*, *28*(6), 352-359. https://doi.org/10.1002/hfm.20745

Nyholm, S. (2018). The ethics of crashes with self-driving cars: A roadmap, I. *Philosophy Compass*, *13*(7), Article e12507. https://doi.org/10.1111/phc3.12507

O'Brien, R. G., & Kaiser, M. K. (1985). MANOVA method for analyzing repeated measures designs: An extensive primer. *Psychological Bulletin*, *97*(2), 316-333. https://doi.org/10.1037/0033-2909.97.2.316

Pak, R., Fink, N., Price, M., Bass, B., & Sturre, L. (2012). Decision support aids with anthropomorphic characteristics influence trust and performance in younger and older adults. *Ergonomics*, *55*(9), 1059-1072. https://doi.org/10.1080/00140139.2012.691554

Scheutz, M., & Malle, B. F. (2021). May machines take lives to save lives? Human perceptions of autonomous robots (with the capacity to kill). In J. Galliott, D. MacIntosh, & J. D. Ohlin (Eds.), *Lethal autonomous weapons: Re-examining the law and ethics of robotic warfare* (pp. 89-101). Oxford University Press. https://doi.org/10.1093/oso/9780197546048.003.0007

Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2017). Psychological roadblocks to the adoption of self-driving vehicles. *Nature Human Behaviour*, *1*(10), 694-696. https://doi.org/10.1038/s41562-017-0202-6

Smith, A., & Anderson, M. (2017). *Automation in everyday life*. Pew Research Center. Retrieved August 12, 2021, from https://www.pewresearch.org/internet/2017/10/04/automation-in-everyday-life/

Spieser, K., Treleaven, K., Zhang, R., Frazzoli, E., Morton, D., & Pavone, M. (2014). Toward a systematic approach to the design and evaluation of automated mobility-on-demand systems: A case study in Singapore. In G. Meyer & S. Beiker (Eds.), *Road vehicle automation* (pp. 229-245). Springer. https://doi.org/10.1007/978-3-319-05990-7_20 (Corrected version of the book published 2018)

Tennant, C., Stares, S., & Howard, S. (2019). Public discomfort at the prospect of autonomous vehicles: Building on previous surveys to measure attitudes in 11 countries. *Transportation Research Part F: Traffic Psychology and Behaviour*, *64*, 98-118. https://doi.org/10.1016/j.trf.2019.04.017

Thomson, J. J. (1976). Killing, letting die, and the trolley problem. *The Monist*, *59*(2), 204-217. https://doi.org/10.5840/monist197659224

Thomson, J. J. (1985). The trolley problem. *The Yale Law Journal*, *94*(6), 1395-1415. https://doi.org/10.2307/796133

Wakabayashi, D. (2018). Self-driving Uber car kills pedestrian in Arizona, where robots roam. *The New York Times*. Retrieved October 15, 2021, from https://www.nytimes.com/2018/03/19/technology/uber-driverless-fatality.html

Waldrop, M. M. (2015). Autonomous vehicles: No drivers required. *Nature*, *518*(7537), 20-23. https://doi.org/10.1038/518020a

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, *52*, 113-117. https://doi.org/10.1016/j.jesp.2014.01.005

Winkler, M., Mehl, R., Erander, H., Sule, S., Buvat, J., KVJ, S., Sengupta, A., & Khemka, Y. (2019). *The autonomous car: A consumer perspective*. Capgemini Research Institute. Retrieved October 19, 2021, from https://www.capgemini.com/wp-content/uploads/2019/05/30min-–-Report.pdf

Wolkenstein, A. (2018). What has the Trolley Dilemma ever done for us (and what will it do in the future)? On some recent debates about the ethics of self-driving cars. *Ethics and Information Technology*, *20*(3), 163-173. https://doi.org/10.1007/s10676-018-9456-6

