Evaluating Autonomous Vehicles' Safety Benefits in Mixed Autonomy Scenarios

Data Collection

What data will you collect or create?

We will collect or create four types of data in this project: (1) reports on the results of our research, (2) computer code to run simulations that validate our proposed AV movement and traffic incident models, (3) results (e.g., expected traffic statistics) from the simulation experiments, and (4) existing datasets on traffic maps from Pittsburgh that will be used in the algorithm simulations.

Team members may also make informal notes as needed to conduct the research, which will be stored in their private notebooks. Data will be stored in standard formats, e.g., csv, latex, Word, or json. We may convert it to proprietary formats as needed to run simulation experiments. The volume of data will be determined as needed as the research progresses.

How will the data be collected or created?

We will periodically write reports on our research that will lead to papers submitted to journals or conferences. Each member of the team will update the reports weekly to include their progress. The report format will be determined by the author as appropriate, and all reports will be organized by week so that everyone can track each other's progress. More informal progress may also be summarized in Powerpoint slide decks, which will later be used to create public presentations of our results.

Computer code to run simulations that validate our proposed algorithms will be structured according to the judgement of the programmer. We expect to use Python-based simulations throughout the project, and we will adhere to standard best practices in software engineering to design and develop our codebase.

Experimental results data will be organized into folders corresponding to different experiments. We will annotate them as appropriate and create summary documents to track the setup and results of each simulation.

The existing datasets that we will use will be stored in the formats provided by our deployment partners or other data owners, in the case of publicly available data. We will keep copies of the data in its original format and alter the format as needed to run simulation experiments and analysis.

We plan to store all data used, collected, or created in this project in appropriate version-controlled repositories shared with all members of the team (e.g., Google Drive folders or Overleaf documents for project reports, Github repos for simulation code and results).

Documentation and Metadata

What documentation and metadata will accompany the data?

We will create readme files and line-by-line annotations for all of our simulation code that fully document how to use it. We will maintain records of the experiments run for at least one year after the project lifetime, including the conditions for each experiment and a summary of the findings.

These records will allow secondary users to understand the experiments we have already run and design new ones. Data that we receive from our deployment partners will be accompanied with summary documents describing the types of data collected and any other relevant details. We will write this documentation so that it is understandable to all team members, to facilitate future projects that might build on this work as well as onboarding of new team members.

Our written reports will often incorporate summaries of our experiments, code, and deployment partner datasets. We plan to prepare one or two research papers that describe our work in detail and would be accessible to secondary users and others unfamiliar with our work.

Ethics and Legal Compliance

How will you manage any ethical issues?

We do not foresee any ethical issues from our planned research. We do not plan to directly collect any data from experiment participants, or from any of our deployment partners. Ethical implications of our results will be incorporated into our policy guidelines in consultation with our deployment partner, the Southwestern Pennsylvania Commission.

How will you manage copyright and Intellectual Property Rights (IP/IPR) issues?

Ownership of all data created under this project will be governed by the existing Safety21 agreement with the Department of Transportation. We plan to open source our simulation code and make it publicly available, to stimulate related research. Papers that are published based on our work will be subject to the copyright policies of the publishers (e.g., IEEE, ACM). Subject to these restrictions, we plan to open source all of our findings, e.g., by publishing paper preprints on free public repositories like arXiv and releasing public versions of our simulation code on sites like Github.

Storage and Backup

How will the data be stored and backed up during the research?

We plan to use cloud storage for the data used in this research. Cloud providers have extensive backup and recovery mechanisms that we can rely on to protect our data. We will use existing CMU accounts (e.g., on Google Drive and Overleaf) for this storage and do not anticipate any monetary charges arising from data storage needs. We will discuss additional storage and backup mechanisms needed for any proprietary datasets provided by our deployment partners at the time they provide the dataset(s).

How will you manage access and security?

With the exception of publicly released code and research reports, all data for this project will be shared only with team members (and appropriate representatives from our deployment partners). We will use password-protected cloud servers (e.g., Google Drive, Github) to store all of the data and ensure that external entities or users cannot access it. We do not plan to work with any confidential data. We will discuss additional access control mechanisms needed for any proprietary datasets provided by our deployment partners at the time they provide the dataset(s).

Selection and Preservation

Which data are of long-term value and should be retained, shared, and/or preserved?

Open-sourced code and published papers or technical reports will remain available for several years after the conclusion of the project, in order to enable follow-on research by ourselves and others. Detailed records of our experimental results will be kept as long as they are needed for our follow-on research. Since we will store all of our data on cloud providers, we do not expect to encounter challenges in indefinitely storing data.

What is the long-term preservation plan for the dataset?

The two types of data that have long-term value are our simulation code and technical reports or research papers resulting from the project. We plan to store these on standard repositories (e.g., Github for code; arXiv for papers), which will be free of charge. Published papers will also be made available by the relevant publishers (e.g., ACM and IEEE), according to their policies.

Data Sharing

How will you share the data?

We will publicly share our simulation code and technical reports by posting them on well-known code and paper repositories. All other data will be restricted to project team members.

Are any restrictions on data sharing required?

We do not expect any significant restrictions on data sharing. We may wait to post some technical reports or simulation code until our research papers based on those results are accepted for publication and the results are finalized.

Responsibilities and Resources

Who will be responsible for data management?

The PI, Carlee Joe-Wong, will have overall responsibility for data management. All team members will be asked to store and document the data they create in accordance with this plan.

What resources will you require to deliver your plan?

We do not anticipate additional resource requirements for data management.
