A Data-Driven Perspective on Human Error as a Cause of Fatal Traffic Accidents

Leon Trochelmann [*1] **Jonathan Ranck** [*2] **Paul-Henrik Heilmann** [*3] **Filippo Albani** [*4]

Abstract

Stepping into the era of automated vehicles, new questions arise around human error and responsibility on the road. This study introduces a new data-driven perspective on human error based on the FARS2021National dataset of fatal U.S. traffic accidents. Our analysis reveals a high occurrence of drivers within these accidents being subject to preventable human error, also demonstrating strong correlations with driver demographics and the time of the accident. The high incidence of preventable risk factors in fatal accidents underscores the necessity of new solutions to enhance road safety in the United States.

1. Introduction

Human error is a contested term with several accepted definitions (Reason, 2000; Woods et al., 2017; Strauch, 2017). In the context of motor vehicle driving it has traditionally been used to differentiate from mechanical error (Stanton & Salmon, 2009), but for example, in the modern context of automated motor vehicles, we may also be interested in considering errors that are specific to human drivers.
In this work, we introduce a new data-driven definition of preventable human error, allowing us to differentiate these human errors from those that may be consequences of the distinct challenges of the driving scenario (Guanetti et al., 2018). We base our definition on the FARS2021National dataset (NHTSA) of fatal traffic accidents in the United States of America and analyse the incidence of preventable human error within it.
Our results show that at least 38.70% of drivers involved in fatal traffic accidents in the United States were subject to preventable human error. We also observe a strong correlation of such human error with the demographics of the involved drivers and the time of day of the accident.
Ultimately, these results reflect a significant number of potentially avoidable traffic fatalities, presenting evidence of the shortcomings of human drivers that motivate improvements of road traffic safety in the United States.

[*]Equal contribution [1]Matrikelnummer 6646000, leon.trochelmann@student.uni-tuebingen.de, MSc Machine Learning [2]Matrikelnummer 6230070, jonathan.ranck@student.uni-tuebingen.de, BSc Physics [3]Matrikelnummer 16648314, paul-henrik.heilmann@student.uni-tuebingen.de, MSc Machine Learning [4]Matrikelnummer 6638113, filippo.albani@student.uni-tuebingen.de, MSc Physics.

Project report for the "Data Literacy" course at the University of Tübingen, Winter 2023/24 (Module ML4201). Style template based on the ICML style files 2023.

2. Data and Methods

We introduce the dataset we considered for our analysis and show how it leads to our definition of preventable human error.

2.1. Dataset

We base our methods on the FARS2021National dataset by the United States of America's National Highway Traffic Safety Administration (NHTSA). The FARS (Fatality Analysis Reporting System) aims to register all fatal car accidents that occur in the USA in any given year.
FARS captures a large amount of information about each accident, including information about all persons and vehicles involved in the accident, as well as the circumstances of the accident itself. Figure 1 illustrates how the recorded cases are distributed between the different states in the U.S. in the year 2021. The statistics for the U.S. population in that year

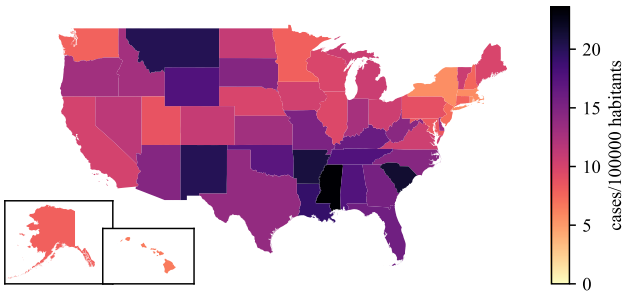


Figure 1. Density of cases of road traffic fatalities in the United States of America in the year 2021 grouped by state and displayed by population density.

were obtained from the U.S. Census Bureau (United States Census Bureau). With a total of 39508 cases, the FARS dataset presents a rich body of information upon which we

base our analysis.
Zoning in on fatal car accidents furthermore reduces ambiguity over what constitutes a crash.

