

BayesianURQ: Bayesian Approach for Unknown Risks Quantification on Highly Automated Driving

Masanori Yokota ¹⁾

1) NTT DATA Automobility Research Center Co., Ltd.
3-1-9 Shin-Yokohama, Kohoku, Yokohama, 222-0033, Japan (E-mail: yokota-masanori@zipc.com)

KEY WORDS: Electronics and control, Autonomous driving system, System Engineering, SOTIF [E1]

Highly automated driving systems operate in complex and open-ended environments, making it difficult to identify all hazardous contexts and triggering conditions in advance. ISO 21448 (SOTIF) requires consideration of risks arising not only from known scenarios but also from unknown scenarios. However, a concrete methodology for quantitatively estimating risks caused by unknown conditions, particularly unknown unknowns, has not yet been established. Conventional approaches often rely on accumulating large-scale real-world driving data and statistically estimating accident rates. For rare events such as fatal accidents, however, hundreds of millions to tens of billions of miles may be required to demonstrate safety, leading to the well-known “billion-mile problem.” Moreover, such demonstrations may need to be repeated after system updates.

This paper proposes BayesianURQ, a Bayesian framework for quantitatively estimating residual risks caused by unknown triggering conditions and unknown contexts. System risk is decomposed into losses (L), hazards (H), contexts (S), component failures (F), and triggering conditions (C). Unknown classes (C_K) and (C_U) are explicitly introduced into this decomposition. Because the conditional failure or hazard probability under an unknown condition cannot be directly evaluated, BayesianURQ conservatively assumes that an unknown triggering condition always causes a failure, *i.e.* $p(F|C_U, S) = 1$, and that an unknown context always leads to a hazard, *i.e.* $p(H|S_U) = 1$. Consequently, the corresponding residual risks can be upper-bounded by the occurrence rates $p(C_U|S)$ and $p(S_U)$ respectively.

For unknown triggering conditions, an obstacle-detection failure in a forward obstacle-avoidance context is considered as an example. Known object categories may include vehicles, motorcycles, bicycles, and pedestrians, while objects outside these categories are temporarily assigned to an unknown category C_U . Assuming mutually exclusive categories, their occurrence probabilities are modeled by a multinomial distribution with a Dirichlet prior. After observing category counts, the posterior distribution is again Dirichlet *i.e.* $\pi | \mathcal{D} \sim \text{Dir}(\alpha_1 + n_1, \dots, \alpha_K + n_K, \alpha_U + n_U)$. Even when no unknown-category sample is observed, *i.e.* $n_U = 0$, the posterior assigns a non-zero occurrence probability to C_U . Thus, unlike the empirical estimate, the proposed approach avoids concluding that the residual risk is zero merely because an unknown category has not yet been observed. Posterior expectations or conservative upper posterior quantiles, such as the 95th percentile, can be used depending on the required level of conservativeness.

For unknown contexts, the occurrence rate is modeled per unit driving distance. Let $\lambda_U [km^{-1}]$ denote the occurrence rate of unknown contexts, and assume that occurrences follow a Poisson process. The posterior distribution is $\lambda_U | D \sim \text{Gamma}(\alpha + n_U, \beta + d)$ with a Gamma prior where d is the observed driving distance and n_U is the number of observed unknown contexts. When $n_U = 0$, a finite residual risk is still obtained. For example, the Jeffreys prior corresponds to $\text{Gamma}(0.5, 0)$, yielding a posterior mean of $1/2d$. The required observation distance for demonstrating a target residual-risk level can therefore be explicitly quantified.

A major feature of BayesianURQ is that unknown-risk estimation can be separated from the performance of a specific automated-driving system. Under the conservative assumption that unknown conditions always cause adverse outcomes, $p(C_U|S)$ and $p(S_U)$ depend mainly on the operational environment rather than on a specific system implementation. Therefore, provided that the operational-environment distribution remains equivalent, accumulated observations may be reused after system updates and may also be obtained from external sources such as infrastructure sensing or other operators. This potentially converts part of the billion-mile problem from repeated system-level validation into an operational-environment observation problem.

Several challenges remain. The distribution of unknown conditions may vary across regions, seasons, and time, requiring segmentation or sequential Bayesian updating with distribution-drift handling. Results also depend on the choice of prior distribution and upper posterior quantile, requiring a balance between statistical objectivity, conservativeness, and practical feasibility. In addition, the framework assumes that unknown conditions can be detected in the observation data; conditions that are difficult for humans or classifiers to identify may violate this assumption. Finally, intentionally generated adversarial conditions are not necessarily governed by stationary environmental occurrence probabilities and should be addressed separately as cybersecurity threats.

BayesianURQ provides a quantitative framework for incorporating unknown unknowns into safety arguments without treating unobserved risks as zero. Future work will evaluate the approach using public traffic datasets and investigate practical methods for prior selection, unknown-condition detection, and distribution-drift management.