
# A Systems Engineering Perspective on Safety of AI-based Systems

## Abstract Presentation for AI for Systems Engineering and Systems Engineering for AI Conference

***Reginald Holmes, MEng, ASEP, PMP***

***Ph.D. Student - Systems Engineering***

***The University of Alabama in Huntsville***

***Dr. Hanumanthrao "Rao" Kannan***

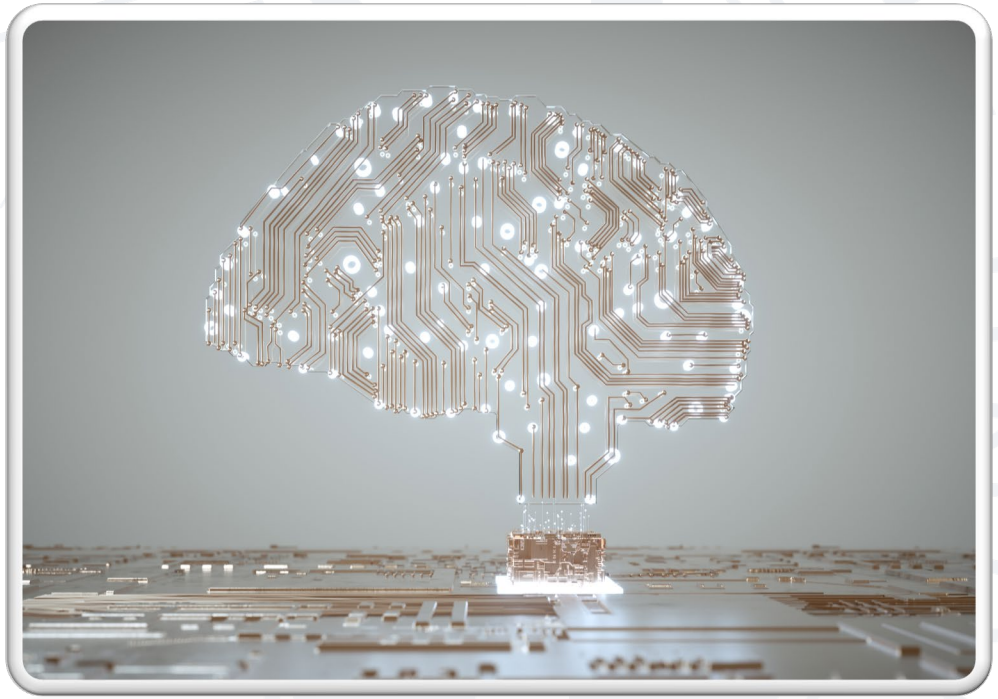
***Assistant Professor***

# “A Tragic Loss”

“For Joshua Brown and the countless others whose stories remind us why we strive for safety..”

# In This Presentation

- Introduction
- Definition of AI-Based Systems
- AI Expansion and Safety Challenge
- Systems Engineering Approach to AI Safety
- Emergent Behavior and Safety
- Overview of Safety in Different Domains
- Desired and Undesired Emergent Behavior
- Robustness, Resilience, Reliability, and Safety
- What's Next?



# Introduction

- AI technologies are rapidly expanding into diverse areas, from space missions and power plants to healthcare.
- This growth introduces unique safety challenges; think space debris in space missions or ensuring reactor safety in power plants.
- Furthermore, AI-based systems complicate these challenges with unpredictable behaviors and "black box" decision-making
- Within the Systems Engineering framework, this research examines the relationship between such emergent behaviors and safety, emphasizing AI's positive and negative implications for safety across industries.

# Definition of AI-Based System

AI-based system (also called AI-enabled system) refers to a software-based system that comprises AI components besides traditional software components (Felderer and Ramler Feb 2021)

Artificial Intelligence is the simulation of human intelligence processes by machines, especially computer systems (Rouse M. 15 Aug 2021)

Artificial intel
computers to
the ability to
language, an
(Rouse M.,15

elligences
ed to emulate (or
ans, such as
ng tasks (Huang

AI as machin
reasoning (S

The field
not just u
entities –
effectively and safely in a wide variety of novel situations (Russell and Norvig 2020)

data

environment and takes action to achieve its goals (Russell and Norvig 2016)

Plenty of literature describes the characteristics of AI and AI-based systems however, there is no generally accepted definition of artificial intelligence (AI)

For this paper, the term AI-based system will refer to software or hardware systems that mimic human-like intelligence through rule-based algorithms, heuristics, or other techniques. They possess properties like intelligent behavior, adaptability, autonomy, learning, decision-making, perception, and communication to achieve predefined goals while meeting ethical requirements.

