$$r(t) = \|v_{\mathrm{des}}\| - \|v_{\mathrm{des}} - v(t)\| - \alpha \sum_i \max\left[h_{\max} - h_i(t), 0\right]$$

Figure 1: The reward function for the system in question consists of three terms. The first term v_{des} is a positive constant that rewards the agent for longer simulation episodes - discouraging vehicle collisions, which terminate simulation runs early. The second term penalizes the agent when the instantaneous overall system velocity $v(t)$ differs from the desired system velocity v_{des}. Finally, the third term sums over each subscribed Connected Autonomous Vehicle and adds a penalty whenever this vehicle is too close to the vehicle immediately in-front - a characteristic known to trigger stop-and-go traffic waves. More details are provided below in the section 'Defined Performance Metrics'.

1 System Details

1.1 System Owner

This may be the designer deploying the system, a larger agency or body, or some combination of the two. The entity completing the report should also be indicated.

This system was developed by the Project Flow core team members, with all deployment, infrastructure, and ongoing management taking managed by Caltrans.

1.2 Dates

The known or intended timespan over which this reward function & optimization is active.

The system discussed here was trained in simulation during 2020, using empirical hyper-parameters (such as inflow traffic rates) collected during 2019. The RL policy was deployed in the real world on a trial basis on the 1st of Jan, 2021, and is presently undergoing initial real-world evaluation and validation.

1.3 Feedback & Communication

Contact information for the designer, team, or larger agency responsible for system deployment.

Any correspondence should be directed to test@example.ca.gov.

1.4 Other Resources

Where can users or stakeholders find more information about this system? Is this system based on one or more research papers?

More information about this specific system can be found in the paper [1], as well as in the associated project website.

General information about the project flow simulation environment can be found in [2] or on the project website and associated GitHub repository.

2 Optimization Intent

2.1 Goal of Reinforcement

A statement of system scope and purpose, including the planning horizon and justification of a data-driven approach to policy design (e.g. the use of reinforcement learning or repeated retraining). This justification should contrast with alternative approaches, like static models and hand-designed policies. What is there to gain with the chosen approach?

The system in question is designed to dissipate stop-and-go traffic waves caused by merging off the California State Route 24 (CA-24) freeway onto Telegraph Avenue in the North Oakland / South Berkeley metropolitan area.

This is achieved by the coordinated actions of any subscribed Connected Autonomous Vehicles (CAVs) operating along the freeway segment in question, acting to 'shepherd' non-autonomous vehicles into patterns of traffic which can locally buffer against stop-and-go traffic waves.

Eligible CAVs, when entering the freeway zone of interest, communicate over the 4G/5G cell network with the central controller hub to 'subscribe' to the traffic management policy, which then sends real-time recommendations to these vehicles about lane selection and preferred acceleration/braking profiles.

The RL policy is trained using a discrete-time road network simulation, with simulation runs lasting 3600s (one hour), and individual steps of 0.2s, giving 1800 steps per full simulation episode. The simulated road network consists of an 800m stretch of the CA-24 freeway containing a single off-ramp merging lane. These temporal and spatial planning horizons were selected because they were deemed large enough to allow emergence of typical driving dynamics based on the average safe following distance between vehicles and driver reaction times along comparable freeway offramps, based on state and federal records of past traffic behavior.

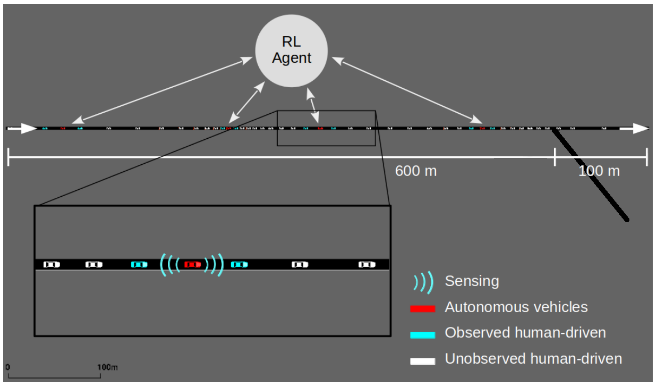

Figure 2: A central RL controller attempts to mitigate stop-and-go traffic waves caused by vehicles entering the freeway *via* on-ramps.

