Transportation Research Part A 94 (2016) 182–193

Contents lists available at ScienceDirect

Transportation Research Part A

journal homepage: www.elsevier.com/locate/tra

Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability?

Nidhi Kalra *, Susan M. Paddock

RAND Corporation, 1776 Main Street, Santa Monica, CA 90401, United States

ARTICLE INFO

Article history:
Received 22 March 2016
Received in revised form 19 July 2016
Accepted 6 September 2016
Available online 23 September 2016

Keywords:
Autonomous vehicles
Vehicle automation
Safety
Reliability
Test driving
Statistical significance

ABSTRACT

How safe are autonomous vehicles? The answer is critical for determining how autonomous vehicles may shape motor vehicle safety and public health, and for developing sound policies to govern their deployment. One proposed way to assess safety is to test drive autonomous vehicles in real traffic, observe their performance, and make statistical comparisons to human driver performance. This approach is logical, but it is practical? In this paper, we calculate the number of miles of driving that would be needed to provide clear statistical evidence of autonomous vehicle safety. Given that current traffic fatalities and injuries are rare events compared to vehicle miles traveled, we show that fully autonomous vehicles would have to be driven hundreds of millions of miles and sometimes hundreds of billions of miles to demonstrate their reliability in terms of fatalities and injuries. Under even aggressive testing assumptions, existing fleets would take tens and sometimes hundreds of years to drive these miles—an impossible proposition if the aim is to demonstrate their performance prior to releasing them on the roads for consumer use. These findings demonstrate that developers of this technology and third-party testers cannot simply drive their way to safety. Instead, they will need to develop innovative methods of demonstrating safety and reliability. And yet, the possibility remains that it will not be possible to establish with certainty the safety of autonomous vehicles. Uncertainty will remain. Therefore, it is imperative that autonomous vehicle regulations are adaptive—designed from the outset to evolve with the technology so that society can better harness the benefits and manage the risks of these rapidly evolving and potentially transformative technologies.

1. Introduction

In the United States, roughly 32,000 people are killed and more than two million injured in crashes every year (Bureau of Transportation Statistics, 2015). U.S. motor vehicle crashes as a whole can pose economic and social costs of more than $800 billion in a single year (Blincoe et al., 2015). And, more than 90 percent of crashes are caused by human errors (National Highway Traffic Safety Administration, 2015)—such as driving too fast and misjudging other drivers' behaviors, as well as alcohol impairment, distraction, and fatigue.

Autonomous vehicles have the potential to significantly mitigate this public health crisis by eliminating many of the mistakes that human drivers routinely make (Anderson et al., 2016; Fagnant and Kockelman, 2015). To begin with,

* Corresponding author.
E-mail addresses: nidhi_kalra@rand.org (N. Kalra), susan_paddock@rand.org (S.M. Paddock).

http://dx.doi.org/10.1016/j.tra.2016.09.010
0965-8564/© 2016 Elsevier Ltd. All rights reserved.

autonomous vehicles are never drunk, distracted, or tired; these factors are involved in 41 percent, 10 percent, and 2.5 percent of all fatal crashes, respectively (National Highway Traffic Safety Administration, 2011; Bureau of Transportation Statistics, 2014b; U.S. Department of Transportation, 2015).[1] Their performance may also be better than human drivers because of better perception (e.g., no blind spots), better decisionmaking (e.g., more-accurate planning of complex driving maneuvers like parallel parking), and better execution (e.g., faster and more-precise control of steering, brakes, and acceleration).

However, autonomous vehicles might not eliminate all crashes. For instance, inclement weather and complex driving environments pose challenges for autonomous vehicles, as well as for human drivers, and autonomous vehicles might perform worse than human drivers in some cases (Gomes, 2014). There is also the potential for autonomous vehicles to pose new and serious crash risks, e.g., crashes resulting from cyber-attacks (Anderson et al., 2016). Clearly, autonomous vehicles present both enormous potential benefits and enormous potential risks.

Given the high stakes, policymakers, the transportation industry, and the public are grappling with a critical concern: How safe should autonomous vehicles be before they are allowed on the road for consumer use? For the answer to be meaningful, however, one must also be able to address a second concern: How safe are autonomous vehicles?

Perhaps the most logical way to assess safety is to test-drive autonomous vehicles in real traffic and observe their performance. Developers of autonomous vehicles rely upon this approach to evaluate and improve their systems,[2] almost always with trained operators behind the wheel who are ready to take control in the event of an impending failure incident.[3] They can analyze the failure incident after the fact to assess what the autonomous vehicle would have done without intervention, and whether it would have resulted in a crash or other safety issue (Google, 2015). Developers have presented data from test driving to Congress in hearings about autonomous vehicle regulation (Urmson, 2016).

But is it practical to assess autonomous vehicle safety through test-driving? The safety of human drivers is a critical benchmark against which to compare the safety of autonomous vehicles. And, even though the number of crashes, injuries, and fatalities from human drivers is high, the rate of these failures is low in comparison with the number of miles that people drive. Americans drive nearly 3 trillion miles every year (Bureau of Transportation Statistics, 2015). The 2.3 million reported injuries in 2013 correspond to a failure rate of 77 reported injuries per 100 million miles. The 32,719 fatalities in 2013 correspond to a failure rate of 1.09 fatalities per 100 million miles.

For comparison, Google's autonomous vehicle fleet, which currently has 55 vehicles, was test-driven approximately 1.3 million miles in autonomous mode and was involved in 11 crashes from 2009 to 2015.[4] Blanco et al. (2016) recently compared Google's fleet performance with human-driven performance. They found that Google's fleet might result in fewer crashes with only property damage, but they could not draw conclusions about the relative performance in terms of two critical metrics: injuries and fatalities. Given the rate of human and autonomous vehicle failures, there were simply not enough autonomously driven miles to make statistically significant comparisons.

In this report, we answer the next logical question: How many miles[5] would be enough? In particular, we first ask:

1. How many miles would autonomous vehicles have to be driven without failure to demonstrate that their failure rate is below some benchmark? This provides a lower bound on the miles that are needed.

However, autonomous vehicles will not be perfect and failures will occur. Given imperfect performance, we next ask:

2. How many miles would autonomous vehicles have to be driven to demonstrate their failure rate to a particular degree of precision?
3. How many miles would autonomous vehicles have to be driven to demonstrate that their failure rate is statistically significantly lower than the human driver failure rate?

[1] This does not mean that 53.5 percent of all fatal crashes are caused by these factors because a crash may involve, but not be strictly caused by, one of these factors, and because more than one of these factors may be involved in a single crash.

[2] Extensive testing on public roads is essential for developing and evaluating autonomous vehicles, given their great complexity and the diversity and unpredictability of conditions in which they need to operate. In contrast, typical automobile components are significantly simpler and their operating conditions can be well defined and recreated in controlled settings, which enables laboratory testing and verification. Curtain-style air bags, for example, are tested with a combination of component tests to assess inflation time, fill capacity, and other responses in a range of temperature conditions and impact configurations, as well as laboratory crash testing to evaluate their performance in collisions (Kaleto et al., 2001).

