Statistical Perspectives on Reliability of Artificial Intelligence Systems

Yili Hong

Joint Research Conference 2024

June 18, 2024

joint work with Jiayi Lian, Li Xu, Jie Min, Yueyao Wang, Laura J. Freeman, and Xinwei Deng

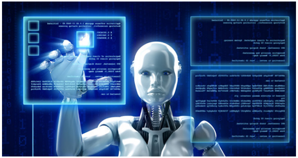
Department of Statistics, Virginia Tech, Blacksburg, VA, USA

Outline

- Background, AI applications, and AI reliability
- AI reliability framework
- The roles of traditional reliability
- Challenges in statistical analysis of AI reliability
- Accelerated tests and AI reliability improvements
- Concluding remarks

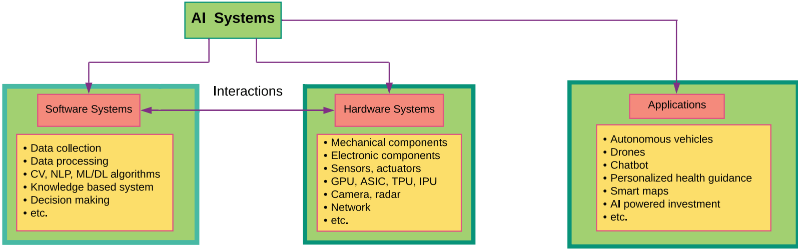
Background

- Artificial intelligence (AI) systems have become increasingly common and the trend will continue.
- To allow for safe, effective, and massive deployment of AI systems, the reliability of such systems need to be addressed.
- The main goal of this talk is to provide statistical perspectives on the reliability of AI systems.
- We also review recent developments in modeling and analysis of AI reliability, and outline statistical research challenges in the area for statisticians.



AI Applications and AI System Framework

- Application areas include information technology, transportation, government, healthcare, finance, and manufacturing. Examples including self-driving cars, drones, robots, and chatbots.
- Autonomous systems are the main applications. Typical examples include autonomous vehicles, industrial robotics, aircraft autopilot systems, and unmanned aircraft.

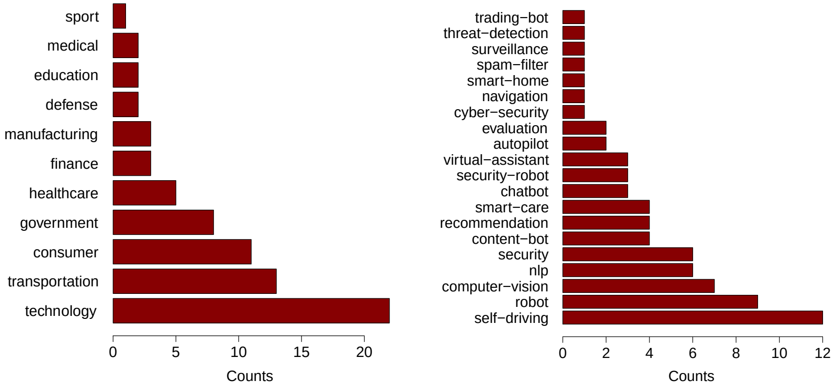


The Importance of AI Reliability

- Failures of AI systems can lead to economic loss and even, in extreme cases, lead to loss of life.
- For example, a failure in the autopilot system of an autonomous car can lead to an accident with loss of life.
- Thus, reliability is critical, especially for autonomous systems.
- From another point of view, the large-scale deployment of AI technologies requires public trust.
- AI reliability falls within the larger scope of AI safety and AI assurance.

AI Incidence Examples

- Cases reported on website “AI Incidence Database.”
- Among the 126 incidents reported up to date, 72 incidents are related to reliability events.



(a) AI Application Sectors (b) AI Systems/Technologies

Examples of Algorithms and Failure Causes

- We also notice that 29 incidents involve deaths or injuries among those 72 events.