As of entry 0.3, it was found that the planning horizon for the system was too short. Following consultation with Caltrans, it was found that increasing the horizon from 500m to 800m would provide a significant increase in simulation performance without exhausting computational resources. Any future changes in computational capabilities will be documented here and compared in light of prior modeling choices and stakeholder commitments.

Simplistic microscopic traffic analysis models preclude the possibility of stable congestion patterns in open road topologies. However, as any driver can attest, these traffic patterns are ubiquitous on many road systems today. Instead, the presence of these traffic patterns in real-world networks is typically attributed to perturbations from bottleneck structures which can be difficult to capture in theoretical analyses (such as lane closures, road works, road debris, *etc*). [1] The ad-hoc nature of these perturbations means that modelling and planning for their occurrence within classical control frameworks may be difficult, motivating more flexible approaches such as Deep Reinforcement Learning.

RL may be indicated in this situation, compared to static supervised ML models, due to the fact that it inherently encompasses multiple types of feedback through the environment specification. For instance, in the case of CA-24, RL may help mitigate the observed phenomenon of excessive traffic on residential streets near highway intersections that is induced by apps like Google Maps and Waze. In the interest of recommending perceived shortcuts to individual human drivers, these apps have in fact been known to induce overload on smaller roadways, generating unnecessary stoppage and possible gridlock. In the case of Los Gatos (where this phenomenon has been previously recorded), the city's Parks and Public Works Director noted that "The apps are not able to respond fast enough to the overload they have created on the roadways" [3]. RL may make real-time monitoring and control of the CA-24 offramp possible, mitigating induced overload effects and stabilizing feedback between traffic behavior and road infrastructure.

2.2 Defined Performance Metrics

A list of "performance metrics" included explicitly in the reward signal, the criteria for why these metrics were chosen, and from where these criteria were drawn (e.g. government agencies, domain precedent, GitHub repositories, toy environments). Performance metrics that are used by the designer to tune the system, but not explicitly included in the reward signal should also be reported here.

The reward signal optimized by this system consists of three performance metrics, outlined in fig. 1. These terms are;

- $\|v_{\text{des}}\|$ - the desired system-level velocity in m/s. This is a positive constant reward to penalize prematurely terminated simulation rollouts caused by vehicle collisions. For the simulated experiments described here, $v_{\text{des}} = 25\text{m/s} = 90\text{kmph} \approx 55\text{mph}$.

- $-\|v_{\text{des}} - v(t)\|$ - the absolute difference between the desired system level velocity and the actual instantaneous system-level velocity in m/s. A non-zero difference incurs a cost for the RL agent.

- $-\alpha \sum_i \max\left[h_{\max} - h_i(t), 0\right]$ - this term sums over each Autonomous Vehicle in the purview of the RL agent, and accrues a cost whenever that vehicle's instantaneous time headway (gap in seconds to the vehicle ahead) is too small (*i.e.* lower than $h_{\max}$). The sum of all headway costs is scaled by a gain factor α. For the simulated experiments described here, $h_{\max} = 1\text{s}$ and $\alpha = 0.1$.

2.3 Oversight Metrics

Are there any additional metrics not included in the reward signal but relevant for vendor or system oversight (e.g. performance differences across demographic groups)? Why aren't they part of the reward signal, and why must they be monitored?

Several other performance metrics are not included in the reward function, but are analysed for the purpose of evaluating the system performance:

- Absolute temporal vehicle density (or *throughput*) - the number of vehicles exiting the controlled region the road network, measured in vehicles/hr. A larger vehicle flow-through rate compared to baseline is seen as a positive effect (assumed to correlate with a decrease in stop-and-go traffic waves, and to indicate that the road network is functioning efficiently).
- Absolute spatial vehicle density (or *network congestion*) - the number of vehicles within a fixed region of the road network, measured in vehicles/m. A larger number of vehicles present on the roadway is seen as a negative effect, indicating increased likelihood of stoppage.
- The average velocity of vehicles in the system. Higher vehicle velocities are seen as a positive effect.
- The average time vehicles spend within a given region of the system. Lower average time is seen as a positive effect.
- The maximum time any vehicle spent within a given region of the system over the course of an experimental evaluation of the system. Lower maximum time is seen as a positive effect.
- Simulated episode length. Simulation episodes are cut short whenever a collision occurs between vehicles - as such, longer episodes are seen as a positive effect.