These systems can analyze large volumes of data, identify patterns, make decisions, and adapt to changes.

# AI Expansion and Safety Challenges

# Understanding AI’s evolving impact on safety

- **Technological Advancements:** AI technologies are rapidly evolving, introducing new capabilities and complexities. Systems Engineers must stay informed about these advancements to assess their potential impact on safety and update safety measures accordingly.

- **Unpredictable Behavior:** As mentioned earlier, AI systems, especially those based on machine learning, can exhibit unpredictable behavior. Understanding how AI's evolving capabilities affect its behavior and potential safety risks is essential for risk assessment and management.

- **Ethical Considerations:** The ethical landscape surrounding AI is evolving. New ethical considerations, such as data privacy, fairness, and accountability, continue to emerge. Engineers must be aware of these evolving ethical standards and integrate them into AI safety practices.

- **Regulatory Changes:** AI safety regulations are also evolving in response to technological developments and societal concerns. Staying informed about changes in regulations and ensuring compliance is essential for AI safety.

- **Continuous Monitoring:** AI systems require continuous monitoring and adaptation. Understanding AI's evolving impact on safety involves keeping AI systems up-to-date, addressing new risks, and ensuring they remain safe in changing environments.

# Can SE principles address the issues with AI safety?

| AI Challenges | Characteristics |
|---|---|
| **Complexity of AI Systems** | AI-based systems are characterized by their **complexity**, often involving intricate algorithms and neural networks. These systems exhibit behaviors that can be **challenging to predict and control** using traditional systems engineering principles designed for more deterministic systems. |
| **Lack of Transparency** | Many AI algorithms, especially deep learning models, operate as "**black boxes**," making their decision-making processes **opaque**. Current systems engineering principles emphasize transparency and comprehensibility, which may not directly apply to AI systems. |
| **Adversarial Attacks** | AI systems are vulnerable to harmful attacks that manipulate input **data**. Specialized methods must be developed to detect and mitigate these attacks. Integrating these techniques into safety protocols is crucial for maintaining the reliability and security of AI-powered solutions. |
| **Continuous Learning and Adaptation** | The **dynamic nature** of AI systems can lead to new safety challenges that traditional systems engineering principles may not adequately address. To overcome these challenges, it is necessary to incorporate ongoing monitoring, adaptation strategies, and reinforcement learning through augmentation. |
| **Regulatory Compliance** | AI safety regulations are rapidly evolving, and systems engineering practices must ensure compliance with emerging standards and regulations. Augmentation should include mechanisms for aligning with these developing regulations. |
| **Data Quality and Security** | The security and quality of data used by AI systems are crucial for safety. It is important to prioritize robust **data** governance, security practices, and measures to address data quality issues in augmentation. |
| **Cascading Failures** | AI systems are interconnected across various domains, posing the risk of cascading failures. A failure in one AI system could trigger a chain reaction, amplifying the magnitude of safety breaches. Augmentation should include strategies to prevent and manage cascading failures. |

# Systems Engineering Gaps?

Although the foundations and formalized standards in Systems Engineering offer a valuable foundation, there are still gaps and areas where increased rigor could improve SE practices. In the context of our research, we will focus on highlighting some of these common gaps:

- INCOSE and ISO-15288 process areas capture best practices to address complicated systems, but there are additional methods required to address the **emergent behavior** found in **Complex Systems** (Castellani, 2014)

- Traditional systems engineering, which relies on prediction and control, faces challenges due to the increasing **complexity of systems** and the need to interconnect existing systems. These challenges arise from one of two key issues: (Pennock M.J., Wade J.P, 2015)
  - **Complexity Challenge**: The growing complexity of systems makes it difficult for system engineers to predict the outcomes of their design decisions accurately. As systems become more intricate, traditional prediction methods become less effective.

- The importance of **systems of systems** in today's global endeavors requires that systems engineers develop methods for analyzing **emergent behavior**, in order to predict favorable and unfavorable consequences and in order to **architect Systems of Systems** to better assure desire results (Osmundson J.S., Huynh T.V., Langford G.O.,2008)

**To address these gaps, research is crucial to establish new foundations for systems engineering.**

# Understanding Safety and Emergent Behavior

- Safety and emergent behavior are interconnected concepts, particularly in the context of complex systems like AI-based technologies, where unexpected interactions can lead to potentially harmful outcomes
- Safety entails the systematic identification, assessment, mitigation, and management of risks, with the primary objective of preventing harm to individuals, the environment, and assets, while also ensuring the reliable and effective operation of complex systems.
- Emergent behavior refers to the phenomenon where a complex system, composed of multiple interconnected components or subsystems, exhibits behaviors that are not explicitly designed or anticipated in the individual components or subsystems. (Osmundson J.S., Huynh T.V., Langford G.O.,2008)
- Emergent behavior can be beneficial or hazardous when unexpected patterns arise from system components' interactions

# Overview of Safety Practices in Different Domains

Safety in engineering spans domains like Systems Engineering (SE), Space Exploration, Nuclear Industry, Industrial Automation, and Healthcare, each posing distinct challenges. These challenges encompass system interconnectedness, space risks, nuclear safety, industrial cybersecurity, and healthcare privacy.

Universal Systematic approach to Safety across these domains:

- **Risk assessment and management**
- **Compliance with regulations and standards**
- **Designing for safety from the start**
- **Ongoing training and education**
- **Regular maintenance and inspections**
- **Emergency response planning**
- **Continuous improvement based on lessons learned**
- **Consideration of human factors in design and operation**

# Desired vs. Undesired Emergent Behavior

- Just as emergent behavior in complex systems can have both desired and undesired consequences, Two-Face's character reflects this duality

- Two-Face, also known as Harvey Dent, was a district attorney in Gotham City who, after a tragic accident, developed a split personality. One side of his personality is rational and law-abiding, while the other side is chaotic and criminal. This duality represents the unexpected emergence of contrasting behaviors within a single individual.

- In the context of emergent behavior and safety, Two-Face serves as a metaphor for the potential consequences of unexpected patterns or behaviors that can arise in complex systems.

- Just as Two-Face's dual nature can lead to both positive and negative outcomes, emergent behavior in technology and engineering can result in desired improvements in performance and safety, as well as unforeseen risks and hazards.

# Emergent Behavior

**Desired**

The desired behavior of a Tesla Model S in "Autopilot" mode is to provide semi-autonomous driving capabilities while requiring active supervision and intervention from the human driver

- The key desired behaviors include:
  - **Lane keeping**
  - **Adaptive cruise control**
  - **Traffic-Aware Cruise Control**
  - **Emergency Collison Avoidance**
  - **Autosteer on Limited-Access Highways**
  - **Navigate on Auto-pilot**

The desired behavior of Autopilot is to assist the driver, enhance safety, and make highway driving more convenient but not to replace the driver's active role.

**Undesired**

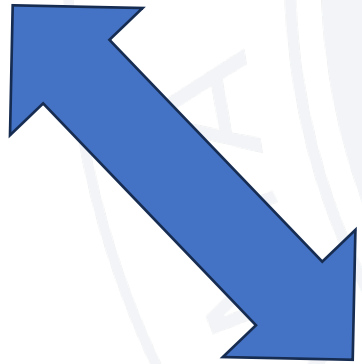
The fatal crash involving a Tesla Model S in 2016, as previously mentioned, was a tragic incident, and it raised questions about the performance and limitations of Tesla's Autopilot system.

- The key undesired behaviors include:
  - **Sensor Limitations**
  - **Failure to brake**
  - **Driver complacency**
  - **Lack of clear system limitations**
  - **Misinterpretation of road scenarios**
  - **Inadequate Response to Human Interventions**

Undesired behaviors can jeopardize AI reliability, robustness, and overall safety. A systems engineering approach is crucial to mitigate these risks.

# The relationship between Robustness, Reliability, Resilience, and Safety

Ensuring safety is the top priority, which includes robustness, reliability, and resilience. Achieving high performance in all these aspects is crucial for complex systems to succeed and last in various fields and industries. Engineers, designers, and practitioners must thoroughly consider these factors when developing and maintaining systems to ensure optimal performance and safety, even amid uncertainties and challenges.