(a) Algorithms

(b) Failure Causes

AI Reliability Framework – The SMART Framework

- **Structure of the system:** Understanding the system structure is a fundamental step in the study of AI reliability.
- **Metrics of reliability:** Appropriate metrics need to be defined for AI reliability so that data can be collected accordingly and reliability can be tracked over time.
- **Analysis of failure causes:** Conducting failure analysis to understand how the system fails (i.e., failure modes) and what factors affect the reliability.
- **Reliability assessments:** Reliability assessments of AI systems include reliability modeling, estimation, and prediction.
- **Test planning:** Test planning methods are needed for efficient reliability data collection.

System Structures

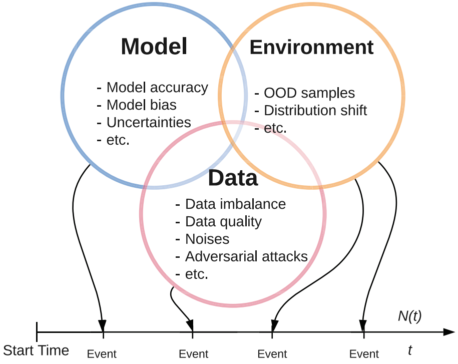
- For AI systems (e.g., autonomous vehicles), we can conceptually divide the overall system into hardware systems and software systems.
- The core of many AI software systems is machine learning/deep learning (ML/DL) algorithms and other rule-based algorithms.
- Hardware reliability is well studied and there are mature methods for testing and assessing hardware reliability.
- Compared to hardware reliability, software reliability is typically more difficult to test.
- In addition, there are two other factors to consider as the AI system structure: hardware-software interaction, and the interaction of the system to the operating environment.

Definition of AI Reliability and Metrics

- Reliability is the probability that a system performs its intended functions under expected conditions for a given period of time.
- The appropriate time scale for measuring AI reliability can be different for different structure levels or AI applications.
- Metrics are needed to characterize reliability for AI systems such as failure rate, event rate, error rate, etc.
- The measurement of the reliability of an AI algorithm is associated with the performance of the AI algorithm (e.g., classification accuracy).
- Overall, there are many metrics for AI algorithm reliability available, but in general we lack universal metrics for algorithm reliability.

Failure Modes and Affecting Factors

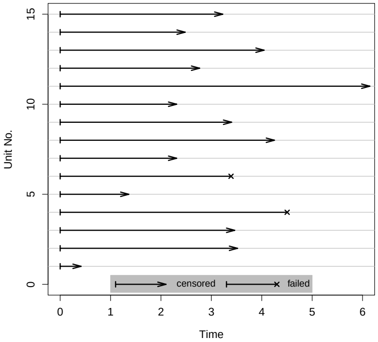
- Hardware failures, software failures, or both.
- The factors that can affect AI reliability can fall into three categories: operating environment, data, and model.



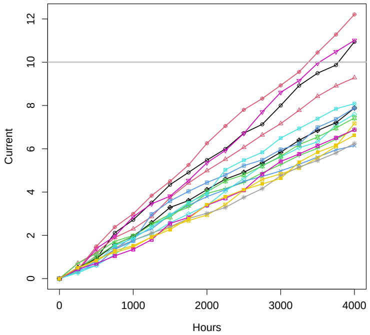
The Roles of Traditional Reliability

- Failure-time data: modeled by lifetime distribution.
- The likelihood is: $L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(t_i; \boldsymbol{\theta})^{\delta_i} \left[1 - F(t_i; \boldsymbol{\theta})\right]^{(1-\delta_i)}$.
- Degradation data: modeled by general path models $y_{ij} = \mathcal{D}(t_{ij}; \boldsymbol{\alpha}, \boldsymbol{\gamma}_i) + \epsilon_{ij}$.
- The likelihood is: $L(\boldsymbol{\theta}|\text{Data}) = \prod_{i=1}^{n} \int_{-\infty}^{\infty} \left[\prod_{j=1}^{n_i} \frac{1}{\sigma_\epsilon^2} \phi_{\text{nor}}\left(z_{ij}\right)\right] \times f_{\text{MVN}}(\boldsymbol{\gamma}_i; \boldsymbol{\Sigma}) d\boldsymbol{\gamma}_i$.
- Recurrent events data: modeled by NHPP model with intensity $\lambda(t; \boldsymbol{\theta}) = \frac{\beta}{\eta}\left(\frac{t}{\eta}\right)^{(\beta-1)}$.
- The likelihood is: $L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left[\prod_{j=1}^{n_i} \lambda(t_{ij}; \boldsymbol{\theta})\right] \exp\left[-\Lambda(\tau_i; \boldsymbol{\theta})\right]$.

Examples of Traditional Data

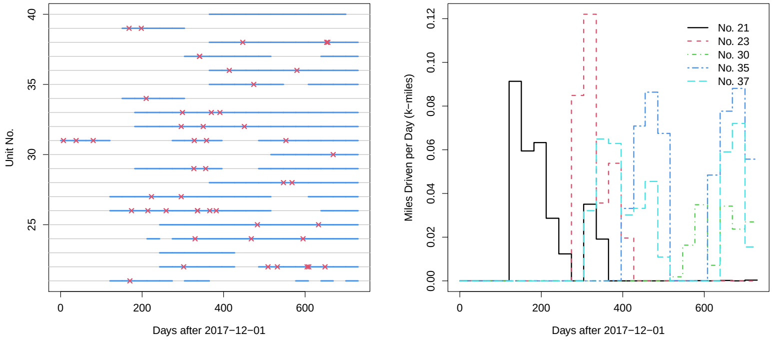


(a) GPU Failure-time Data



(b) Laser Degradation Data

Disengagement Events in Autonomous Vehicle



(a) Disengagement Events (b) Miles Driven per Day

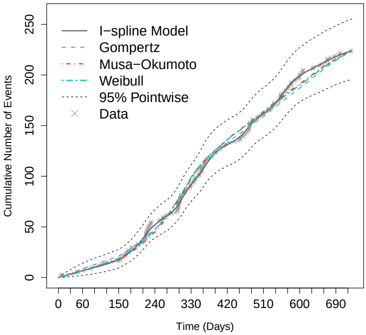
Relationship with Software Reliability

- Software reliability is an area of traditional reliability that is closely related to AI reliability.
- In modeling software reliability, usually a software reliability growth model (SRGM) based on NHPP is built.
- Traditional software reliability focuses on software bugs, but AI failures may not necessarily be caused by bugs.
- e.g., a less accurate outcome of a predictive model may lead to the crash of self-driving cars without any bugs.

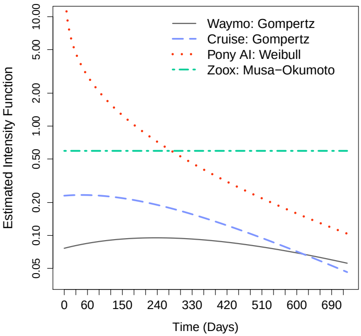
Table: List of commonly used parametric forms for SRGM.

Model	$\Lambda_0(t;\boldsymbol{\theta})$	Parameters
Musa-Okumoto	$\theta_1^{-1}\log(1+\theta_2\theta_1 t)$	$\boldsymbol{\theta}=(\theta_1,\theta_2)'$ $\theta_1>0, \theta_2>0$
Gompertz	$\theta_1\theta_3^{\theta_2^t}-\theta_1\theta_3$	$\boldsymbol{\theta}=(\theta_1,\theta_2,\theta_3)'$ $\theta_1>0, 0<\theta_2,\theta_3<1$
Weibull	$\theta_1[1-\exp(-\theta_2 t^{\theta_3})]$	$\boldsymbol{\theta}=(\theta_1,\theta_2,\theta_3)'$ $\theta_1>0, \theta_2>0, \theta_3>0$

- Min et al. (2022) analyzed disengagement event data from manufacturers Waymo, Cruise, Pony AI and Zoox for the period from December 1, 2017 to November 30, 2019.
- Spline model was used to model the BCIF: $\Lambda_0(t;\boldsymbol{\theta}) = \Lambda_0(t) = \int_0^t \lambda_0(s;\boldsymbol{\theta})ds$.