In addition, the qualitative nature of stop-and-go traffic waves (size in terms of space and time duration and severity as measured by the average space-time slope of a wave) is assessed using microscopic vehicle space-time graphs such as those shown in fig. 3.

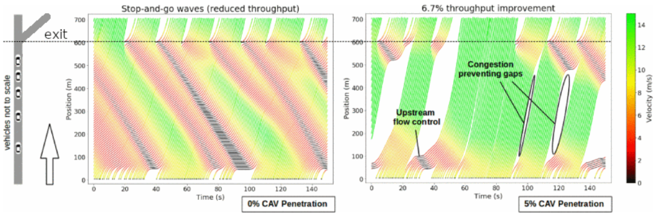

Figure 3: Space-time microscopic vehicle trace graphs such as these allow qualitative assessment of the system-level state of simple road networks at a glance. Here, stop-and-go traffic waves can be seen as red or black diagonal lines propagating through the traffic flow.

2.4 Known Failure Modes

A description of any prior known instances of "reward hacking" or model misalignment in the domain at stake, and description of how the current system avoids this.

Sim-to-real dynamics misalignment. The emergent dynamics of the simulated model and environment could potentially be misaligned with real-world dynamics (a 'sim-to-real' policy transfer problem). This failure mode was exhibited in the initial version of the system (as documented in change log entry v0.3) - the initially designed planning horizon was found to be too short (500m), which did not allow space for the requisite stop-and-go traffic dynamics to emerge around the freeway entry point. This issue was brought to light because the performance of the system in terms of average reward once deployed was not as high as predicted in simulation, triggering a technical review of the system. Two possible solutions were considered - (a) re-visiting the parameter distributions used for the IDM (which controls the non-automated vehicles in the simulation environment), (b) or adjusting the planning horizon. In a review with Caltrans engineers and the system designers, it was deemed that the IDM parameter distributions were in fact representative of the target section of CA-24, based on empirical data from 2019, and so the planning horizon was expanded from 500m to 800m. Thus far, since this updated version of the system was deployed, the sim-to-real performance gap issue appears to have been resolved, suggesting the updated planning horizon adequately allows the simulated dynamics to reflect real-world dynamics.

Selective behavior throttling. The system was found to decrease throughput and increase congestion for diesel-powered vehicles. This feature was first documented in change log entry v0.3, but not labeled as a known failure mode until entry v0.6. This failure mode was exhibited in all previous versions of the system documented originally in log v0.1 It was highlighted following citizen complaints. No solution has been implemented as of entry v0.6. Two solutions have been proposed - (a) a city ordinance limiting diesel-powered vehicle travel on residential streets in the adjoining city of Emeryville (at present out of scope for the system), (b) or adjusting the policy parameters' training environment so that the controller behaves appropriately around diesel-powered vehicles in the future. This resolution is pending the recommendation of the Diesel Vehicle Taskforce to be presented at a future regular meeting.

3 Institutional Interface

3.1 Deployment Agency

What other agency or controlling entity roles, if any, are intended to be subsumed by the system? How may these roles change following system deployment?

The system in question is developed by the Project Flow core development team. The deployment infrastructure and ongoing management are operated by the California Department of Transportation (Caltrans), in coordination with the city departments of Oakland and Berkeley.

Our RL system is designed to manage the flow of traffic immediately surrounding an exit point off the CA-24 freeway (see fig. 4) - as such, the system operates in a functionally similar way to traffic control signals that are sometimes used to regulate vehicles entering or exiting freeways.

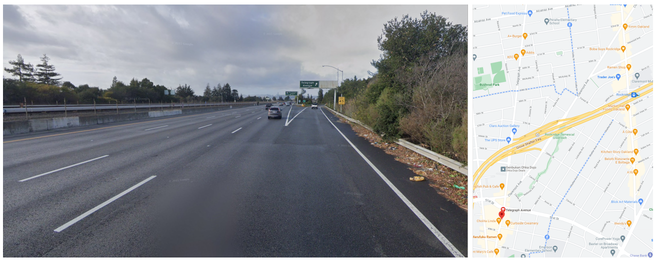

Figure 4: The freeway exit from CA-24 to telegraph avenue, which this system is designed to manage.