[3] Some states, such as California, require trained drivers to be behind the wheel of any autonomous vehicle driving on public roads (California Vehicle Code, 2012).

[4] Two of these crashes involved injury and none involved a fatality. Seven of the crashes did not reach a level of severity that would warrant a Department of Motor Vehicles report (Blanco et al., 2016).

[5] Note that not all miles of road are created equal. The miles used to demonstrate autonomous vehicle safety must represent the full range of conditions (climate, terrain, congestion, etc.) in which humans drive, and be proportionally distributed as well. That is, if 10 percent of human-driven miles occur in snow, so too must the autonomous vehicle test miles.

We answer each of these questions with straightforward statistical approaches. Given that fatalities and injuries are rare events, we will show that fully autonomous vehicles[6] would have to be driven hundreds of millions of miles and sometimes hundreds of billions of miles to demonstrate their reliability in terms of fatalities and injuries. Under even aggressive testing assumptions, existing fleets would take tens and sometimes hundreds of years to drive these miles—an impossible proposition if the aim is to demonstrate their performance prior to releasing them on the roads for consumer use.

These results demonstrate that developers of this technology and third-party testers cannot simply drive their way to safety. Instead, they will need to develop innovative methods of demonstrating safety and reliability. This is a rapidly growing area of research and development. We hope the data and figures in this paper will serve as a useful reference in developing those alternative methods, and a benchmark and method for assessing their efficiency.

The next three sections provide an explanation, analysis, and results for each of these questions. We end with a summary and discussion of results and draw conclusions about their implications for stakeholders of autonomous vehicle technology.

2. How many miles would autonomous vehicles have to be driven without failure to demonstrate that their failure rate is below some benchmark?

2.1. Statistical method

We can answer this question by reframing failure rates as reliability rates and using success run statistics based on the binomial distribution (O'Connor and Kleyner, 2012). If the per-mile failure rate of a vehicle is F, then the reliability R is $1 - F$ and can be interpreted as the probability of not having a failure in any given mile. In practice, unless the technology is truly perfect, there likely will be failures during testing.[7] However, a simple "no failures" scenario (see Eq. (1)) can be used to estimate a lower bound on the number of failure-free miles, n, that would be necessary to establish the reliability of autonomous vehicles with confidence level C: [8]

$$C = 1 - R^n \tag{1}$$

This is useful if, for example, a developer has driven autonomous vehicles for a certain number of failure-free miles and wishes to know the reliability (or, equivalently, the failure rate) that can be claimed at a particular level of confidence. Alternatively, for a given confidence C and reliability R we can solve for n, the number of miles required with no failures:

$$n = \ln(1 - C)/\ln(R) \tag{2}$$

This equation is usually used to show the survival of a product based on duration of use (Kleyner, 2014).

2.2. Example calculation

To demonstrate that fully autonomous vehicles have a fatality rate of 1.09 fatalities per 100 million miles (R = 99.9999989%) with a C = 95% confidence level, the vehicles would have to be driven 275 million failure-free miles. With a fleet of 100 autonomous vehicles being test-driven 24 h a day, 365 days a year at an average speed of 25 miles per hour, this would take about 12.5 years.

[6] Note that the term "autonomous vehicle" can refer to different degrees of autonomy. The Society of Automotive Engineers International (2014), for example, defines three levels of automated driving. Vehicles with "conditional automation" can drive themselves in certain conditions but may request human intervention. Vehicles with "high automation" can drive themselves in certain conditions without requiring human intervention. Vehicles with "full automation" can drive under all roadway and environmental conditions in which a human can drive. The numerical results in this report assess the miles needed to demonstrate the reliability of this last class of fully autonomous vehicles. Therefore, we use the total fatality, injury, and crash rates of human drivers in the United States as benchmarks against which to compare autonomous vehicle performance. However, the statistical approaches described in this report can be used to compare the reliability for any autonomy mode. Doing so would require changing the human performance benchmarks against which these other modes are compared.

[7] In the case of an imperfect vehicle, the probability C of a driverless car with reliability R having k failures while driving N miles is: $C = 1 - \sum_{i=0}^{k} \frac{N!}{i!(N-i)!} R^{N-i}(1-R)^{i}$.

[8] In reliability testing, the confidence level of $100(1 - \alpha)\%$ is the probability that the true failure rate is within some range $[0, U]$, where U is a random variable for the upper bound; or, equivalently, the probability of the true success rate (reliability) is within $[1 - U, 1]$ (Darby, 2010). Once data are observed and u is estimated, then u is no longer random, and the interval $[1 - u, 1]$ either does or does not contain the true failure rate. Thus, the interval is described in terms of a confidence level rather than a probability (Martz and Waller, 1982).

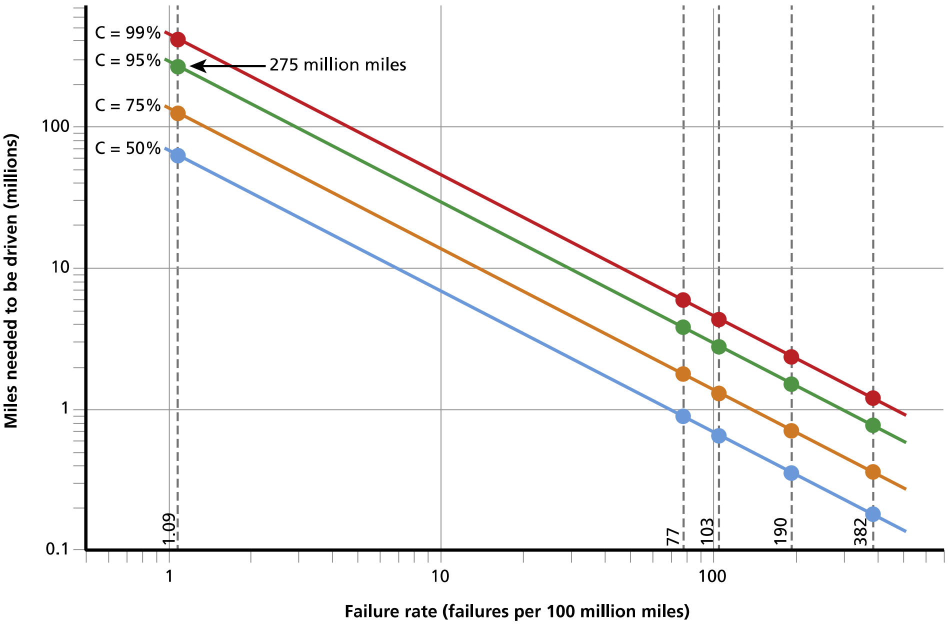

Fig. 1. Failure-free miles needed to demonstrate maximum failure rates. *SOURCE:* Authors' analysis. *NOTE:* The four diagonal lines show results for different levels of confidence. The five dashed vertical reference lines indicate the failure rates of human drivers in terms of fatalities (1.09), reported injuries (77), estimated total injuries (103), reported crashes (190), and estimated total crashes (382).

2.3. Results

Fig. 1 shows how many failure-free miles fully autonomous vehicles would have to be driven to demonstrate maximum failure rates to different levels of confidence. We chose a range of 1–400 failures per 100 million miles to include the range of fatality, injury, and crash rates for human drivers. For reference, we show the failure rate of human drivers as dashed vertical lines.[9] Reference lines are shown for fatalities (1.09), reported injuries (77), and reported crashes (190) per 100 million miles. It is also known that injuries and crashes may be significantly underreported: one study suggests by 25 percent and 60 percent, respectively (Blincoe et al., 2015). Therefore, we have also shown reference lines that could reflect a truer estimate of human-driven injuries (103) and crashes (382) per 100 million miles. The 275-million mile data point corresponding to the 95% confidence level is annotated in Fig. 1. We assess sensitivity to different levels of confidence because different fields use different standards that result in large differences in the number of required miles. While 95% and 99% confidence levels are widely used, the automotive industry sometimes uses a 50% confidence level for vehicle components (Misra, 2008). The diagonal lines represent C = 50%, 75%, 95%, and 99%.

This analysis shows that for fatalities it is not possible to test-drive autonomous vehicles to demonstrate their safety to any plausible standard, even if we assume perfect performance. In contrast, one could demonstrate injury and crash reliability to acceptable standards based on driving vehicles a few million miles. However, it is important to recognize that this is a theoretical lower bound, based on perfect performance of vehicles. In reality, autonomous vehicles will have failures—not only commonly occurring injuries and crashes in which autonomous vehicles have already been involved, but also fatalities. Our second and third questions quantify the miles needed to demonstrate reliability through driving given this reality.

[9] These rates reflect failures from all motor vehicles, including cars and light trucks, motorcycles, large trucks, and buses. One could restrict comparisons to only some subsets of these data, e.g., omitting motorcycle fatalities, which occur at a rate that is 20 times higher than the overall fatality rate (about 23 fatalities per 100 million miles driven) (Bureau of Transportation Statistics, 2014a). This would not change the statistical methods shown here, but the miles needed to demonstrate comparative levels of performance would change. For example, by omitting motorcycle fatalities, the remaining human-driven fatality rate would decrease and so the miles needed to demonstrate comparable autonomous vehicle performance would increase. We use overall failure rates because there is the potential for all travel to occur in autonomous vehicles and for autonomous vehicles to affect the safety of all road users. It is possible, for example, that current motorcyclists may in the future choose to travel by autonomous passenger vehicles for safety or other reasons, or that autonomous passenger vehicles may lead to fewer motorcycle fatalities, even if motorcycles remain human driven.

3. How many miles would autonomous vehicles have to be driven to demonstrate their failure rate to a particular degree of precision?

3.1. Statistical method

To estimate the true autonomous vehicle failure rate, we must count the number of events (failures) that occur for a given distance driven. The failure rate is estimated as $\hat{\lambda} = x/n$, where x is the number of observed failures observed over n miles driven. We can describe the precision of the failure rate estimate using the width of a $100(1-\alpha)\%$ confidence interval (CI).[10] If the number of failures is expected to be greater than 30, then a normal approximation to the Poisson distribution can be used. An approximate CI for the failure rate is:

$$\left(\frac{x - z_{1-\alpha/2}\sqrt{x}}{n}, \frac{x + z_{1-\alpha/2}\sqrt{x}}{n}\right) \tag{3}$$

where $z_{1-\alpha/2}$is $100*(1-\alpha/2)$th quantile of standard normal distribution.[11] The half-width of the CI is

$$\frac{z_{1-\alpha/2}\sqrt{x}}{n}$$

and it provides an estimate of the precision of the failure rate estimate, $\hat{\lambda} = x/n$. We can calculate the precision relative to the failure estimate rate as

$$\frac{\frac{z_{1-\alpha/2}\sqrt{x}}{n}}{\frac{x}{n}}$$

which simplifies to

$$\frac{z_{1-\alpha/2}}{\sqrt{x}}$$

If δ is our desired degree of precision (e.g., if we wish to estimate the failure rate to within 20%, $\delta = 0.2$) then the number of failures one must observe to estimate the failure rate with a precision of δ is:

$$x = \left(\frac{z_{1-\frac{\alpha}{2}}}{\delta}\right)^2 \tag{4}$$

If the assumed failure rate (prior to data collection) is λ_* (Mathews, 2010), then Equation (5) implies the number of miles that must be driven is:

$$x = \frac{\left(\frac{z_{1-\frac{\alpha}{2}}}{\delta}\right)^2}{\lambda_*} \tag{5}$$

3.2. Example calculation

We can demonstrate this as follows. Given some initial data on its safety performance, suppose we assume that a fully autonomous vehicle fleet had a true fatality rate of 1.09 per 100 million miles. We could use this information to determine the sample size (number of miles) required to estimate the fatality rate of the fleet to within 20% of the assumed rate using a 95% CI. We apply Eq. (4) to estimate the number of fatalities we would need to observe before having this level of precision in the fatality rate estimate: $(1.96/0.20)^2 = 96$. (Here, 1.96 is the z-score associated with a two-sided 95% CI for the standard normal distribution.) We apply Eq. (5) to determine how many miles of driving this would require:

$$x = \frac{\left(\frac{1.96}{0.2}\right)^2}{1.09 \times 10^{-8}} = 8,811,009,174$$

This is approximately 8.8 billion miles. With a fleet of 100 autonomous vehicles being test-driven 24 h a day, 365 days a year at an average speed of 25 miles per hour, this would take about 400 years.

[10] In this context, a $100(1-\alpha)\%$ CI is an estimate of the random interval (L, U) that contains the true failure rate λ with probability $(1-\alpha)$. If l and u are the estimates of random variables L and U, then (l, u) is called the CI for λ with confidence coefficient $(1-\alpha)$ (DeGroot, 1986). A $100(1-\alpha)\%$ CI can be interpreted as follows: If one were to run the experiment that generated the data and conduct the analysis repeatedly, in $100(1-\alpha)\%$ of the samples the $100(1-\alpha)\%$ CIs calculated in each of those experiments would contain the true mean.

[11] If the number of events is fewer than 30, an exact CI could alternatively be calculated (Ulm, 1990).

3.3. Results

Fig. 2 shows how many miles fully autonomous vehicles would have to be driven to estimate the failure rate to different degrees of precision with 95% confidence. The number of miles that must be driven to achieve a given level of precision in the failure rate estimate decreases as the failure rate increases. The diagonal lines represent 5%, 10%, and 20% precision. As in Fig. 1, we show for reference the failure rate of human drivers as dashed vertical lines for fatalities (1.09), reported injuries (77), estimated total injuries (103), reported crashes (190), and estimated total crashes (382) per 100 million miles. The 8.8-billion mile data point corresponding to this example is annotated in Fig. 2.

These results show that it may be impossible to demonstrate the reliability of high-performing autonomous vehicles (i.e., ones with failure rates comparable to or better than human failure rates) to any reasonable degree of precision. For instance, even if the safety of autonomous vehicles is low—hundreds of failures per 100 million miles, which is akin to human-driven injury and crash rates—demonstrating this would take tens or even hundreds of millions of miles, depending on the desired precision. For low failure rates—1 per 100 million miles, which is akin to the human-driven fatality rate—demonstrating performance to any degree of precision is impossible—requiring billions to hundreds of billions of miles. These results show that as autonomous vehicles perform better, it becomes harder—if not impossible—to assess their performance with accuracy because of the extreme rarity of failure events.

4. How many miles would autonomous vehicles have to be driven to demonstrate that their failure rate is statistically significantly lower than the human driver failure rate?

4.1. Statistical method for significance testing

Setting up the statistical significance test requires that we specify the null hypothesis that we are testing, which is that the failure rate, λ, is greater than or equal to λ_0. Here, we set $\lambda_0 = H$, the human driver failure rate.[12] We also must specify an alternative hypothesis, which we specify as $\lambda < H$. In the context of significance testing, α is the significance level, or Type 1 error rate of the test, which is defined as the probability of rejecting the null hypothesis when the null hypothesis is true—in other words, a false positive. In the context of autonomous vehicles, a false positive would occur if data suggest that autonomous vehicles perform better than human drivers, when in fact they do not—a dangerous proposition for policymakers, technology developers, the insurance industry, and of course consumers.

To be able to test the null hypothesis with significance level α, one can examine the upper confidence bound from Eq. (3) for whether the estimated failure rate is lower than the human driver rate, i.e., [13]

$$\frac{x + z_{1-\alpha}\sqrt{x}}{n} < H.$$

If so, then the null hypothesis can be rejected at the αth significance level. To assess when the confidence bound would be expected to be less than H requires a guess of the autonomous vehicle failure rate we expect, λ_{alt}. We set $\lambda_{alt} = (1 - A)\,H$.[14] To determine how many failures (and miles) would be required to show this, we can solve for x and n:

$$x = \left(\lambda_{alt}\frac{z_{1-\alpha}}{\lambda_0 - \lambda_{alt}}\right)^2 \tag{6}$$

$$n = \lambda_{alt}\left(\frac{z_{1-\alpha}}{\lambda_0 - \lambda_{alt}}\right)^2 \tag{7}$$

4.2. Example calculation for significance testing

We can demonstrate this as follows. Suppose a fully autonomous vehicle fleet had a true fatality rate that was A = 20% lower than the human driver fatality rate of 1.09 per 100 million miles, or 0.872 per 100 million miles. We apply Eq. (7) to determine the number of miles that must be driven to demonstrate with 95% confidence that this difference is statistically significant:

$$n = 0.872 \times 10^{-8}\left(\frac{1.645}{1.09 \times 10^{-8} - 0.872 \times 10^{-8}}\right)^2 = 4{,}965{,}183{,}486$$

It would take approximately 5 billion miles to demonstrate this difference. With a fleet of 100 autonomous vehicles test-driven 24 h a day, 365 days a year at an average speed of 25 miles per hour, this would take about 225 years.

[12] We assume H is a known benchmark rather than an estimate of some unknown quantity. This gives us the best-case estimate number of miles that need to be driven. We discuss the implications of this choice after Fig. 4.

[13] The subscript on z is $1 - \alpha$ here because this is a one-sided hypothesis test.

[14] If we were interested in testing the alternative hypothesis that $\lambda > H$, then we would compare the lower confidence bound with H: $\frac{x - z_{1-\alpha}\sqrt{x}}{n} > H$.

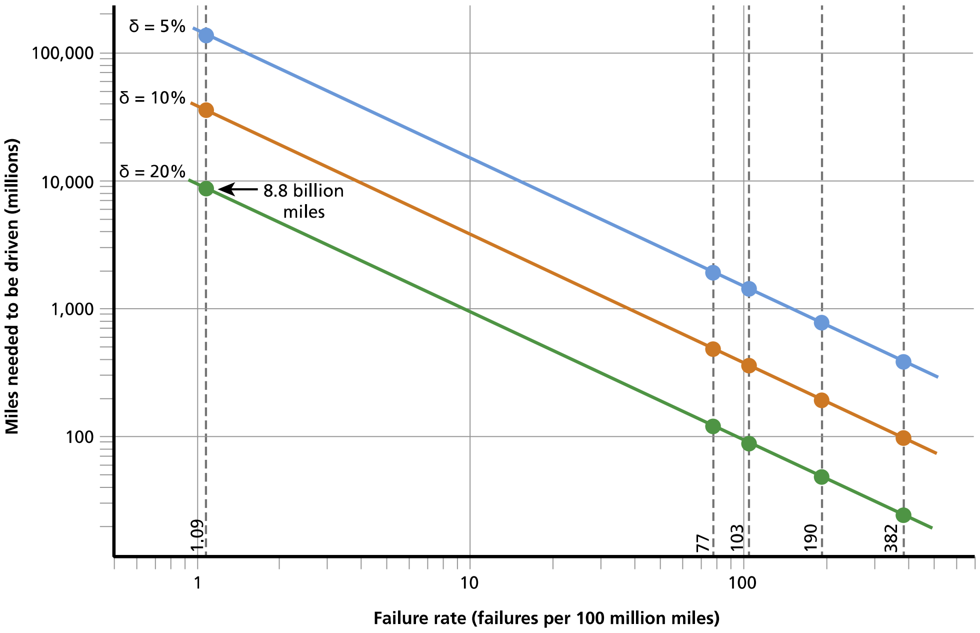

Fig. 2. Miles needed to demonstrate failure rates to a particular degree of precision. *SOURCE:* Authors' analysis. *NOTE:* These results use a 95% CI. The three diagonal lines show results for different levels of precision δ, defined as the size of the CI as a percent of the failure rate estimate. The five dashed vertical reference lines indicate the failure rates of human drivers in terms of fatalities (1.09), reported injuries (77), estimated total injuries (103), reported crashes (190), and estimated total crashes (382).

4.3. Results for significance testing

Fig. 3 shows how many miles fully autonomous vehicles would have to be driven to demonstrate that their failure rate is statistically significantly lower than the human driver failure rate with 95 percent confidence, given different values of A. The different lines represent performance relative to the human driver fatality, reported injury, estimated total injury, reported crash, and estimated total crash rates. Note that the miles needed to be driven approaches infinity as the difference between the human rate and autonomous vehicle rate approaches 0, i.e., as $A \rightarrow 0$. The 5-billion mile data point for this example is annotated in Fig. 3.

4.4. Statistical method for significance and power

Setting the significance level, α, accounts for one of the two types of errors we could make in significance testing: rejecting the null hypothesis when in fact it is true (Type I error). A limitation of determining the sample size as shown above is that it does not take into consideration Type II error, β, which is the second type of error that might occur: not rejecting the null hypothesis when the alternative is true. In the context of autonomous vehicles, a Type II error would mean that data suggest that autonomous vehicles do not perform better than human drivers, when in fact they do. While perhaps less concerning to stakeholders, this also would be a serious error as it could delay the introduction of potentially beneficial technology and needlessly perpetuate the risks posed by human drivers.

The power of the test, $100(1-\beta)\%$, is the probability of correctly rejecting the null hypothesis in favor of the alternative. The power of the test for a given number of miles, n, and hypothesized and assumed rates, λ_0 and λ_{alt}, respectively, is:

$$Power = \Phi\left(\frac{\lambda_0 - \lambda_{alt}}{\sqrt{\lambda_{alt}/n}} - z_{1-\alpha}\right) \tag{8}$$

where Φ(.) is the cumulative standard normal distribution.[15] Building upon our running example, a study with a significance level of $\alpha = 0.05$ and a requirement to drive approximately 5 billion miles would have 50% power to reject the null hypothesis.

One may instead want to know how many miles autonomous vehicles need to be driven to avoid Type I errors and Type II errors with some probability. Using the normal approximation for the distribution of fatalities, the number of miles, n, required to achieve $100(1-\beta)\%$ power at the α significance is:

[15] For a one-sided test in the other direction, the numerator of the first component of Φ(.) would be $\lambda_{alt} - \lambda_0$.

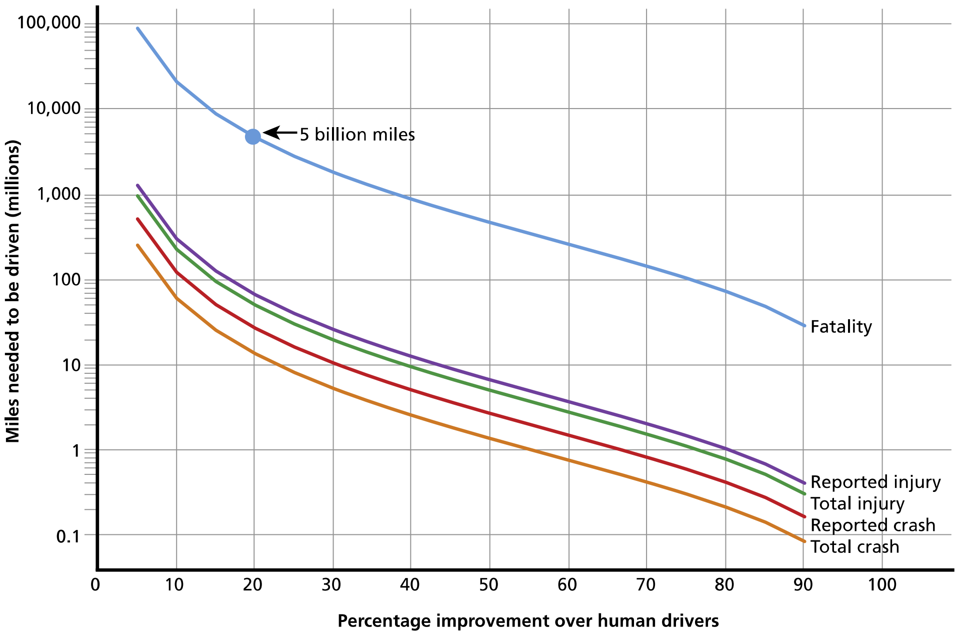

Fig. 3. Miles needed to demonstrate with 95% confidence that the autonomous vehicle failure rate is lower than the human driver failure rate. *SOURCE:* Authors' analysis. *NOTE:* The results depend upon the estimated failure rate of autonomous vehicles. This is shown on the horizontal axis and defined as a percent improvement over the human driver failure rate. The comparison can be made to the human driver fatality rate, reported injury rate, estimated total injury rate, reported crash rate, and estimated total crash rate.

$$n = \lambda_{alt}\left(\frac{z_{1-\alpha} + z_{1-\beta}}{\lambda_0 - \lambda_{alt}}\right)^2 \tag{9}$$

4.5. Example calculation for significance and power

Continuing our example, we apply Eq. (9) to determine the number of miles that autonomous vehicles must be driven to determine with 95% confidence and 80% power (i.e., β = 0.2) that their failure rate is 20% better than the human driver fatality rate:

$$n = 0.872 \times 10^{-8}\left(\frac{1.645 + 0.842}{1.09 \times 10^{-8} - 0.872 \times 10^{-8}}\right)^2 = 11,344,141,710$$

Autonomous vehicles would have to be driven more than 11 billion miles to detect this difference. With a fleet of 100 autonomous vehicles being test-driven 24 h a day, 365 days a year at an average speed of 25 miles per hour, this would take 518 years—about a half a millennium.

4.6. Results for significance and power

Fig. 4 shows how many miles fully autonomous vehicles would have to be driven to demonstrate with 95% confidence and 80% power that their failure rate is A% better than the human driver failure rate. The different lines represent performance relative to the human driver fatality, reported injury, estimated total injury, reported crash, and estimated total crash rates. The 11-billion mile data point for this example is annotated in Fig. 4.

These results show that the closer autonomous vehicles are to human performance, the more miles are required to demonstrate that the differences are statistically significant. This makes sense—the closer two population means are to each other, the more samples will be needed to determine if they are significantly different. For example, if autonomous vehicles improve fatality rates by 5% rather than 20%, the number of miles required to demonstrate a statistically significant improvement with 95% confidence and 80% power is almost ludicrous: 215 billion miles. It would take a fleet of 100 vehicles nearly 10,000 years to achieve this. Indeed, for no improvement in fatality rates between 5% and 90% would it be practical to drive the requisite number of miles with 100-vehicle fleets. For injuries and crashes, until autonomous vehicles are substantially better than human drivers (25% or greater improvement), the miles required to demonstrate significant differences over human drivers would be impractically large.