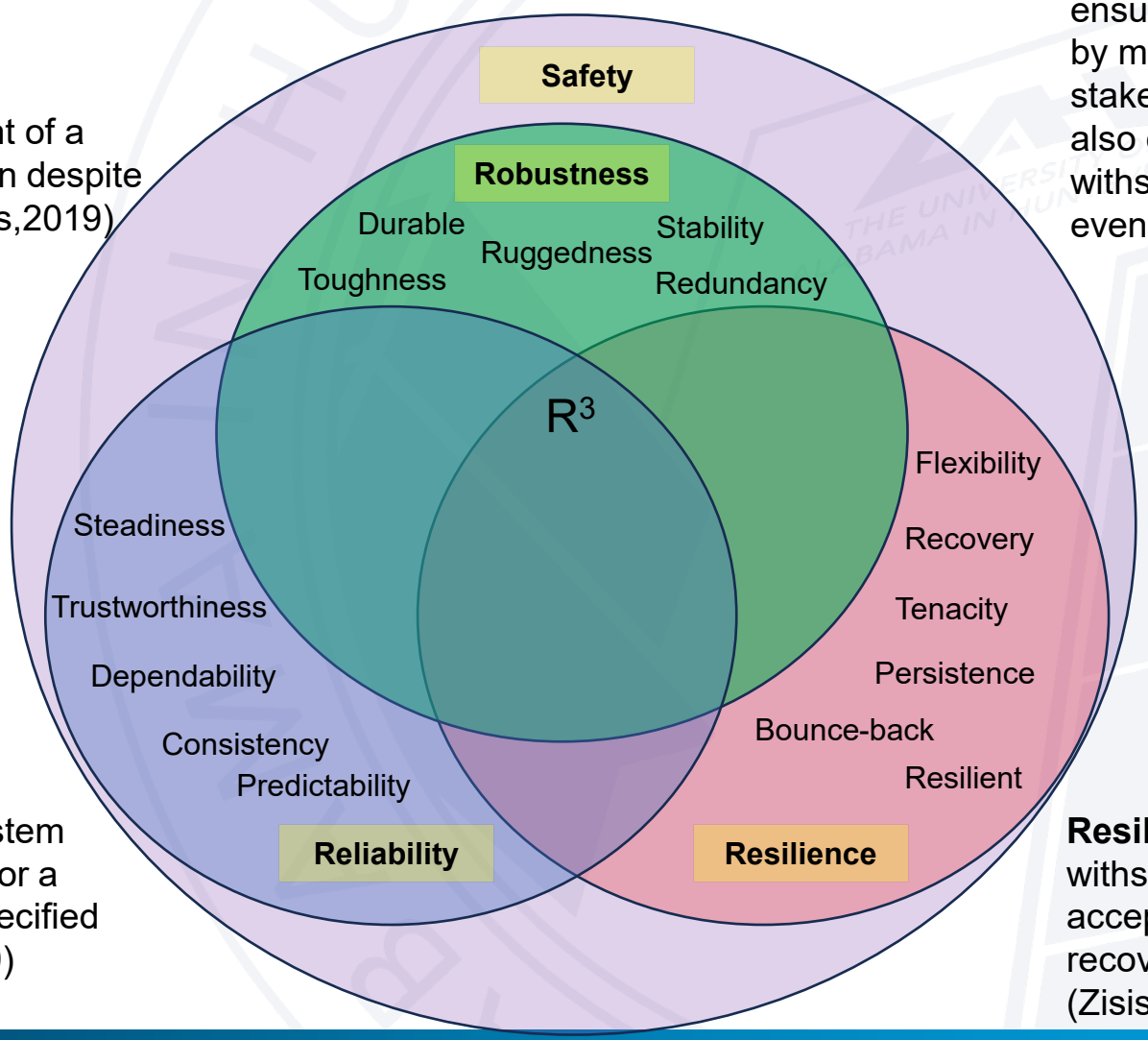


The **"$R^3$" Concept** that deals with Robustness, Reliability, and Resilience is a framework for enhancing the performance and dependability of complex systems, including AI systems, critical infrastructure, and other technologies

# The $R^3$ Concept in relation top Safety

**Safety (Top-level Objective)** – The **$R^3$ concept** supports the overall objective of ensuring the safe operation of the system by minimizing the risk of harm to users, stakeholders, and the environment while also enhancing the system's ability to withstand and recover from unexpected events and disruptions.