This system simultaneously encroaches upon, and expands the capabilities of Caltrans. As the sensing infrastructure, computational capacity, and deployed RL software is centrally managed by a control facility operated by Caltrans, this system serves to provide both (a) an enhanced level of road surveillance for the relevant freeway section, through the remote sensing capabilities of subscribed CAVs, as well as (b) a 'control lever' through which Caltrans can actually influence traffic operations in and around the relevant freeway section (although this influence is delegated to an RL policy).

3.2 Stakeholders

What other interests are implicated in the design specification or system deployment, beyond the designer? What role will these interests play in subsequent report documentation? What other entities, if any, does the deployed system interface with whose interests are not intended to be in scope?

By automating the partial management of this section of the freeway via the RL environment framing and policy structure, the system serves to remake direct oversight of the road network on a new layer of abstraction. This indirection raises potential risks from inappropriate information flow, in particular monopolization of the freeway offramp by the RL controller. Monopolization may generate unstable dynamics leading up to or following the planning horizon (i.e. CA-24 freeway lanes and gridlock along Telegraph Avenue), or unequal access for road users whose behaviors are harder to anticipate (such as public buses, groups of motorcycles, bicycles, and pedestrians experiencing homelessness), or whose dynamics do not conform to the modelling assumptions of the system designers (e.g. heavy vehicles with atypical acceleration profiles). To counter these risks, new coordination is required between Caltrans and the city departments of Oakland and Berkeley.

Diesel vehicle drivers. As of entry 0.6, the behavior throttling generated by the RL controller was found to change the traffic patterns of diesel vehicles. A *Diesel Vehicle Taskforce* was created to help organize this constituency and identify needed changes to the controller to sufficiently reduce inappropriate behavior throttling.

Nearby homeowners. As of entry 0.6, residents of the adjoining city of Emeryville had complained to the Public Works Departments of Berkeley and Oakland about the new traffic flows indirectly generated by the RL controller. Following the creation of the *Diesel Vehicle Taskforce* these departments will coordinate with Emeryville officials about the recommended changes to the controller and monitor future complaints as needed.

3.3 Explainability & Transparency

Does the system offer explanations of its decisions or actions? What is the purpose of these explanations? To what extent is the policy transparent, i.e. can decisions or actions be understood in terms of meaningful intermediate quantities?

The system contains no explicit explainability modules. However, Figure 1 makes makes the reward function transparent in terms of meaningful simulation parameters. Expressed in non-technical language, these are *continuous avoidance of vehicle collisions*, *consistent vehicle velocity*, and *steady following distance.* These terms, and corresponding parameters, are regularly shared with the city departments of Oakland and Berkeley per stakeholder agreements.

3.4 Recourse

Can stakeholders or users contest the decisions or actions of the system? What processes, technical or otherwise, are in place to handle this?

As of v0.2, the city departments of Oakland and Berkeley can review and contest system performance every six weeks, per agreement with Caltrans.

4 Implementation

4.1 Reward Details

How was the reward function engineered? Is it based on a well-defined metric? Is it tuned to represent a specific behavior? Are multiple terms scaled to make one central loss, and how was the scaling decided?

As recorded in Figure 1, the reward function combines well-defined metrics for avoiding collisions, steady speeds, and maintaining safe following distances to other vehicles. Reward parameters were agreed on by stakeholders according to specific desired behaviors.

4.2 Environment Details

Description of states, observations, and actions with reference to planning horizon and hypothesized dynamics/impact. What dynamics are brought into the scope of the optimization via feedback? Which dynamics are left external to the system, as drift? Have there been any observed gaps between conceptualization and resultant dynamics?