Yadron, D., & Tynan, D. (2016). Tesla driver dies in first fatal crash while using autopilot mode. *The Guardian*. Retrieved October 15, 2021, from https://www.theguardian.com/technology/2016/jun/30/tesla-autopilot-death-self-driving-car-elon-musk

Young, A. D., & Monroe, A. E. (2019). Autonomous morals: Inferences of mind predict acceptance of AI behavior in sacrificial moral dilemmas. *Journal of Experimental Social Psychology*, *85*, Article 103870. https://doi.org/10.1016/j.jesp.2019.103870

Maike M. Mayer holds a master's degree in psychology from Heinrich Heine University Düsseldorf, Düsseldorf, Germany, where she is also a lecturer and a PhD candidate at the Department of Experimental Psychology.

Axel Buchner is a professor for cognitive and industrial psychology at the Department of Experimental Psychology, Heinrich Heine University Düsseldorf, Germany. He received his PhD in psychology from Bonn University, Bonn, Germany, in 1992 and his venia legendi in psychology from Trier University, Trier, Germany, in 1998.

Raoul Bell is a lecturer and postdoc researcher at the Department of Experimental Psychology, Heinrich Heine University Düsseldorf, Düsseldorf, Germany, where he also received his PhD in psychology in 2007 as well as his venia legendi in 2013.

Erklärung über den Eigenanteil an den in der Dissertation enthaltenen Einzelarbeiten

Meine Dissertationsschrift umfasst zwei Fachartikel mit insgesamt sieben Experimenten. Für jeden Fachartikel ist im Folgenden aufgeführt, welche Autorinnen und Autoren bei der Planung der Experimente, bei der Umsetzung der Experimente, bei der Datenauswertung und bei dem Verfassen der Manuskripte mitgearbeitet haben. Der überwiegende Teil der Arbeit lag jeweils bei der Erstautorin des Artikels.

Mayer, M. M., Bell, R., & Buchner, A. (2021). Self-protective and self-sacrificing preferences of pedestrians and passengers in moral dilemmas involving autonomous vehicles. *PLoS ONE, 16*(12), Article e0261673. https://doi.org/10.1371/journal.pone.0261673

Planung: Mayer, M. M., Bell, R., & Buchner, A.
Umsetzung: Mayer, M. M., Bell, R., & Buchner, A.
Auswertung: Mayer, M. M., Bell, R., & Buchner, A.
Manuskript: Mayer, M. M., Bell, R., & Buchner, A.

Mayer, M. M., Buchner, A., & Bell, R. (2022). Men, machines, and double standards? The moral evaluation of the actions of autonomous vehicles and human drivers in road-accident scenarios. *Manuscript submitted for publication.*

Planung: Mayer, M. M., Buchner, A., & Bell, R.
Umsetzung: Mayer, M. M., Buchner, A., & Bell, R.
Auswertung: Mayer, M. M., Buchner, A., & Bell, R.
Manuskript: Mayer, M. M., Buchner, A., & Bell, R.

Erklärung an Eides Statt

Hiermit versichere ich an Eides Statt, dass ich die Dissertation mit dem Titel »Autonome Fahrzeuge und moralische Dilemmas: Einflüsse der Perspektive, sozialer Erwünschtheit und der Handelnden« selbstständig und ohne unzulässige fremde Hilfe unter Beachtung der »Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Heinrich-Heine-Universität Düsseldorf« erstellt habe.

Ich versichere insbesondere:

(1) Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt.

(2) Alle wörtlich oder dem Sinn nach aus anderen Texten entnommenen Stellen habe ich als solche kenntlich gemacht; dies gilt für gedruckte Texte ebenso wie für elektronische Ressourcen.

(3) Die Arbeit habe ich in der vorliegenden oder einer modifizierten Form noch nicht als Dissertation vorgelegt – sei es an der Heinrich-Heine-Universität Düsseldorf oder an einer anderen Universität.

Datum: 23. Mai 2022

Name: Maike M. Mayer

Unterschrift: