Die heikle Herausforderung, Maschinen Moral beizubringen: Ethische Dilemmas autonomer Fahrzeuge

Edmond Awad, Jean-François Bonnefon, Azim Shariff und Iyad Rahwan

Selbstfahrende Fahrzeuge: Sicher, aber nicht zu hundert Prozent Autonome, selbstfahrende Fahrzeuge (AFs) werden ausgiebig getestet und trainiert und haben bereits Tausende von Kilometern im realen Straßenverkehr zurückgelegt. Heikle Zwischenfälle sind bemerkenswert selten. Wenn jedoch etwas passiert – und insbesondere bei Todesfällen – hagelt es weltweit Schlagzeilen, und viele Menschen fragen sich, ob autonome Fahrzeuge tatsächlich sicher sind und man ihnen jemals vertrauen kann. Experten sind sich hingegen einig, dass autonome Fahrzeuge tatsächlich Nutzen stiften, indem sie die Verkehrseffizienz steigern, die Umweltverschmutzung reduzieren und bis zu 90 % der Verkehrsunfälle vermeiden – solche, die durch Fahrfehler, Müdigkeit, Trunkenheit oder andere menschliche Faktoren verursacht werden. Obwohl die Sicherheit ständig verbessert wird und Verletzungen und Todesfälle deutlich reduziert werden können, wird es niemals gelingen, Unfälle komplett auszuschließen. Und wenn ein Crash droht, müssen die Roboterautos schwierige Entscheidungen treffen.

Wie reagiert man am besten, wenn ein Crash unmittelbar bevorsteht? Stellen Sie sich beispielsweise Situationen vor, wie sie Abbildung 1 darstellt. Das autonome Fahrzeug kann entweder vermeiden, mehrere Fußgänger zu verletzen, indem es ausweicht und dabei einen anderen Passanten opfert (A), oder es steht vor der Wahl, die eigenen Insassen zu opfern, um einen (B) oder mehrere (C) Fußgänger zu retten.

Obwohl diese Szenarien sehr unwahrscheinlich sind, kann man sie nicht ausschließen, wenn Millionen von AFs unterwegs sind. Außerdem werden ähnliche Trade-Offs in weniger extremen Szenarien sogar häufiger auftreten: Auch wenn es nicht um Tod oder Leben geht, muss das Auto wählen, für welche Gruppe es mehr Risiko eingeht und für welche weniger.