The RL observation space consists of traffic features which are locally observed by subscribed CAVs (see fig. 2). That is, for each subscribed CAV i, the RL agent observes the speeds $v_{i,\mathrm{lead}}$, $v_{i,\mathrm{lag}}$ and bumper-to-bumper time headways $h_{i,\mathrm{lead}}$, $h_{i,\mathrm{lag}}$ of the vehicles immediately preceding and following the CAV, as well as the currently occupied lane l_i, and ego speed v_i of the CAV itself. The action space for the RL policy consists of a vector of bounded acceleration recommendations a_i, one for each subscribed CAV i. Importantly, although the policy may request a certain acceleration a_i, the system design is such that the CAV locally maintains control authority, so the actions may not necessarily be followed exactly - for this reason they are referred to as action recommendations. This effect is modelled by adding stochastic Gaussian action noise in the simulation environments.

As the number of subscribed CAVs can vary over time, the RL policy is designed with a fixed upper number of subscribed CAVs n. When an $n + 1^{\mathrm{th}}$ CAV attempts to subscribe to the RL system when entering the freeway region, the subscription offer is declined, and the vehicle enters a queue. When the next CAV exits the controlled freeway region, the subscription-waiting CAV at the front of the queue is then subscribed into the policy. When there are less than n CAVs subscribed, zero-padding is used in the RL observation vector.

4.3 Measurement Details

How are the components of the reward and observations measured? Are measurement techniques consistent across time and data sources? Under what conditions are measurements valid and correct? What biases might arise during the measurement process?

Observations are measured using a mix of LiDAR, radar, and camera sensors on fleet vehicles. These measurements are compared across vehicles and over time to ensure consistency. Observed metrics are validated against simluation parameters for following distance and expected velocity according to the terms of the reward function.

Sensor bias may arise due to blocked cameras, extreme weather, or other unanticipated situations in which one or more sensors are blocked. A mix of sensor types is used across vehicles to help ensure redundancy in case of malfunction.

4.4 Algorithmic Details

The key points on the specific algorithm(s) used for learning and planning. This includes the form of the policy (e.g. neural network, optimization problem), the class of learning algorithm (e.g. model-based RL, off-policy RL, repeated retraining), the form of any intermediate model (e.g. of the value function, dynamics function, reward function), technical infrastructure, and any other considerations necessary for implementing the system. Is the algorithm publicly documented and is code publicly available? Have different algorithms been used or tried to accomplish the same goal?

The RL system uses a Deep Neural Network policy. Specifically, the controller is a diagonal Gaussian Multi Layer Perceptron policy with three hidden layers of size 32 with rectified linear unit nonlinearities and bias terms. The Gaussian diagonal variance terms are learned as part of the policy parameters.

The RL policy was trained in simulation using the Trust Region Policy Optimization (TRPO) policy gradient RL algorithm [4]. The discount factor was set as $\gamma = 0.999$, which corresponds to a reward half-life of $\sim$ 700 steps, or slightly over 2 minutes. The TRPO step size was set at 0.01.

4.5 Data Flow

How is data collected, stored, and used for (re)training? How frequently are various components of the system retrained, and why was this frequency chosen? Could the data exhibit sampling bias, and is this accounted for in the learning algorithm? Is data reweighted, filtered, or discarded? Have data sources changed over time?

Per v0.2, every system component is retrained at least every six weeks, corresponding to public performance reports. Specific system components pertaining to perception, motion planning, control, or route navigation are retrained at the discretion of Caltrans. As of v0.6 (latest version), no known issues with sampling bias have arisen, and data sources have not been changed since the specification proposed and simulated in v0.1.

4.6 Limitations

Discussion and justification of modeling choices arising from computational, statistical, and measurement limitations. How might (or how have) improvements in computational power and data collection change(d) these considerations and impact(ed) system behavior?

As of v0.3, the planning horizon was updated from 500m to 800m. This was not motivated by technical limitations, but by observed discrepancies between observed system performance and predictions from simulation training.

No fundamental changes in computational power or data collection have been made as of v0.6 (latest version).

Future improvements in vehicle sensing may permit an even longer planning horizon (1000m or more). This may result in improved oversight metrics on throughput and network congestion. Caltrans officials have determined this change would not result in improvements on defined performance metrics as of v0.6 (latest version).

4.7 Engineering Tricks

RL systems are known to be sensitive to implementation tricks that are key to performance. Are there any design elements that have a surprisingly strong impact on performance? For example, state-action normalization, hard-coded curricula, model-initialization, loss bounds, or more?