2.2. Defining Preventable Human Error

To analyse preventable human error, we define it based on a selection of variables that we consider to both reflect error and arise from the human condition specifically. This selection notably excludes risk factors that may not be preventable, such as disabilities. It furthermore disregards risk factors that do not arise from the human condition, such as the weather.
Finally, we choose a conservative approach of only selecting risk factors which are generally known to cause many traffic accidents. While these risk factors are not guaranteed to be the true underlying cause of the accident, they can generally be accepted as likely causes.
First, we consider a driver record as speeding if FARS2021National reflects that the driver went over the speed limit. Second, we consider it to be driving without a license if the driver did not have a valid license. Third, we consider driving under the influence, subject to the judgement of law enforcement. Figure 2 illustrates the total number of records that have these attributes and the overlap between the variables.

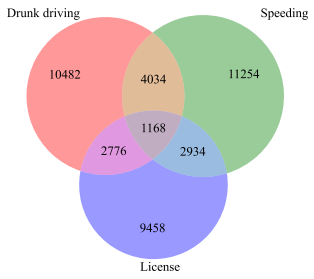


Figure 2. driver counts for the three variables we consider as preventable human error in FARS2021National.

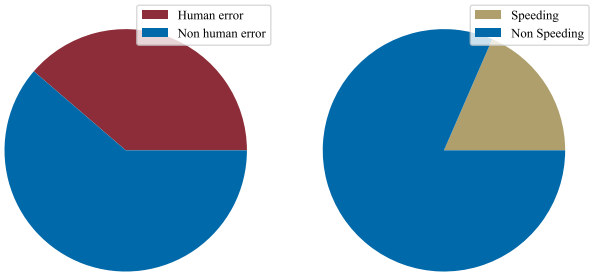
Finally, we introduce preventable human error as a new variable that is TRUE if one of these indicators is given and FALSE otherwise. We will consider only this definition of preventable human error in our analysis.
The dataset includes several records where the value of one or more of these variables is missing, of which 3558 were related to speed, 0 to drinking and 2410 to unlicensed driving. For our analysis, we treated all such records as not speeding/drinking/unlicensed driving to preserve the conservative nature of our estimate.

3. Experiments and Results

To demonstrate the significance of preventable human error and show how it interacts with other variables, we analyse rates of occurrence within the total set of drivers involved in fatal accidents. First, we analyse the total occurrence of preventable human error in our dataset. Second, we investigate how preventable human error is correlated with other variables.
The FARS datasets contain a wide range of variables, many of which have no significant correlation with human error. We present several significant correlations our analysis yielded, but we do not claim completeness. A full correlation analysis with every variable in the FARS is beyond the scope of this work.

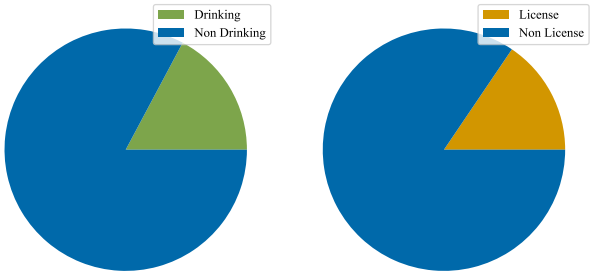
3.1. Occurrence

Figure 3 illustrates the general occurrence of preventable human error.



(a) Incidence of drivers subject to preventable human error.

(b) Incidence of drivers subject to speeding.



(c) Incidence of drivers subject to drinking.

(d) Incidence of drivers subject to unlicensed driving.

Figure 3. Incidence of preventable human errors as a percentage out of the total set of drivers.

Figure 3a demonstrates approximately 38.70% of all drivers in our dataset to be subject to preventable human error. For the individual variables, we observe ca. 18.48% of cases to be subject to speeding, 17.21 % to drinking and

15.53% to unlicensed driving as seen in figures 3b, 3c and 3d respectively.

3.2. Correlations

We find that preventable human error is strongly correlated with the demography of the drivers and the time of the accident.
We begin by demonstrating the demographics of preventable human error by conditioning on age and sex. We plot the percentages of driver records subject to human errors as an overlapping bar chart with binned age groups for the sake of readability, displayed in figure 4. We also show the amounts of records attributed to each age group for context.