(a) Waymore

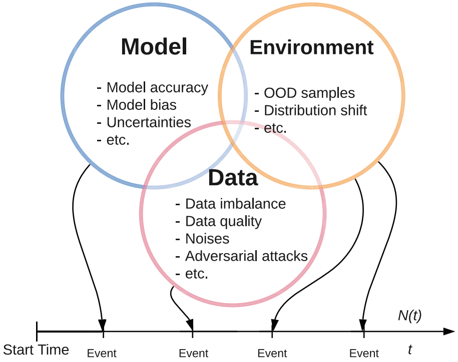


(b) Estimated Baseline Intensity

Challenges in Statistical Analysis of AI Reliability

- Need a general framework for AI reliability modeling
- Out-of-distribution detection
- The effect of data quality and algorithm
- Adversarial attacks
- Model accuracy and uncertainty quantification

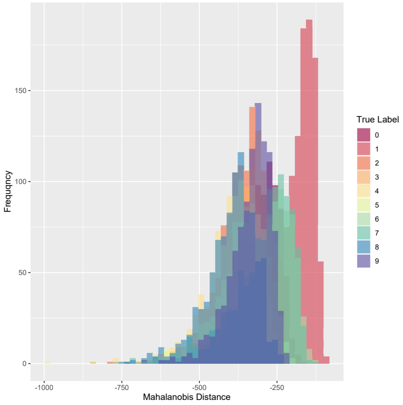
AI Reliability Modeling Framework



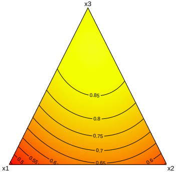
- A general intensity function for the counting process is proposed as $\lambda[t; \boldsymbol{x}(t), \boldsymbol{z}] = \sum_{j=1}^{k} \lambda_j[t; \boldsymbol{x}(t)] \cdot p_j(\boldsymbol{z}; \boldsymbol{\beta}_j)$.
- The probability is modeled as $p_j(\boldsymbol{z}; \boldsymbol{\beta}_j) = \frac{\exp(\boldsymbol{z}'\boldsymbol{\beta}_j)}{1+\exp(\boldsymbol{z}'\boldsymbol{\beta}_j)}$.

Out-of-Distribution Detection

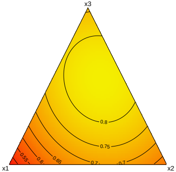
- OOD observations are data that never appear in the training set.
- In classification problems, many ML tasks assume the labels in the test set all appear in the training set.
- However, it is possible that we encounter a new class in the test dataset.
- Xu et al. (2024) developed an OOD detection method based on Mahalanobis distance: $M(\boldsymbol{x}_i) = \max_j \left\{ -[f(\boldsymbol{x}_i) - \widehat{\mu}_j]' \widehat{\boldsymbol{\Sigma}}^{-1} [f(\boldsymbol{x}_i) - \widehat{\mu}_j] \right\}$.



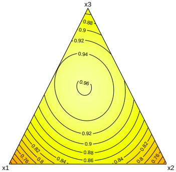
- Lian et al. (2021) used a mixture experimental design to study the effect of data quality and algorithms.
- The performance of AI algorithms is measured by the area under the receiver operating characteristic curves.