As of v0.4, the system was observed to conduct "behavior throttling" when in the vicinity of diesel-powered vehicles. No engineering tricks were implemented to fix this performance discrepancy, but new oversight metrics for diesel-powered vehicle throughput were added for purpose of future monitoring and reporting. No other surprising performance impacts have been noted as of v0.6 (latest version).

5 Evaluation

5.1 Evaluation Environment

How is the system evaluated (and if applicable, trained) prior to deployment (e.g. using simulation, static datasets, etc.)? Exhaustive details of the offline evaluation environment should be provided. For simulation, details should include description or external reference to the underlying model, ranges of parameters, etc. For evaluation on static datasets, considering referring to associated documentation (e.g. Datasheets *[5]).*

The RL model is developed in the Project Flow AV simulation test-bed.

For training the RL agent, non-autonomous vehicles are modelled using the Intelligent Driver Model (IDM) [6] - a microscopic traffic simulation car-following model in which the accelerations of a human vehicle α are a function of the bumper-to-bumper time headway h_α, velocity v_α, and relative velocity with the preceding vehicle $\Delta v = v_l - v_\alpha$, *via* the following equation;

$$f(h_\alpha, v_l, v_\alpha) = a \left[1 - \left(\frac{v_\alpha}{v_0} \right)^\delta - \left(\frac{s^*(v_\alpha, \Delta v_\alpha)}{h_\alpha} \right)^2 \right],$$

where s^* is the desired headway of the vehicle, calculated according to

$$s^*(v_\alpha, \Delta v_\alpha) = \max \left(0, v_\alpha T + \frac{v_\alpha \Delta v_\alpha}{2\sqrt{ab}} \right),$$

where s_0, v_0, T, a, b are given parameters empirically calibrated to match typical traffic in the highway region of interest, and to simulate stochasticity in driver behaviour, exogenous Gaussian noise calibrated to match findings in [7] is added to accelerations.

5.2 Offline Evaluations

Present and discuss the results of offline evaluation. For static evaluation, consider referring to associated documentation (e.g. Model Cards *[8]). If applicable, compare the behaviors arising from counterfactual specifications (e.g. of states, observations, actions).*

As of v0.3, planning horizon was updated and expanded to 800m from 500m. Previous fleet behaviors were found to deviate from desired thresholds for following distance and constant acceleration/deceleration.

As of v0.6 (latest version), the system behaviors were found to lie within desired thresholds on key performance metrics.

5.3 Evaluation Validity

To what extent is it reasonable to draw conclusions about the behavior of the deployed system based on presented offline evaluations? What is the current state of understanding of the online performance of the system? If the system has been deployed, were any unexpected behaviors observed?

The RL system was initially designed in a simulation environment with a closed network topology (a ring road with length 1400m, 700m of which is controlled by the RL agent. This was done as a means to test the robustness of the policy architecture and training paradigm - a type of transfer learning (from a theoretically simple closed topology to the more complex open topology). With this counterfactual environment specification, it was observed that the policy performs well, and after transfer to the open topology environment there was little decrease in policy performance, providing confidence in the policy design choices.

5.4 Performance standards

What standards of performance and safety is the system required to meet? Where do these standards come from? How is the system verified to meet these standards?

The 'gold standard' for this problem is defined as the average condition of the traffic before and after the CA-24 exit prior to implementation of the RL system. In this domain, this standard is not actually 'optimal' behaviour, in the sense that the RL controller has the capability to out-perform this existing standard of performance.

6 System Maintenance

6.1 Reporting Cadence

The intended timeframe for revisiting the reward report. How was this decision reached and motivated?

The most important commitment is for a regular set of meetings to be scheduled between relevant city departments and the Caltrans officials tasked with overseeing the RL controller. The cadence and structure of meetings should reflect the policy priorities of the city departments, particularly the Public Works Department (including the Transportation Division that oversees traffic engineering) and the Housing and Community Services Department (which administers a subsidized transportation program for seniors and disabled persons). In this way, the gains in traffic efficiency and safety made possible through deep RL's flexibility can be leveraged in the interests of those municipalities most likely to be impacted by the intervention.