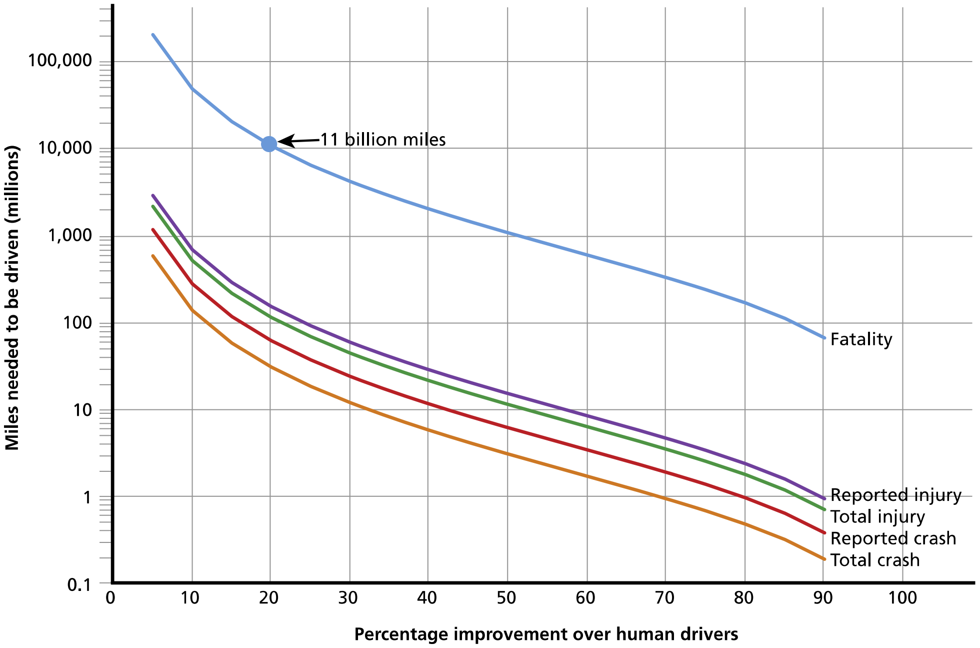

Fig. 4. Miles needed to demonstrate with 95% confidence and 80% power that the autonomous vehicle failure rate is lower than the human driver failure rate. *SOURCE:* Authors' analysis. *NOTE:* The results depend upon the estimated failure rate of autonomous vehicles. This is shown on the horizontal axis and defined as a percent improvement over the human driver failure rate. The comparison can be made to the human driver fatality rate, reported injury rate, estimated total injury rate, reported crash rate, and estimated total crash rate.

It is possible that if the policy question were to differ from how we have framed it that fewer miles could be driven to examine the reliability of autonomous vehicles. For example, suppose there was a consensus that autonomous vehicles should be allowed on the roads, provided their performance was no more than some (small) amount worse than human-driven cars, but that it was expected that their performance was actually better than human-driven cars. In this case, a test of non-inferiority could be conducted and the sample size planned accordingly (Chow et al., 2008).

Yet even these results are optimistic. We have intentionally framed this analysis to calculate the fewest number of miles that would need to be driven to demonstrate statistically significant differences between autonomous vehicles and human drivers. First, developers are likely to improve autonomous vehicles as testing reveals shortcomings of the technology. The performance of the vehicle will change between the start and the end of a multiyear testing time frame, hopefully for the better. However, this may mean that still more miles are required to prove safety because the technology will have changed.

Second, recall that we treat H as a known benchmark against which we can do a one-sample test. Yet H is not a known benchmark for three key reasons. First, the performance of human drivers in 2013 or any particular year is not the benchmark of concern. The concern is whether autonomous vehicle performance is better than human driver performance, and a single year's failure data is only an estimate of the true rate of human driver failures. Second, injuries and crashes are significantly underreported and there is conflicting evidence about the rate of underreporting. Experiments in which injuries and crashes are accurately recorded could yield different rates. Third, human driver performance is changing. Motor vehicle fatality rates have fallen in the past several decades. In 1994, there were 1.73 fatalities per 100 million miles compared with 1.09 fatalities per 100 million miles in 2013 (Bureau of Transportation Statistics, 2015). Much of the decline can be attributed to improvements in vehicle designs (Farmer and Lund, 2015), which could continue. Thus, the benchmark of human driver performance is a moving target. So, if we compare the performance of human drivers against autonomous vehicles in some time frame, there is uncertainty about whether the comparison would hold moving into the future. For all of these reasons, it would be appropriate to treat H as uncertain and use a two-sample hypothesis test, which would require even more failures to be observed and miles to be driven. This suggests it is not possible to drive our way to answers to one of the most important policy questions about autonomous vehicles: Are they safer than human drivers?

5. Discussion and conclusions

This report frames three different questions about the number of miles that autonomous vehicles would have to be driven as a method of statistically demonstrating their reliability. We lay out the formulas for answering these questions and present results for fully autonomous vehicles that can serve as a reference for those interested in statistically testing their reliability.

Table 1
Examples of miles and years needed to demonstrate autonomous vehicle reliability.

	How many miles (years[a]) would autonomous vehicles have to be driven…	Benchmark failure rate		
		(A) 1.09 fatalities per 100 million miles?	(B) 77 reported injuries per 100 million miles?	(C) 190 reported crashes per 100 million miles?
Statistical question	(1) without failure to demonstrate with 95% confidence that their failure rate is at most…	275 million miles (12.5 years)	3.9 million miles (2 months)	1.6 million (1 month)
	(2) to demonstrate with 95% confidence their failure rate to within 20% of the true rate of…	8.8 billion (400 years)	125 million (5.7 years)	51 million (2.3 years)
	(3) to demonstrate with 95% confidence and 80% power that their failure rate is 20% better than the human driver failure rate of…	11 billion (500 years)	161 million (7.3 years)	65 million (3 years)

[a] We assess the time it would take to complete the requisite miles with a fleet of 100 autonomous vehicles (larger than any known existing fleet) driving 24 h a day, 365 days a year, at an average speed of 25 miles per hour.

Table 1 provides illustrative results from our analysis. The three numbered rows show sample results for each of our three statistical questions about the miles needed to demonstrate safety. Sample results are shown for each of three benchmark failures rates noted in the lettered columns. These correspond to human-driven (A) fatality rates, (B) reported injury rates, and (C) reported crash rates. The results also show in parentheses the number of years it would take to drive those miles with a fleet of 100 autonomous vehicles driving 24 h a day, 365 days a year, at an average speed of 25 miles per hour. For example, one can ask, “How many miles (years) would autonomous vehicles have to be driven (row 2) to demonstrate with 95% confidence their failure rate to within 20% of the true rate of (column A) 1.09 fatalities per 100 million miles?” The answer is 8.8 billion miles, which would take 400 years with such a fleet.

The results show that autonomous vehicles would have to be driven hundreds of millions of miles and sometimes hundreds of *billions* of miles to demonstrate their reliability in terms of fatalities and injuries. Under even aggressive testing assumptions, existing fleets would take tens and sometimes hundreds of years to drive these miles—an impossible proposition if the aim is to demonstrate their performance *prior* to releasing them on the roads. Only crash performance seems possible to assess through statistical comparisons of this kind, but this also may take years. Moreover, as autonomous vehicles improve, it will require many millions of miles of driving to statistically verify changes in their performance.