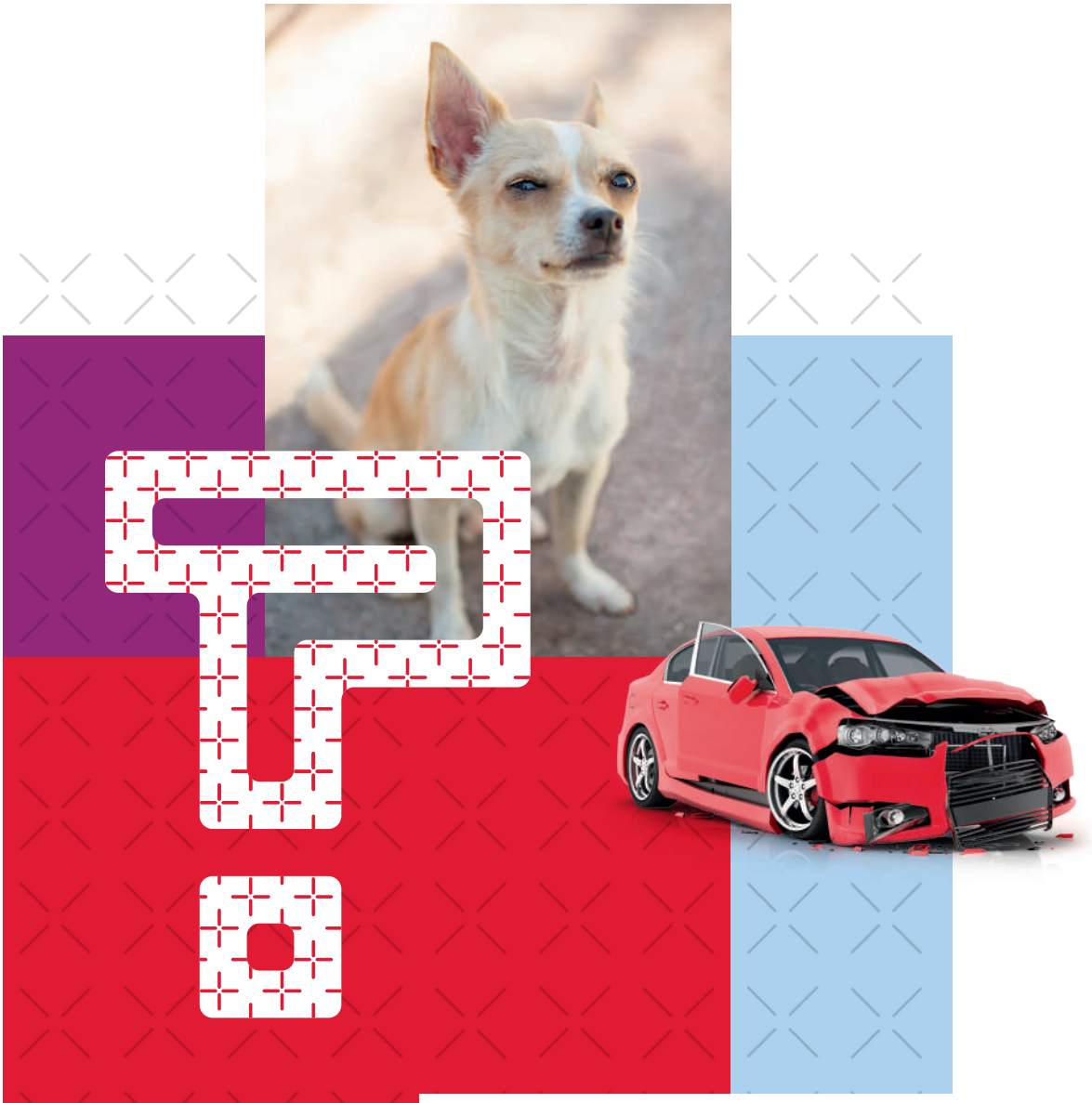
KEYWORDS
Ethik, Entscheidungsfindung, KI, Autonome Fahrzeuge, Moralische Maschinen

AUTOREN
Edmond Awad
The Media Lab, Institute for Data, Systems and Society, Massachusetts Institute of Technology, Cambridge, MA, USA
awad@mit.edu

Jean-François Bonnefon
Toulouse School of Economics (TSM-R, CNRS), Université Toulouse-1 Capitole, Toulouse, France
jean-francois.bonnefon@tse-fr.eu

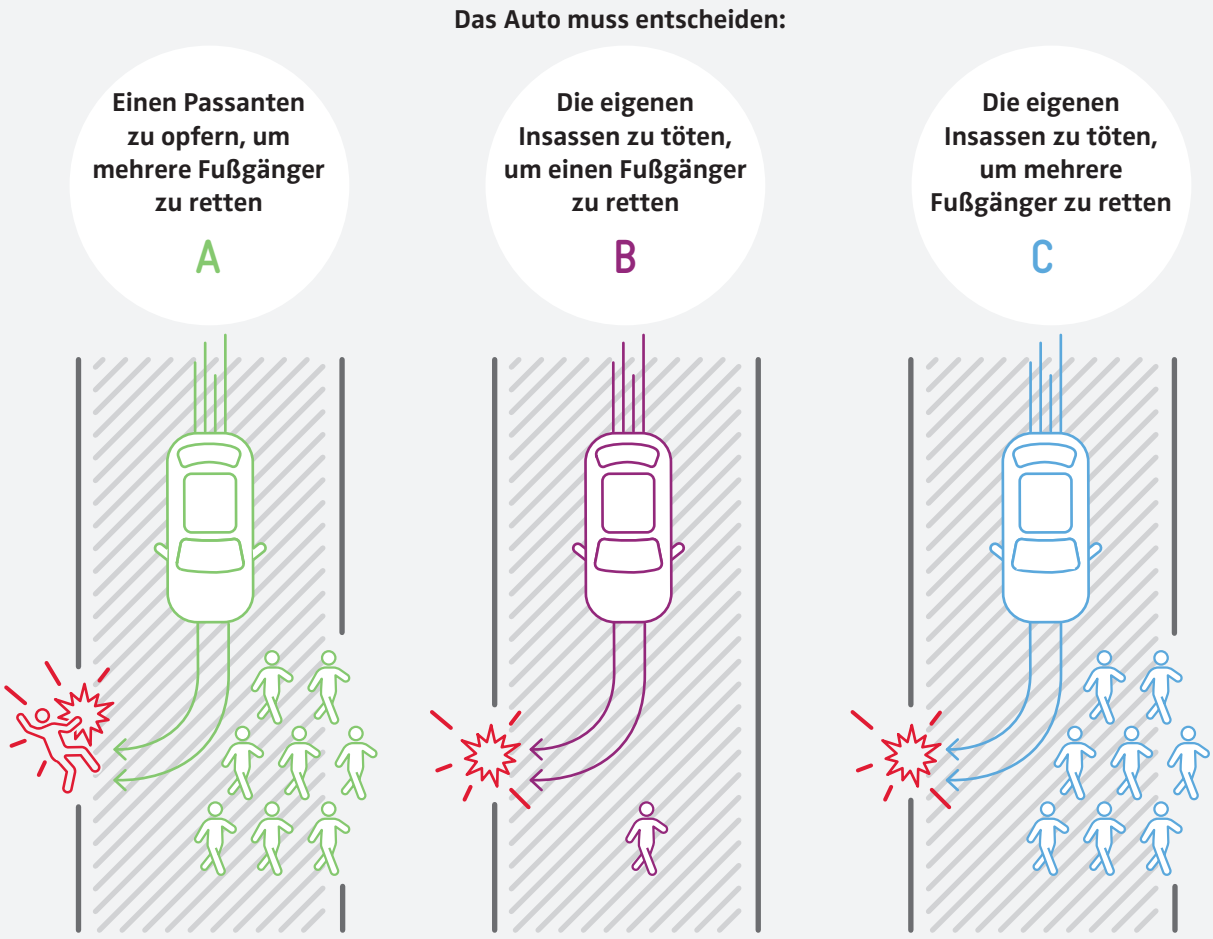
Azim Shariff
Department of Psychology, University of British Columbia, Vancouver, Canada
Canadashariff@psych.ubc.ca

Iyad Rahwan
The Media Lab, Massachusetts Institute of Technology, Cambridge, MA, USA, Center for Humans and Machines, Max-Planck Institute for Human Development, Berlin, Germany
irahwan@mit.edu



Die Konzeption ethisch autonomer Maschinen ist eine der schwierigsten Herausforderungen in der aktuellen Entwicklung der künstlichen Intelligenz.

ABBILDUNG 1 > Drei Verkehrssituationen, bei denen ein unvermeidbarer Unfall unmittelbar bevorsteht



Die AF-Programmierung muss Entscheidungsregeln liefern, was in solchen Situationen zu tun ist. Während ein menschlicher Fahrer in Sekundenbruchteilen spontan reagiert, muss ein autonomes Fahrzeug im Vorfeld bewusst programmiert werden, und irgendwer muss die Regeln dafür definieren, bevor AFs zu einem globalen Massenprodukt werden.

Algorithmen zur Steuerung von AFs müssen moralischen Prinzipien folgen, die ihre Entscheidungen in Situationen unvermeidlichen Schadens leiten. Aber was ist in solchen Fällen eine moralisch richtige Entscheidung und ein gutes Entscheidungsprinzip? Wie soll die künstliche Intelligenz (KI) für solche Momente programmiert werden? Hersteller und Regulierungsbehörden müssen drei potenziell inkompatible Ziele erreichen: konsistent sein, keine öffentliche Empörung hervorrufen und keine Käufer verschrecken. Unsere Studie zu „moralischen Maschinen" ist der Versuch herauszufinden, wie die Menschen über alternative Entscheidungsmöglichkeiten denken, die selbstfahrende Fahrzeuge mittels KI treffen müssen (siehe Box 1).

Wen retten – Fahrzeuginsassen oder Fußgänger? ✕

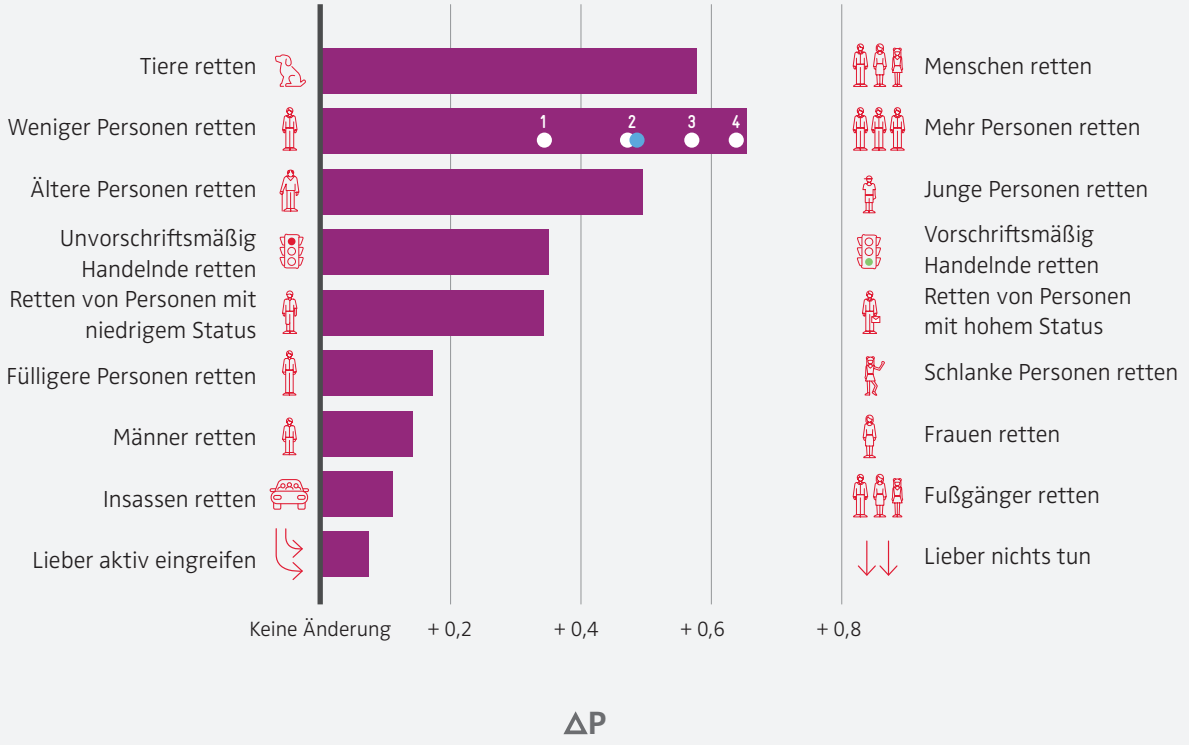
Eine weitere Online-Studie unter US-Bürgern setzte sich noch intensiver mit der Komplexität des Themas KI-gesteuerter Entscheidungsfindung in Gefahrensituationen auseinander. Diese Studie untersuchte den Trade-off zwischen der Rettung von Fahrer und Fahrgästen gegenüber der Rettung von Fußgängern und anderen Verkehrsteilnehmern – das in Abbildung 1 veranschaulichte Dilemma. Im Prinzip begrüßten die Teilnehmer

BOX 1

Erforschung moralischer Präferenzen – Das „Moral Machine Experiment"

Mit einer Gruppe von MIT-Forschern haben wir uns das Ziel gesetzt, gesellschaftliche Erwartungen und gewünschte ethische Prinzipien für das Verhalten von Maschinen zu erheben. Zu diesem Zweck haben wir die „Moral Machine" entwickelt, eine Online-Experimentierplattform zur Untersuchung von moralischen Dilemmas autonomer Fahrzeuge. Diese Plattform sammelte 40 Millionen Entscheidungen bei unvermeidlichen Unfällen. Mehr als zwei Millionen Menschen aus 233 Ländern und Regionen nahmen online an unserem mehrsprachigen „Serious Game" teil und zeigten damit, welche Schäden den meisten Menschen erträglicher erschienen. Die klarsten globalen Präferenzen ergab die Umfrage beim Schutz von Menschenleben gegenüber Tieren, beim Sichern vieler Menschenleben gegenüber wenigen und bei der Bevorzugung von jungen gegenüber älteren Menschen (siehe die ersten drei Präferenzen in Abbildung 2).

ABBILDUNG 2 > Globale Präferenzen zugunsten der Entscheidung auf der rechten Seite



ΔP ist der Unterschied zwischen der Wahrscheinlichkeit, dass Personen mit den Eigenschaften der rechten Seite gerettet werden und der Wahrscheinlichkeit, dass Personen mit den Eigenschaften der linken Seite gerettet werden, aggregiert über alle anderen Merkmale. Bei der Anzahl der Personen sind die Werte für jede zusätzliche Person dargestellt (1 bis 4 Personen). Der mittlere Wert entspricht in etwa dem Ergebnis bei 2 Personen (= blauer Punkt).

BOX 2

Kulturelle Unterschiede bei moralischen Präferenzen

Während demographische Merkmale wie Alter, Geschlecht, Einkommen, Bildung oder politische und religiöse Ansichten keine großen Unterschiede zeigten, spielte der kulturelle Hintergrund bei der Bewertung eine erkennbare Rolle. Einige der Unterschiede sind nachfolgend aufgeführt:

› Geographisch nahe gelegene Länder zeigten ähnlichere moralische Präferenzen, wobei sich drei dominante Cluster ergaben: der Westen, der Osten und der Süden.

› Teilnehmer aus kollektivistischen, östlichen Kulturen wie China und Japan verschonten die Jungen gegenüber den Alten weniger stark als Länder im südlichen Cluster, in dem die mittel- und südamerikanischen Länder dominierten.

› Teilnehmern aus individualistischen Kulturen, wie Großbritannien und den USA, war es wichtiger bei sonst gleichen Bedingungen eine höhere Anzahl an Menschenleben zu retten – möglicherweise, weil der Einzelne dort generell mehr zählt.

› Teilnehmer aus ärmeren Ländern mit schwächeren öffentlichen Institutionen erwiesen sich als toleranter gegenüber Fußgängern, die bei Rot die Straße überquerten, als bei den korrekt Handelnden.

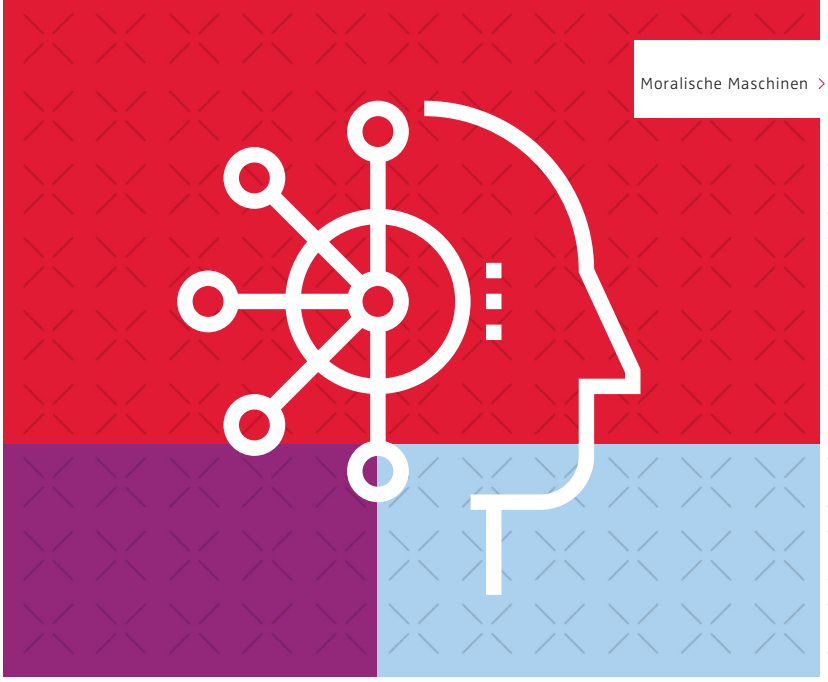
› Teilnehmer aus Ländern mit hoher wirtschaftlicher Ungleichheit tendierten stärker dazu, Personen mit hohem bzw. niedrigem sozialen Status unterschiedlich zu behandeln.

› Wir fanden auch ein paar sonstige Auffälligkeiten: Im Cluster Süd gab es beispielsweise eine starke Präferenz für die Rettung von Frauen gegenüber Männern und von schlanken gegenüber fülligeren Personen.

utilitaristische AFs, bei denen die Zahl der Opfer minimiert wurde. Die moralische Zustimmung nahm mit der Anzahl der zu rettenden Leben zu. Die Zustimmung der Teilnehmer, Insassen zu opfern, war sogar dann noch leicht positiv, wenn sie sich selbst und ein Familienmitglied als Insassen des AFs vorstellen mussten. Die Konsumenten wünschen jedoch vor allem, dass die anderen Konsumenten AFs mit einem utilitaristisch ausgerichteten Algorithmus kaufen, während sie selbst ein autonomes Fahrzeug bevorzugen würden, das um jeden Preis die eigenen Insassen schützt. Darüber hinaus lehnten die Studienteilnehmer die Durchsetzung utilitaristischer Vorschriften für AFs ab und würden ein AF, bei dem nicht die Insassen an erster Stelle stehen, weniger gern kaufen. Das moralische Dilemma bedingt also ein soziales Dilemma, das es zu lösen gilt.

Mögliche Maßnahmen zur Lösung des ethischen Dilemmas von selbstfahrenden Fahrzeugen ✕ Fahrzeuge, die über Tod oder Leben selbstständig und ohne unmittelbare menschliche Eingriffsmöglichkeiten entscheiden können, sind eine neue globale Herausforderung. Das Problem betrifft nicht nur einen Nischenmarkt, sondern täglich alle Verkehrsteilnehmer, egal ob sie mit dem Auto, dem Rad oder zu Fuß unterwegs sind. Bevor wir AFs unsere Straßen überlassen, müssen Produzenten deshalb nicht nur technische, sondern auch gesellschaftliche Herausforderungen meistern:

› **Führen allgemeiner Diskussionen über die Ethik künstlicher Intelligenz** ✕ Alle Stakeholdergruppen sollten die Herausforderungen der „Maschinenethik" als einzigartige Chance nutzen, um gemeinschaftlich zu entscheiden, was richtig und falsch ist. Danach sollten wir sicherstellen, dass Maschinen im Gegensatz zu Menschen den vereinbarten moralischen Präferenzen ausnahmslos folgen. Wir werden vermutlich keine universelle Übereinstimmung erzielen, wie die Umfrage zu moralischen Maschinen zeigt, aber dass sich weite Teile der Welt doch ziemlich einig sind, ist ermutigend.