As of entry 0.2, the cadence of meetings was decided as approximately every six weeks between Caltrans and the Public Works Departments of Berkeley and Oakland. This timeframe was motivated by the policy priorities of both city departments with the consent of Caltrans. Meetings may deviate from this schedule slightly (e.g. twice per quarter / eight times per year) at the discretion of both city departments, but will not be held without all three agencies present.

Documentation of the planned meeting schedule for the year–and any break in this schedule due to special events, municipal elections, or holidays–should be the first item included in the changelog of the updated reward report.

As of entry 0.2 and per agreement with key development parties, the model is to be retrained every six weeks following each regular meeting. Training data is to be updated at the discretion of Caltrans, and shared with Public Works departments at each regular meeting.

At a minimum, these meetings should review the real-world implementation to confirm that the RL controller is operating safely and as intended by Caltrans per the environment specification. Caltrans officials will also document shifts in the oversight metrics that, while not explicitly factored into the reward signal, were deemed of interest prior to implementation (related to *throughput* and *congestion*. This documentation may be included in subsequent updates to the reward report at the discretion of Caltrans, wherever it is deemed relevant for oversight of the RL controller.

Of special importance is the need to reinterpret public works priorities in light of the real-world implementation. For example, Berkeley's subsidized transportation program might be reevaluated in light of system effects, or expanded to cover a wider group of stakeholders. Caltrans will invite comment on the system implementation in light of city departments' ex ante assumptions about the traffic domain. This bureaucratic oversight may be complemented by requests for public comment from citizens, civil society advocates, and other members of the public at the discretion of the city governments of Berkeley and Oakland. At the discretion of Caltrans, records of this public comment may be included in subsequent reward reports where deemed relevant for understanding changes to the planning horizon, environment specification, or list of known failure modes.

6.2 Update Triggers

Specific events (projected or historic) significant enough to warrant revisiting this report, beyond the cadence outlined above. Example triggers include a defined stakeholder group empowered to demand a system audit, or a specific metric (either of performance or oversight) that falls outside a defined threshold of critical safety.

The most important ground for review of this deployed RL system will be any vehicle collisions or near-miss incidents in the controlled region of the CA-24 freeway. This is because such events may compromise the entire motive of the RL controller in the first place. These may serve as grounds for changing the specification or altering the institutional agreements between Caltrans and the Public Works Departments of both municipalities, at their own discretion.

At the discretion of Caltrans, any shift in the oversight metrics deemed pressing or significant may also trigger a new reward report. Here and below, the threshold for "significant" is to be decided by agreement between Caltrans and Public Works Departments. The updated report should note the magnitude of the observed shift, the specification already deployed at the time the shift was observed, and Caltrans officials' own best evaluation of why the shift occurred. If possible, the officials should propose alternative specifications (or roll back to a prior one) that would mitigate the shift or at least bring it into alignment with the documented priorities of the Public Works Departments. These alternatives could then be interpreted and evaluated at the next regular meeting according to institutional prerogatives.

Other review grounds include:

- Discrepancies between prior reward reports and system behavior as observed in the real world.
- Discrepancies between prior reward reports and system behavior as observed in simulated environments of interest to policymakers.
- A security breach resulting in loss of data or other infrastructure components that violates the terms of agreement between relevant agencies.
- Substantial changes in the distribution of CAVs using the CA-24 freeway exit - including changes in the capabilities of the vehicles (e.g. increased levels of autonomy) and/or changes in group statistics (e.g. make or model, absolute number, temporal distribution, *etc.*)
- A new mode of transport with significant observed throughput at the CA-24 offramp, but unknown distribution of traffic behaviors.
- Any change in the schedule of meetings between Caltrans and Public Works Departments corresponding to regular future updates of reward reports.
- A new ordinance (passed by either city) or statute (adopted by Caltrans) that alters the design assumptions of the deployed specification as documented in prior reward reports.
- A significant shift in the personnel makeup of the Public Works Departments of Berkeley or Oakland.
- A plebiscite leading to basic reforms of municipal governance in either city.

6.3 Changelog

Descriptions of updates and lessons learned from observing and maintaining the deployed system. This includes when the updates were made and what motivated them in light of previous reports. The changelog comprises the central difference between reward reports and other forms of machine learning documentation, as it directly reflects their intrinsically dynamic nature.