Our results confirm and quantify that developers of this technology and third-party testers cannot drive their way to safety. Our findings support the need for alternative methods to supplement real-world testing in order to assess autonomous vehicle safety and shape appropriate policies and regulations. These methods may include but are not limited to accelerated testing (Nelson, 2009), virtual testing and simulations (Chen and Chen, 2010; Khastgir et al., 2015; Olivares et al., 2015); mathematical modeling and analysis (Hojjati-Emami et al., 2012; Kianfar et al., 2013); scenario and behavior testing (California Department of Motor Vehicles, 2015; Sivak and Schoettle, 2015); and pilot studies (ANWB, 2015), as well as extensive focused testing of hardware and software systems.

And yet, even with these methods, it may not be possible to establish the safety of autonomous vehicles prior to making them available for public use. Uncertainty will remain. This poses significant liability and regulatory challenges for policymakers, insurers, and developers of the technology, and it would be a cause for concern among the public. It also suggests that pilot studies may be an essential intermediate step for understanding autonomous vehicle performance prior to widespread use. Such pilot studies would need to involve public-private partnerships in which liability is shared among developers, insurers, the government, and consumers.

Simultaneously, the technology will evolve rapidly, as will the social and economic context in which it is being introduced. In fast-changing contexts such as these, regulations and policies cannot take a one-shot approach. Therefore, in parallel to creating new testing methods, it is imperative to begin developing approaches for planned adaptive regulation (Eichler et al., 2015; Walker et al., 2010).

Such regulation is designed from the outset to generate new knowledge (e.g., through pilot studies), review that knowledge (e.g., through scheduled safety review boards), and use that knowledge to evolve with the technology (e.g., by modifying safety requirements). This can help society better harness the benefits and manage the risks of these potentially transformative technologies.

Conflict of interest

During the development of this report and at time of publication, Nidhi Kalra’s spouse served as Principal Engineer for Google’s driverless car project. Neither her spouse nor Google had any influence on the report.

Acknowledgments

We would like to acknowledge Anita Chandra and James Anderson for their advice and support, and for helping shepherd this work through the publication process efficiently. We thank Lance Tan and Jeffrey Chappelle for their administrative support and edits on the draft, and Maria Vega for her excellent editorial help. We are enormously grateful to Carolyn Rutter at RAND and Constantine Samaras at Carnegie Mellon University for their insightful reviews. We thank Bonnie Ghosh-Dastidar for connecting Nidhi and Susan for this work. Funding for this study was provided by philanthropic contributions from RAND supporters and income from operations. We especially appreciate the support from the RAND Justice, Infrastructure, and Environment Advisory Board and the National Cooperative Highway Research Program.

References

Anderson, James M., Kalra, Nidhi, Stanley, Karlyn D., Sorensen, Paul, Samaras, Constantine, Oluwatola, Oluwatobi A., 2016. Autonomous Vehicle Technology: A Guide for Policymakers. RAND Corporation, RR-443-2-RC, Santa Monica, Calif. As of January 24, 2016 http://www.rand.org/pubs/research_reports/RR443-2.html.

ANWB, 2015. Experiments on Autonomous and Automated Driving: An Overview 2015. As of March 3, 2016. http://www.anwb.nl/bestanden/content/assets/anwb/pdf/over-anwb/persdienst/rapport_inventarisatie_zelfrijdende_auto.pdf.

Blanco, Myra, Atwood, Jon, Russell, Sheldon, Trimble, Tammy, McClafferty, Julie, Perez, Miguel, 2016. Automated Vehicle Crash Rate Comparison Using Naturalistic Data. Virginia Tech Transport Institute. As of March 3, 2016 http://www.apps.vtti.vt.edu/PDFs/Automated%20Vehicle%20Crash%20Rate%20Comparison%20Using%20Naturalistic%20Data_Final%20Report_20160107.pdf.

Blincoe, Lawrence, Miller, Ted R., Zaloshnja, Eduard, Lawrence, Bruce A., 2015. The Economic and Societal Impact of Motor Vehicle Crashes 2010 (Revised). National Highway Traffic Safety Administration, DOT HS 812 013, Washington, D.C. As of March 3, 2016 http://www-nrd.nhtsa.dot.gov/pubs/812013.pdf.

Bureau of Transportation Statistics, 2014a. Motorcycle Rider (Operator) Safety Data, Table 2–22. U.S. Department of Transportation, Washington, D.C. As of March 3, 2016: <http://www.rita.dot.gov/bts/sites/rita.dot.gov.bts/les/publications/national_transportation_statistics/html/table_02_22.html>.

Bureau of Transportation Statistics, 2014b. Occupant and Non-Motorist Fatalities in Crashes by Number of Vehicles and Alcohol Involvement (Updated July 2014), Table 2–20, Washington, D.C.: U.S. Department of Transportation, 2014b.

Bureau of Transportation Statistics, 2015. Motor Vehicle Safety Data, Table 2–17. Research and Innovative Technology Administration, U.S. Department of Transportation, Washington, D.C. As of March 3, 2016: <http://www.rita.dot.gov/bts/sites/rita.dot.gov.bts/les/publications/national_transportation_statistics/html/table_02_17.html>.

California Department of Motor Vehicles, 2015. Express Terms Title 13, Division 1, Chapter 1 Article 3.7—Autonomous Vehicles, 2015. As of March 3, 2016. <https://www.dmv.ca.gov/portal/wcm/connect/ed6f78fe-fe38-4100-b5c2-1656f555e841/AVExpressTerms.pdf?MOD=AJPERES>.

California Vehicle Code, 2012. California Vehicle Code, Division 16.6. As of November 25, 2013: <http://leginfo.legislature.ca.gov/faces/codes_displaySection.xhtml?lawCode=VEH&sectionNum=38750>.

Chen, Suren, Chen, Feng, 2010. Simulation-based assessment of vehicle safety behavior under hazardous driving conditions. J. Transport. Eng. 136 (4), 304–315. http://dx.doi.org/10.1061/(ASCE)TE.1943-5436.0000093. As of March 3, 2016:.

Chow, Shein-Chung, Shao, Jun, Wang, Hansheng, 2008. Sample Size Calculations in Clinical Research. Chapman & Hall/CRC Biostatistics Series, Boca Raton, Fla.

Darby, John L., 2010. Sample Sizes for Con Dence Limits for Reliability. Sandia National Laboratories, SAND2010-0550, Albuquerque, N.M. As of March 3, 2016 http://prod.sandia.gov/techlib/access-control.cgi/2010/100550.pdf.

DeGroot, Morris H., 1986. Probability and Statistics. Addison-Wesley Publishing Company Inc, Reading, Mass.