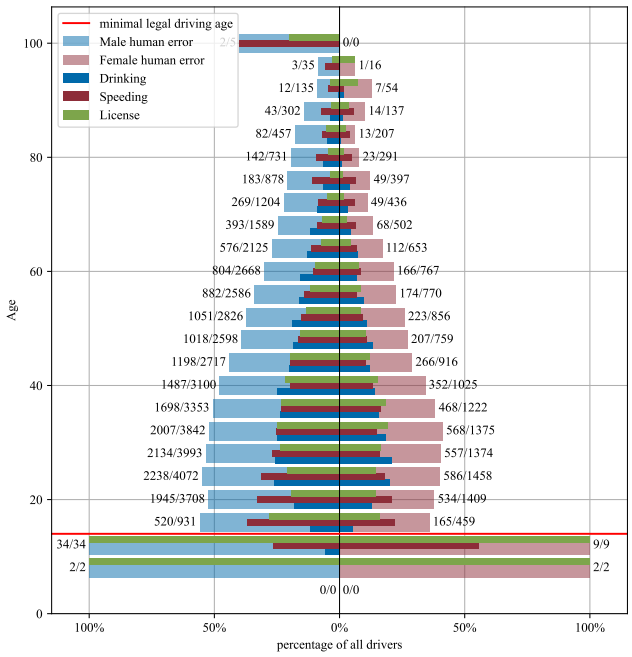


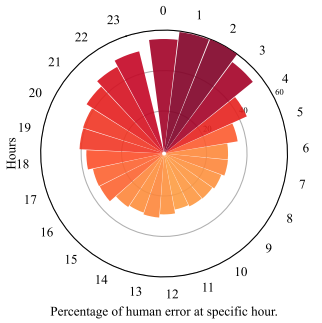
Figure 4. Correlation of preventable human error with age and sex as a percentage of the total set of drivers. The male population is displayed on the left and the female population is on the right. Each type of error is displayed in a different colour. The bin size is 4. The numbers of records considered for each age group are explicitly displayed at the end of the bars, as a fraction of records with preventable human error out of all driver records. The absolute minimum legal driving age of 14 is highlighted in red and located precisely between the respective bins.

Trivially, the rate of preventable human error is 1 for all age groups below the absolute minimum legal driving age in the USA. We observe that the incidence of preventable human error is overall higher in the male population. Additionally, the incidence of preventable human error decreases with older age groups.
We observe different peaks in the percentage and total number of drivers subject to preventable human errors in the male and female populations respectively. The percentage of preventable human error reaches its peak in the 14 to 17 age group for the male population and in the 30 to 33 age group for the female population. Meanwhile, the total number of driver records indicating preventable human error has a peak in the 22 to 25 age group for males and a significantly lower peak again in the 30 to 33 age group for females.
There are significantly fewer total drivers subject to preventable human error in the female population as compared to the male population in nearly every age group. Due to the low number of total cases in the 98 to 101 age group, it may be considered an outlier.
Regarding the individual error variables, we observe a gradual decrease in the percentage of speeding in older age groups, whereas the percentage of drinking displays a rise and fall. The percentage of unlicensed driving behaves somewhat irregularly, as it displays a series of increases and decreases throughout the age ranges, but ultimately also exhibits the mass of its distribution lying within the younger age groups.

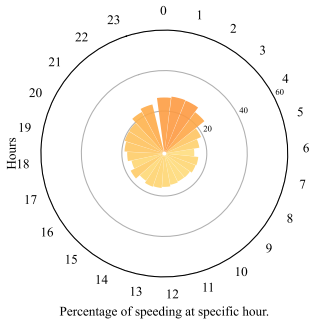
The time of day also has a significant correlation with preventable human error. We plot the percentage of drivers subject to preventable human error conditioned on the time of the accident in hourly intervals, displayed in 5. We see a significant increase in drunk driving during the late evening and early night, see figure 5c, whereas speeding and and unlicensed driving are more evenly distributed.
To examine the overall time distribution of drivers that were subject to preventable human error, we plot the number of drivers and also expand our time intervals to consider the days of the week, seen in figure 6.
The distribution of drivers subject to preventable human error illustrated in 6a exhibits a clear similarity to the number of total drivers displayed in figure 6b for reference. Specifically, the distribution exhibits a concentration in the later hours of the day and over the weekend.
Comparing the distributions side by side, one can observe that preventable human error is overall less present in the evenings on weekdays but equally concentrated towards the weekend. The distribution of preventable human error also displays less density during the mornings than the overall distribution.