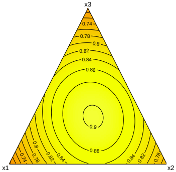
(a) CNN + Bone Marrow



(b) CNN + KEGG



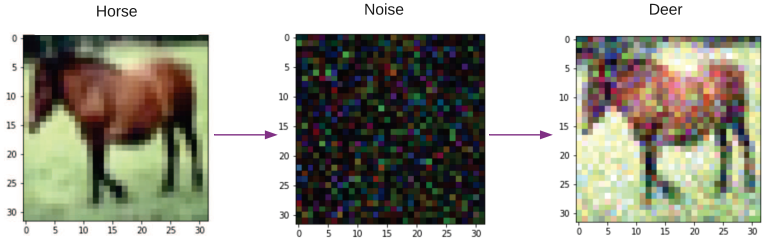
(a) XGBoost + Bone Marrow



(b) XGBoost + KEGG

Adversarial Attacks

- The research on AA focuses on finding adversarial points.
- One needs to solve the following optimization problem, $\min_{\boldsymbol{x}^*} \|\boldsymbol{x}^* - \boldsymbol{x}\| \quad \text{s.t.} \quad f(\boldsymbol{x}^*) \neq f(\boldsymbol{x})$.
- Adversarial attacks can lead to misclassification, which can further lead to reliability issues.
- To ensure the accuracy of the AI application, efforts should be made to prevent or mitigate the impacts of AA.
- It is necessary to detect AA and to study how AA affects the reliability of AI systems.



Model Accuracy and Uncertainty Quantification

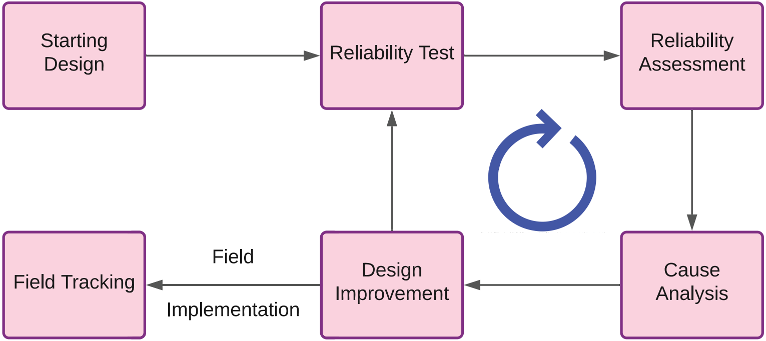
- An ML/DL model has to be accurate enough so that the model can be applied in the field.
- Thus model accuracy is a key factor to reliability.
- One question that is often asked is how much should trust on the model accuracy, which leads to the uncertainty quantification (UQ) problem.
- Quantifying the uncertainty of ML models is the key to understand the reliability of model prediction, especially for critical AI tasks.
- As an example, variational inference can be used to conduct UQ.
- The variation distribution is found through: $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \mathrm{KL}[q(\boldsymbol{\eta}; \boldsymbol{\theta}) \| p(\boldsymbol{\eta} | \mathcal{X}, \boldsymbol{y})]$.

Accelerated Tests for AI Systems

- In traditional reliability analysis, accelerated tests (AT) are widely used to obtain information in a timely manner for products that can last for years or even decades.
- The widely used methods for accelerations in the traditional reliability setting are use-rate acceleration, aging acceleration, and stress acceleration.
- The failure of software systems is usually use driven. Thus testing under high use rate can speed up the test cycles.
- To increase the stress on the AI systems, one way is to use input-data acceleration.
- In addition, testing the systems under AA can be viewed as a form of input-data acceleration.
- Operating environment acceleration, which is to test the AI systems under the OOD situation, can also be considered.

AI Reliability Improvements

- The ultimate goal of statistical reliability analysis is to improve designs for reliable AI systems.
- The flow chart below shows some steps for AI reliability improvement.



Concluding Remarks

- We provide statistical perspectives on the reliability analysis of AI systems.
- The objective is to provide general discussion coupled with concrete illustrations.
- We provide a statistical framework and failure analysis for AI reliability.
- One challenge is the limited public availability in reliability data from AI systems, which is common for all systems and products because reliability data are usually proprietary and sensitive.
- It is ideal to build data repository for AI reliability datasets.
- The paper is published by *Quality Engineering*, Volume 35, Pages 56-78, 2023.

Thank You!