Eichler, H.G., Baird, L.G., Barker, R., Bloechl-Daum, B., Borlum-Kristensen, F., Brown, J., Chua, R., Del Signore, S., Dugan, U., Ferguson, J., Garner, S., Goettsch, W., Haigh, J., Honig, P., Hoos, A., Huckle, P., Kondo, T., Le Cam, Y., Leufkens, H., Lim, R., Longson, C., Lumpkin, M., Maraganore, J., O'Rourke, B., Oye, K., Pezalla, E., Pignatti, F., Raine, J., Rasi, G., Salmonson, T., Samaha, D., Schneeweiss, S., Siviero, P.D., Skinner, M., Teagarden, J.R., Tominaga, T., Trusheim, M. R., Tunis, S., Unger, T.F., Vamvakas, S., Hirsch, G., 2015. From adaptive licensing to adaptive pathways: delivering a flexible life-span approach to bring new drugs to patients. Clin. Pharmacol. Ther. 97 (3), 234–246. http://dx.doi.org/10.1002/cpt.59. March As of April 1, 2016.

Fagnant, Daniel J., Kockelman, Kara, 2015. Preparing a nation for autonomous vehicles: opportunities, barriers and policy recommendations. Transport. Res. Part A: Policy Pract. 77, 167–181. http://dx.doi.org/10.1016/j.tra.2015.04.003. July As of March 3, 2016.

Farmer, Charles M., Lund, Adrian K., 2015. The effects of vehicle redesign on the risk of driver death. Traffic Injury Prevent (October). As of March 3, 2016 http://www.iihs.org/bibliography/topic/2073.

Gomes, Lee, 2014. Hidden Obstacles for Google's Self-Driving Cars: Impressive Progress Hides Major Limitations of Google's Quest for Automated Driving. Massachusetts Institute of Technology. As of March 3, 2016 https://www.technologyreview.com/s/530276/hidden-obstacles-for- googles-self-driving-cars/.

Google, 2015. Google Self-Driving Car Testing Report on Disengagements of Autonomous Mode. As of March 3, 2016. <https://www.dmv.ca.gov/portal/wcm/connect/dff67186-70dd-4042-bc8c-d7b2a9904665/GoogleDisengagementReport2014-15.pdf?MOD=AJPERES>.

Hojjati-Emami, Khashayar, Dhillon, Balbir, Jenab, Kouroush, 2012. Reliability prediction for the vehicles equipped with advanced driver assistance systems (ADAS) and passive safety systems (PSS). Int. J. Indust. Eng. Comput. 3 (5), 731–742. http://dx.doi.org/10.5267/j.ijiec.2012.08.004. As of April 1, 2016.

Kaleto, Helen, Winkelbauer, David, Havens, Chris, Smith, Michael, 2001. Advancements in Testing Methodologies in Response to the FMVSS 201U Requirements for Curtain-Type Side Airbags. Society of Automotive Engineers International. Technical Paper 2001–01- 0470. As of April 1, 2016. http://dx.doi.org/10.4271/2001-01-0470.

Khastgir, Siddartha, Birrell, Stewart A., Dhadyalla, Gunwant, Jennings, Paul A., 2015. Development of a Drive-In Driver-In- e-Loop Fully Immersive Driving Simulator for Virtual Validation of Automotive Systems. Paper presented at IEEE 81st Vehicular Technology Conference, Glasgow, Scotland, May 11–14, 2015. As of April 1, 2016: http://dx.doi.org/10.1109/VTCSpring. 2015.7145775.

Kianfar, R., Falcone, P., Fredriksson, J., 2013. Safety veri cation of automated driving systems. IEEE Intell. Transp. Syst. Mag. 5 (4), 73–86. http://dx.doi.org/10.1109/MITS.2013.2278405. Winter As of April 1, 2016:.

Kleyner, Andre, 2014. How stress variance in the automotive environment will a ect a 'true' value of the reliability demonstrated by accelerated testing. Soc. Automotive Eng. Int. 7 (2), 552–559. http://dx.doi.org/10.4271/2014-01-0722. As of March 3, 2016.

Martz, H.F., Waller, R.A., 1982. Bayesian Reliability Analysis. John Wiley & Sons, Hoboken, N.J.

Mathews, Paul., 2010. Sample Size Calculations: Practical Methods for Engineers and Scientists. Mathews, Malnar, and Bailey Inc, Fairport Harbor, Ohio.

Misra, Krishna.B. (Ed.), 2008. Handbook of Performability Engineering. Springer, New York.

National Highway Traffic Safety Administration, 2011. Traffic Safety Facts: Crash Stats. National Center for Statistics and Analysis, Washington, D.C. DOT HS 811 449. As of March 3, 2016. <http://www-nrd.nhtsa.dot.gov/pubs/811449.pdf>.

National Highway Traffic Safety Administration, 2016. Traffic Safety Facts, A Brief Statistical Summary: Critical Reasons for Crashes Investigated in the National Motor Vehicle Crash Causation Survey. National Center for Statistics and Analysis, U.S. Department of Transportation, Washington, D.C. DOT HS 812 115. As of March 3, 2016. <http://www-nrd.nhtsa.dot.gov/pubs/812115.pdf>.

Nelson, Wayne B., 2009. Accelerated Testing: Statistical Models, Test Plans, and Data Analysis. John Wiley & Sons, Hoboken, N.J.

O'Connor, Patrick., Kleyner, Andre., 2012. Practical Reliability Engineering. John Wiley & Sons, Hoboken, N.J.

Olivares, Stephanie, Rebernik, Nikolaus., Eichberger, Arno., Stadlober, Ernst., 2015. Virtual stochastic testing of advanced driver assistance systems. In: Schulze, Tim., Müller, Beate., Meyer, Gereon. (Eds.), Advanced Microsystems for Automotive Applications 2015: Smart Systems for Green and Automated Driving. Springer, New York.

Sivak, Michael, Schoettle, Brandon, 2015. Should We Require Licensing Tests and Graduated Licensing for Self-Driving Vehicles? Transportation Research Institute, University of Michigan. Technical Report UMTRI-2015-33.

Society of Automotive Engineers International, 2014. Automated Driving: Levels of Driving Automation Are Defined in New SAE International Standard J3016. As of March 3, 2016: <http://www.sae.org/misc/pdfs/automated_driving.pdf>.

U.S. Department of Transportation, 2015. Distracted Driving 2013: Summary of Statistical Findings, DOT HS 812 132.

Ulm, Kurt., 1990. A simple method to calculate the confidence-interval of a standardized mortality ratio (SMR). Am. J. Epidemiol. 131 (2), 373–375. Feb.

Urmson, Chris, 2016. "Hands Off: e Future of Self-Driving Cars" testimony before the Senate Committee on Commerce, Science and Technology hearing, Washington, D.C. March 15, 2016. As of March 22, 2016: <http://www.commerce.senate.gov/public/_cache/files/5c329011-bd9e-4140-b046-a595b4c89eb4/BEADFE023327834146FF4378228B8CC6.google-urmson-testimony-march152016.pdf>.

Walker, Warren E., Marchau, Vincent A.W.J., Swanson, Darren, 2010. Addressing deep uncertainty using adaptive policies: introduction to section 2. Technol. Forecast. Soc. Chang. 77 (6), 917–923. http://dx.doi.org/10.1016/j.techfore.2010.04.004. July As of March 3, 2016:.