4. Discussion & Conclusion

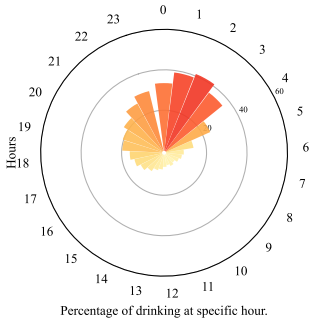
In this work, we have introduced a data-driven definition of preventable human error and demonstrated its high occurrence based on a vast dataset of fatal car accidents.
Our results show that preventable risk factors have a significant incidence in fatal motor vehicle accidents. They also show that the incidence is strongly correlated with the



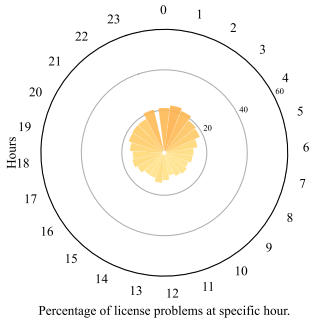
(a) Incidence of preventable human error by the time of day.



(b) Incidence of speeding by the time of day.

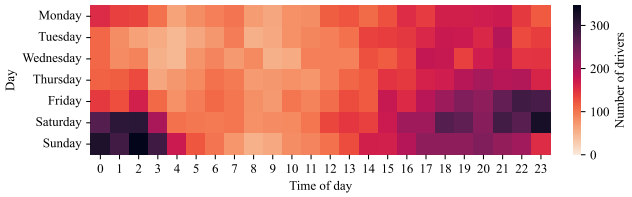


(c) Incidence of drunk driving by the time of day.

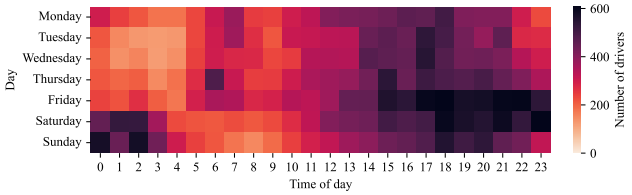


(d) Incidence of unlicensed driving by the time of day.

Figure 5. The correlation of time of day in hours with preventable human error as a percentage of the set of all drivers at an hour of the day. The scale is presented through three equidistant circles, each representing a 20% distance.



(a) Number of drivers involved in fatal accidents and subject to preventable human error.



(b) Numbers of total drivers involved in fatal accidents.

Figure 6. The total amounts of drivers involved in fatal accidents plotted by the time of day and day of the week concerning the crash. To visually compensate for the overall difference in the number of drivers, we use the same colour bar over a different range of drivers for the two plots respectively

demography of drivers and the time of the accident.

Due to our conservative approach to variable selection, these rates of occurrence can also be seen as lower bounds on the true rate of irresponsible driving in fatal car accidents, as we did not consider difficult-to-assess factors like road rage and reckless driving.

Our analysis is based on traffic data that stems exclusively from the USA in the year 2021. We conducted our analysis at the federal level, such that the statistics over individual states may differ. Furthermore, different rates of occurrence may come about in other countries or at different points in time.

Overall, the presence of significant quantities of preventable human error motivates changes in the motorway traffic system in the USA, whether these solutions be automotive, legislative or societal. Code, full results and documentation are available at https://github.com/PaulHenrik/DataLiteracy_FARS.

Contribution Statement

Leon conceptualised the project, assisted with visualisations and authored the report. Jonathan was highly involved in the data exploration, merging, filtering and visualisation. Paul worked on exploring the data and creating visualisations. Filippo worked on data preparation and was the primary editor of the report. These statements were written individually.

References

Guanetti, J., Kim, Y., and Borrelli, F. Control of connected and automated vehicles: State of the art and future challenges. *Annual reviews in control*, 45:18–40, 2018.

NHTSA. Fatality analysis reporting system. https://www.nhtsa.gov/research-data/fatality-analysis-reporting-system-fars. Accessed: 2024-01-29.

Reason, J. Human error: models and management. *Bmj*, 320(7237):768–770, 2000.

Stanton, N. A. and Salmon, P. M. Human error taxonomies applied to driving: A generic driver error taxonomy and its implications for intelligent transport systems. *Safety Science*, 47(2):227–237, 2009.

Strauch, B. *Investigating human error: Incidents, accidents, and complex systems*. CRC Press, 2017.

United States Census Bureau. County population totals and components of change: 2020-2022. https://www.census.gov/data/datasets/time-series/demo/popest/2020s-counties-total.html. Accessed: 2024-01-29.

Woods, D., Dekker, S., Cook, R., Johannesen, L., and Sarter, N. *Behind human error*. CRC Press, 2017.