Erarbeitung eines neuen gesellschaftlichen Vertrags

Vor mehr als hundert Jahren haben Automobile begonnen, die Straßen der Welt zu erobern. Damals wurden die ersten gesetzlichen Vorgaben eingeführt, die das Verhalten von Autofahrern und Fußgängern sowie die Produktionsstandards der Hersteller regelten. Dieses Regelwerk wurde laufend weiterentwickelt und stellt insgesamt ein Verkehrssystem dar, dem die Gesellschaft im Großen und Ganzen vertraut. Die Integration autonomer Fahrzeuge wird sehr bald einen neuen Gesellschaftsvertrag mit klaren Richtlinien dafür erfordern, wer für verschiedene Arten von Unfällen verantwortlich ist, wie die Überwachung und Durchsetzung von Regeln erfolgen soll und wie man Vertrauen zwischen allen Beteiligten schaffen kann. Dieser Prozess wird ähnlich transformativ sein wie damals, aber wahrscheinlich in einem viel kürzeren Zeitraum stattfinden.

Vertrauensfördernde Maßnahmen zur Vorbereitung der Öffentlichkeit setzen

Das moralische Dilemma, wer bei lebensbedrohlichen Vorfällen gerettet werden soll, führt zu einem sozialen Dilemma. Die Menschen erkennen den utilitaristischen Ansatz als den ethischeren an, und als Mitbürger wollen sie, dass Autos so viele Menschenleben retten wie möglich. Als Konsumenten hingegen möchten sie ein Auto, das sie selbst am besten schützt. Sowohl die Umsetzung der einen als auch der anderen Strategie birgt für Hersteller gewisse Risiken: Setzen sie auf Selbstschutz, riskieren sie öffentliche Empörung, setzen sie hingegen auf eine utilitaristische Strategie, könnten sie Konsumenten abschrecken. Damit sich die Menschen sicher fühlen und AFs vertrauen können, brauchen wir einen öffentlichen Diskurs darüber, dass AFs ganz generell zu einer Verringerung der Unfallquote führen und dadurch auch das Risiko für Fahrgäste reduzieren. Andernfalls könnte die intensive Medienberichterstattung über seltene Unfälle die Risikowahrnehmung potenzieller Insassen verzerren und die positiven, weitaus größeren Sicherheitseffekte irrational überschatten.

Die nächste Zeit wird spannend. Die Konzeption ethisch autonomer Maschinen ist eine der schwierigsten Herausforderungen in der aktuellen Entwicklung der künstlichen Intelligenz. Da wir im Begriff sind, Millionen von Fahrzeugen mit Entscheidungsautonomie auszustatten, ist eine ernsthafte Beschäftigung mit der Moral von Algorithmen dringlicher denn je. Unser datenbasierter Ansatz zeigt, wie der Bereich der experimentellen Ethik wichtige Erkenntnisse über die moralischen, kulturellen und rechtlichen Standards liefern kann, die Menschen von den Algorithmen selbstfahrender Fahrzeuge erwarten. Und selbst wenn wir diese Probleme angehen und schließlich lösen, bleiben weitere Herausforderungen im Zusammenhang mit AFs nach wie vor brisant, wie z.B. Hacking, Haftungsfragen und die Verdrängung menschlicher Arbeitskräfte. Es ist und bleibt interessant!

LITERATURHINWEISE

Awad, E.; Dsouza, S.; Kim, R.; Schulz, J.; Henrich, J.; Shariff, A.; Bonnefon, J.K. und Rahwan, I. (2018): „The Moral Machine Experiment", Nature. 563. doi 10.1038/s41586-018-0637-6.

Bonnefon, J.-F.; Shariff, A. und Rahwan, I. (2016): „The Social Dilemma of Autonomous Vehicles", Science. 352. doi: 10.1126/science.aaf2654.

Bonnefon J.-F.; Shariff A. und Rahwan, I. (2019): https://emea01.safelinks.protection.outlook.com Proceedings of the IEEE, Vol. 107, 502-504.

Shariff, A.; Bonnefon, J.-F. und Rahwan, I. (2017): „Psychological roadblocks to the adoption of self-driving vehicles", Nature Human Behaviour https://doi.org/10.1038/s41562-017-0202-6