- v0.1 (08/Oct/2020) - Initial reward report was drafted based on the system developed and tested in simulation only.
- v0.2 (01/Jan/2021) - System is deployed to the real-world environment in a ongoing evaluation capacity, reward report updated to reflect this fact. Reporting cadence decided to be every six weeks based on agreement between Caltrans and the city departments of Oakland and Berkeley. Intended feedback section was updated to include plans for regular model retraining and data sharing agreements. No other substantial changes.
- v0.3 (14/Feb/2021) - Planning horizon for the system was updated from a 500m stretch of freeway to a 800m stretch of freeway. The planning horizon was updated because the deployed system's performance was not in line with predictions from simulation training. Consultation with Caltrans traffic engineers and the system developers suggested that the stretch of highway used in simulation may be too short to sufficiently exhibit typical driving dynamics induced by the IDM, and it was suggested to extend the planning horizon and re-train the agent, before re-deploying the policy. Failure modes section was updated to reflect these observations. Computation footprint section was updated to reflect this change.
- v0.4 (01/April/2021) - Caltrans officials reported to Public Works Departments of Berkeley and Oakland that the system undergoes "behavior throttling" when interacting with diesel-powered vehicles within 800m of the CA-24 offramp. It was decided to add new metrics for diesel-powered vehicle throughput and congestion to the list of oversight metrics. Due to no observed increase in accidents or driver complaints, no changes to performance metrics or environment specification were made at this time.
- v0.5 (15/May/2021) - Meeting was convened according to the regular schedule. Oversight metrics were presented and discussed. Officials

noted a significant decline in diesel-powered vehicle throughput and congestion on the CA-24 offramp. No other substantial changes.

- v0.6 (12/June/2021) - Emergency meeting was called by the Public Works Departments of Berkeley and Oakland in response to a rapid uptick in complaints from residents about the growing frequency of diesel-powered vehicles driving through residential areas in the vicinity of Emeryville, which is located west of the CA-24 exit. Residents have complained about a slight uptick in air pollution and large increase in noise pollution due to the vehicles. Caltrans officials consulted the changelog of previous reward reports and determined that diesel-driven vehicles were being excessively disincentivized from driving on the CA-24 offramp due to behavior throttling. It was decided to convene a *Diesel Vehicle Taskforce* to examine the problem and communicate with drivers of heavy vehicles to identify what new incentives or adjustments were needed to the controller to reduce behavior throttling beneath the desired threshold. It was agreed that the Diesel Vehicle Taskforce issue a report recommending these changes no later than two regular meetings from the present time. Stakeholders section was updated to name these distinct groups (diesel vehicle drivers, nearby homeowners) and reflect these changes.

References

[1] A. R. Kreidieh, C. Wu, and A. M. Bayen, "Dissipating stop-and-go waves in closed and open networks via deep reinforcement learning," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, 2018, pp. 1475–1480.

[2] C. Wu, A. R. Kreidieh, K. Parvate, E. Vinitsky, and A. M. Bayen, "Flow: A modular learning framework for mixed autonomy traffic," *IEEE Transactions on Robotics*, pp. 1–17, 2021.

[3] J. Peterson, "Google apps causing gridlock in downtown Los Gatos," https://www.mercurynews.com/2018/06/01/google-apps-causing-gridlock-for-downtown-los-gatos/, 2018, [Online; accessed 2-January-2022].

[4] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *International conference on machine learning*. PMLR, 2015, pp. 1889–1897.

[5] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii, and K. Crawford, "Datasheets for datasets," *Communications of the ACM*, vol. 64, no. 12, pp. 86–92, 2021.

[6] M. Treiber, A. Hennecke, and D. Helbing, "Congested traffic states in empirical observations and microscopic simulations," *Physical review E*, vol. 62, no. 2, p. 1805, 2000.

[7] M. Treiber and A. Kesting, "The intelligent driver model with stochasticity-new insights into traffic flow oscillations," *Transportation research procedia*, vol. 23, pp. 174–187, 2017.

[8] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, "Model cards for model reporting," in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 220–229.