# DEPENDABLE AND SECURE AI-ML

(AI60006)

Course Overview & Basic Introduction

https://portswigger.net/daily-swig/trojannet-a-simple-yet-effective-attack-on-machine-learning-models

https://portswigger.net/daily-swig/machine-learning

# Course Structure

- Introduction to dependable AI:
  - Resilience, robustness, safety and security
- Reliable Neural Networks:
  - Fault Models, Assessing Fault Tolerance, Redundancy,
  - Reliability during the Learning Phase.
- Methodology for Fault Tolerance:
  - Fault Locations, Fault Manifestations, Fault Coverage.
- Low-cost fault-mitigation techniques:
  - improving the dependability through software testing
  - Accelerators against Soft Errors and Permanent Faults.
- Measuring the Reliability of Reinforcement Learning Algorithms,
- Generative adversarial networks (GAN)

Advanced Topic (if time permits):

Game-Theoretic Methods for Robustness, Security, and Resilience

Fuzzing for vulnerability detection

Integrity checks and monitoring

Provable safety and provable Defense

Formal Scenario Based Testing of Autonomous Vehicles:From Simulation to the Real World

**Study Material:**

**https://drive.google.com/drive/folders/1at__tMA7NzkDNo3xCnCZHu-L4OQI8db2?usp=sharing**

# Course Structure

- Secure AI: Privacy concerns in ML and DL,
- Adversarial models:
  - Honest-but-curious adversary model, semi-honest entity, active adversary model.
- Attacks against ML/DL:
  - Evasion/Adversarial attack, Poisoning, Inference, Trojans, Backdoor attacks with Case Study
- Differential Privacy basics:
  - Properties of Differential Privacy: privacy preservation,
  - Sensitivity, randomization, composition, and stability;
- Differential Privacy in Supervised Learning, Differential Privacy in Unsupervised Learning

- Federated machine learning:
  - Model Training in Federated Learning and optimisation,
- Privacy-Preservation in centralised FL framework:
  - Attack Models on FL, Privacy-preservation solutions

- Homomorphic encryption and machine learning :
  - Basics of homomorphic encryption, Secure hyperplane decision, Naïve Bayes, and decision trees-polynomial approximations ,
  - Division-Free Integer Algorithms for Classification,
  - Homomorphic evaluation of deep neural networks,
  - Case study on medical data

- **Assignment 1 :** Adversarial Robustness Toolbox (ART) for TrustedAI (IBM) (Python)
- **Assignment 2 :** Pydp (Python), IBM differential privacy tool (Python), Google DP (Java/C)
- **Assignment 3 :** Encrypted ML with homomorphic library

**Grading plan (Tentative):**

- Class Test 1, 2: 20/100
- Assignments : 20/100
- MidSem & EndSem : 60/100

# Dependable NN

**Accuracy :** The prediction accuracy is the basic ability of a trustworthy model. Trustworthy NNs are expected to generate accurate output, consistent with the ground truth, as much as possible;

**Reliability:** Trustworthy NNs should be resilient and secure. In

other words, they must be robust against different potential threats, such as inherent noise, distribution shift, and adversarial attacks;

**Explainability:** The model itself must allow explainable for the prediction, which can

help humans to enhance understanding, make decisions and take further actions;

**Privacy protection:** Trustworthy NNs are required to ensure full privacy of

the models as well as data privacy.

# Pillars of Security

- **Confidentiality** is satisfied if data or objects are not read by an unauthorized party.
- **Integrity** is satisfied if data or objects are not changed (written) or generated by an unauthorized party.
- **Authenticity** is satisfied if an author of data or an object is who it claims to be.
- **Availability** is satisfied if data, objects, or services are available.

# Can AI-based Components be Part of Dependable Systems?

# Dependable Systems


Dependable Systems can be found in many forms and application domains, but especially in transportation systems, medical systems and recently in the domain of IoT and Industry 4.0.*

*Industry 4.0 has been defined as "a **name for the current trend of automation and data exchange in manufacturing technologies**, including cyber-physical systems, the Internet of things, cloud computing and cognitive computing and creating the smart factory"

# “Dependability”

- **Reliability** how often is the system allowed to fail
- **Availability** to which extend is the system usable, when it is needed
- **Maintainability** how intense is the maintenance of the system
- **Safety** how much the environment be secured against the system
- **Security** how much the system be protected against the environment

# **Safety Integrity Level** : Standard

- Safety Integrity Levels (SIL) define the criticality of the component,
- Each SIL requires different development techniques as well as testing or verification methods and techniques.

- The SILs are defined by the probability of failure, a risk reduction factor (can the risk of failure be reduced by a certain amount, using multiple instances, redundancy, etc), probability of failure per hour and the meantime between failure.

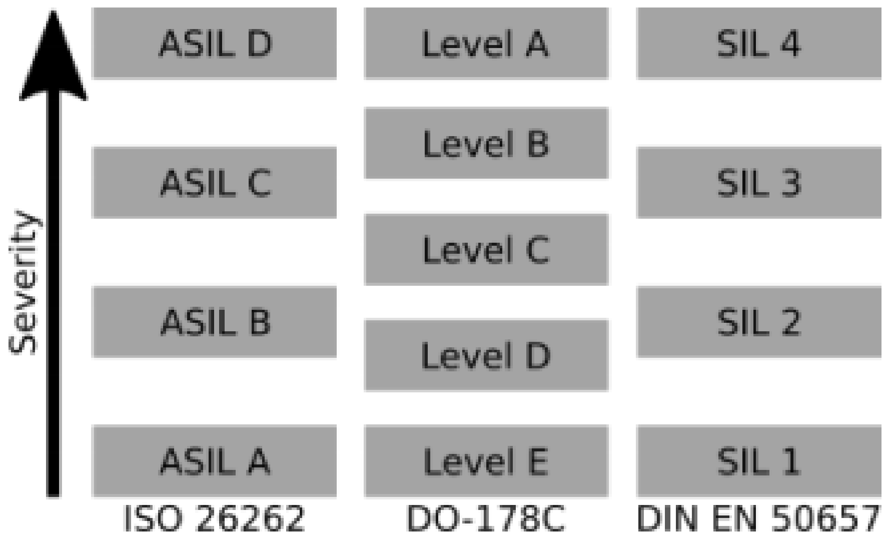
| **SIL** Safety Integrity Level | **PFDavg** Average probability of failure on demand per year (low demand mode) | **RRF** Risk Reduction Factor | **PFDavg** Average probability of failure on demand per hour (high demand or continuous mode) |
|---|---|---|---|
| **SIL 4** | $\geq 10^{-5}$ and $< 10^{-4}$ | 100000 to 10000 | $\geq 10^{-9}$ and $< 10^{-8}$ |
| **SIL 3** | $\geq 10^{-4}$ and $< 10^{-3}$ | 10000 to 1000 | $\geq 10^{-8}$ and $< 10^{-7}$ |
| **SIL 2** | $\geq 10^{-3}$ and $< 10^{-2}$ | 1000 to 100 | $\geq 10^{-7}$ and $< 10^{-6}$ |
| **SIL 1** | $\geq 10^{-2}$ and $< 10^{-1}$ | 100 to 10 | $\geq 10^{-6}$ and $< 10^{-5}$ |

**IEC 61508**
**Automatic protection systems called safety-related system**

# Standards: **Safety integrity level (SIL)**

SIL1 being the lowest and SIL4 the highest severity level:

- For example in SIL4 :
  - high coverage of branches in the source code of a component
  - ensures adequate testing of the most critical components in the system.
  - standard procedure in avionic or automotive applications.




- Civil avionic systems : regulated by DO178c
- train applications DIN EN 50657
- Medical devices are certified under IEC 82304, 2018 [4]
- Automotive under ISO 26262 [5].
- Each of these standards defines strict requirements with the goal to ensure the the functional safety of each component

# Ensuring dependability in critical systems

- **Analytical Approaches:** strict and rigorous review of specification, design and implementation.
- **Constructive Approaches:** These techniques and patterns can be used as a guideline to ensure safety during the design and implementation phase of a project, for example safety cases. These scenarios can be used to directly derive the design or even parts of the implementation of the system.
- Fault tolerant system with **redundancy** concepts can be implemented to increase the reliability and availability of the system.
- In addition a **fault containment** strategy can be developed. If a fault occurs the consequences spread only to specific predefined boundaries, as a result, the system can stay intact.

## Risk mitigation mechanisms

From a more technical point of view dependability properties of a system can be improved by adding risk mitigation mechanisms:

- watchdog or
- brownout detection.

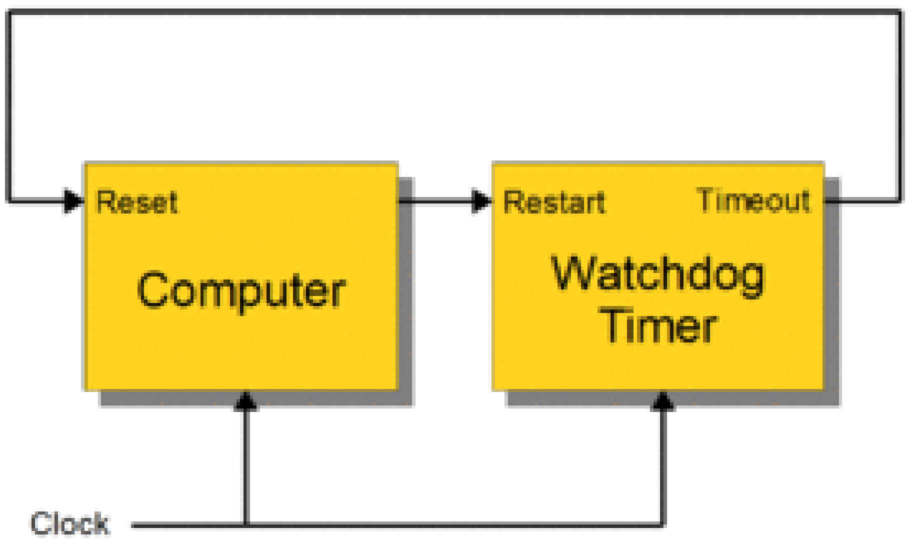
# Watchdog

- The hardware component of a watchdog is a counter:
  - set to a certain value

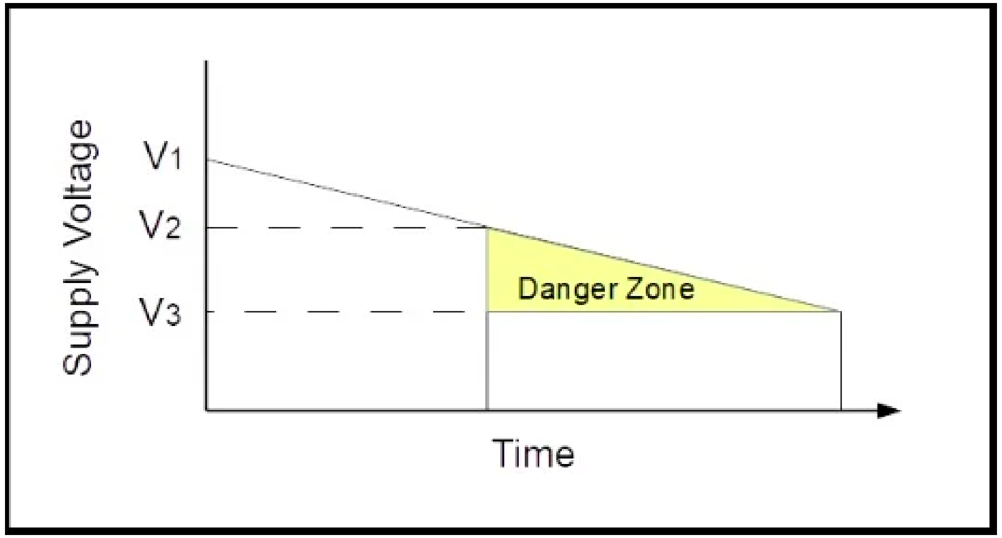
    then counts down towards zero.

- It is the responsibility of the software:
  to set the count to its original value so that it never reaches zero.

- If timer reaches zero:
  - it is assumed that the software has failed in some manner
  - CPU is reset.

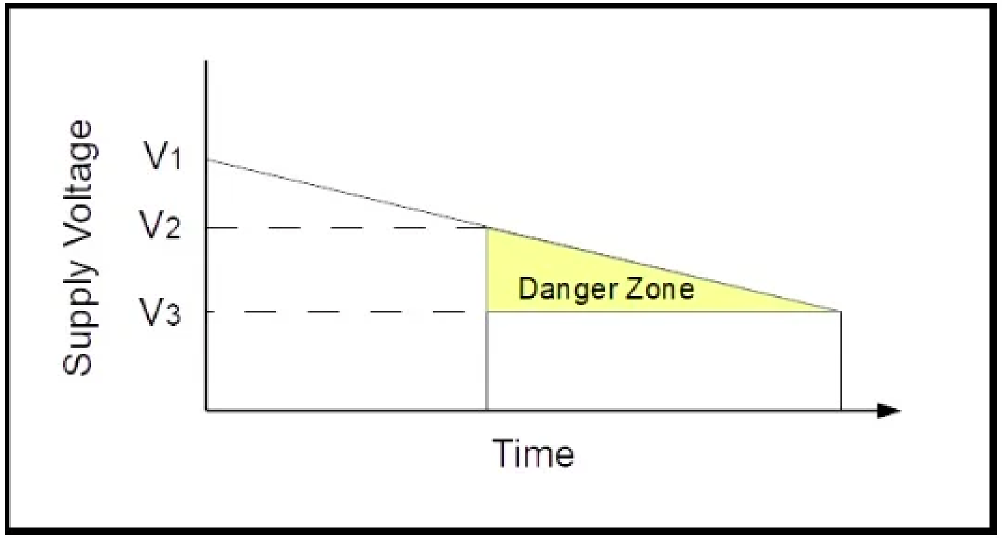


# Brownout Detection



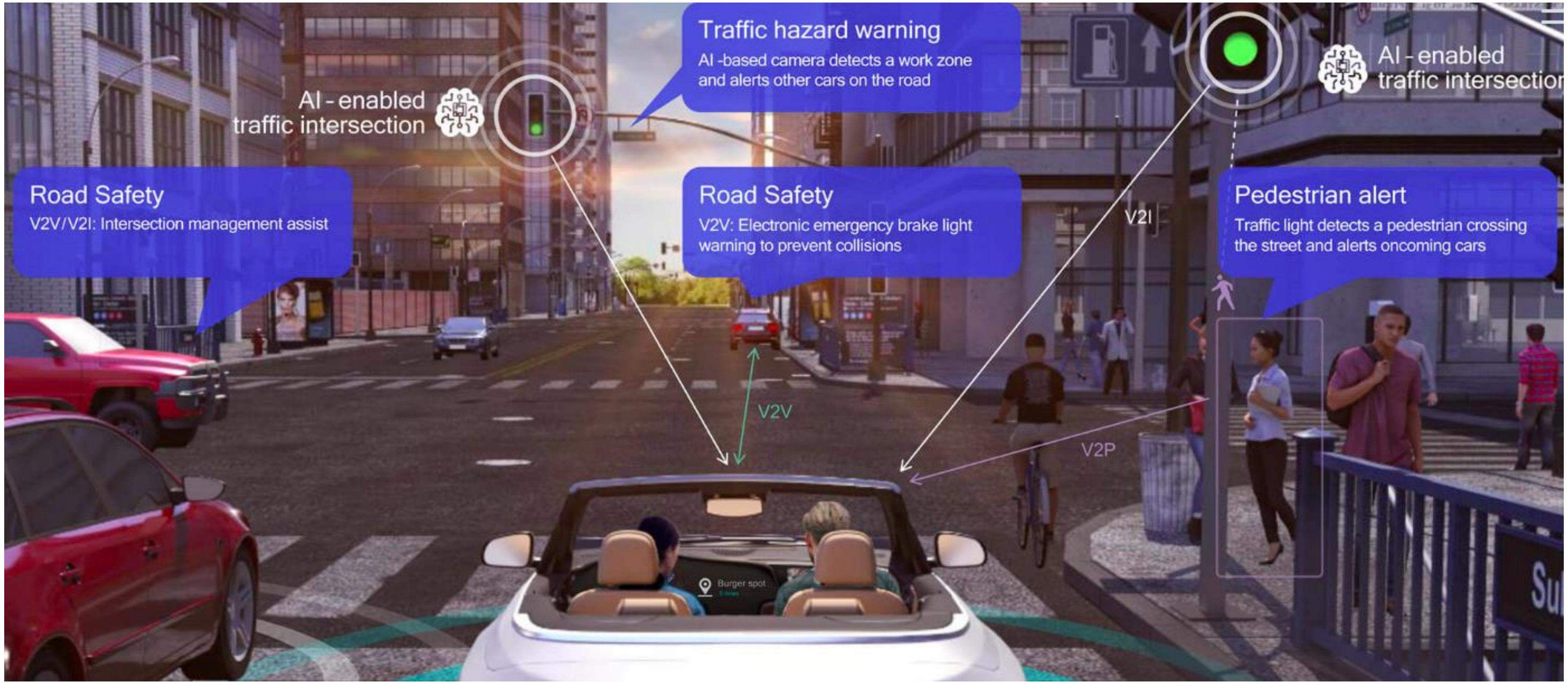

- A “brown out” of a microcontroller is a partial and temporary reduction in the power supply voltage below the level required for reliable operation.
- Many microcontrollers have a protection circuit which detects when the supply voltage goes below this level
- puts the device into a reset state to ensure proper startup when power returns. This action is called a “Brown Out Reset” or BOR.

## Brownout Detection




- V1: is the normal power supply voltage.
- V2: is the point where the microcontroller may not operate reliably.
- V3: a point where operation stops entirely.
- Between V2 and V3 is a “danger zone” where things can go wrong and operation is unreliable.
- The device could work correctly for years while the power supply goes in and out of the danger zone and then there is a failure.
- The BOR level is set above V2 and replaces the danger zone with a reset of the device.
- Reset is not good but (usually) better than uncertain.

# Artificial Intelligence in Critical Systems




- Autonomous driving : prominent example for systems incorporating critical components derived using ML.
- The capability of an **AI system to react in complex scenarios in a short amount of time** is unique
- Enables the system to identify pedestrians or traffic signs in a fraction of classic image recognition methods used before.

# Problems with AI-ML

- ML methods are b**ased on probabilities**.
- They are **stochastic principles**, which can only estimate the correct answer with a specific certainty.
- Even though the ML algorithms might be 100% sure that the outcome is correct, the answer can still be wrong. This can e.g. happen if the **quality of the training** data is too low, or the data does not even contain all possible scenarios