**Robustness** is the measure or extent of a systems' ability to continue to function despite faults in its subsystems or parts (Zisis,2019)




**Reliability** is the probability that a system will perform in a satisfactory manner for a given period when it is used under specified operating conditions (IEEE,Zisis,2019)

**Resilience** is a system's ability to withstand a major disruption within acceptable degradation parameters and recover with a satisfactory timeframe (Zisis,2019)

# Research goal for AI-Based Systems

Our goal is to establish a solid foundation for SE that enables the implementation of methodologies to ensure the safety of AI-based systems. We aim to enhance the safety of these systems by introducing an augmented Systems Engineering approach that prioritizes robustness, reliability, and the ability to recover from unexpected events. Our objective is to minimize potential harm to users, stakeholders, and the environment by effectively mitigating any undesired emergent behaviors in AI-based systems.

**Robustness:** The AI system should be robust, meaning it can perform well even when faced with unexpected or challenging conditions. This includes handling situations it was not explicitly trained for and avoiding erratic or unsafe behavior

Self-driving cars need to be **robust** to handle unexpected situations, such as heavy rainfall or sudden road closures. They must also quickly adapt to changes such as pedestrians appearing or temporary road signs.

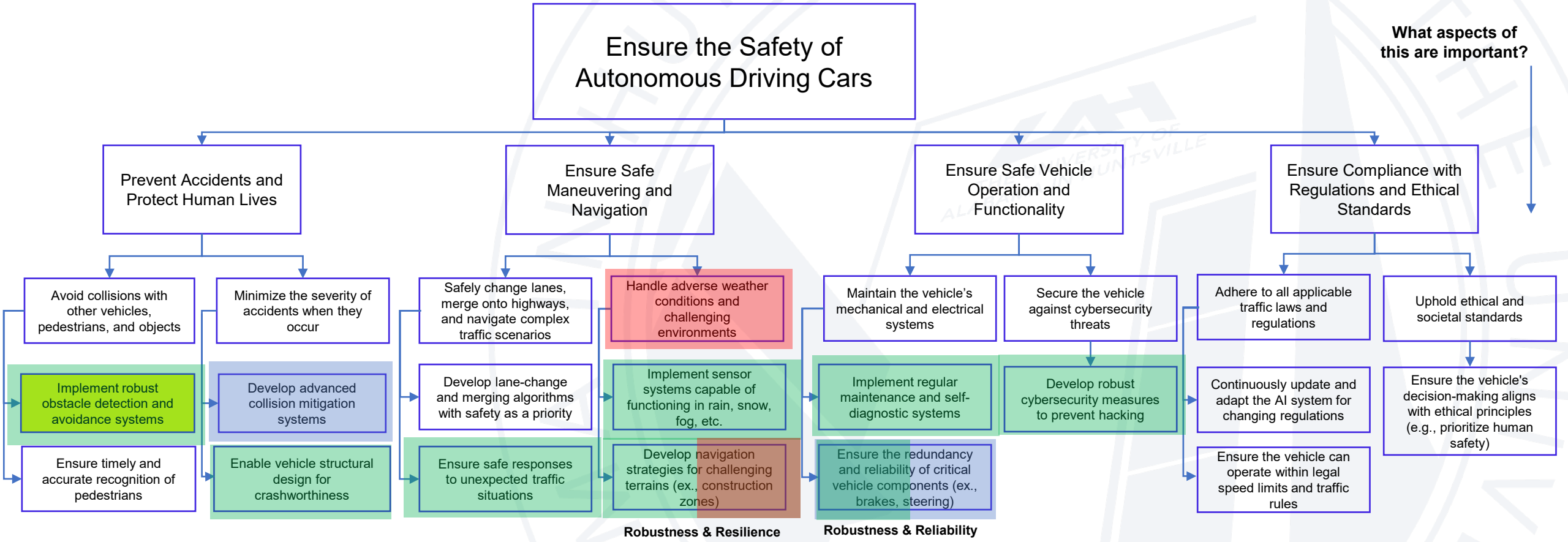
**Reliability:** The AI system should be reliable, consistently providing accurate and dependable results. It should avoid failures, errors, or unintended consequences, especially in critical or high-stakes applications where reliability is paramount

A **reliable** self-driving car drives securely, follows traffic rules, and makes good decisions during trips. This fosters trust in the vehicle.

**Resilience:** The AI system should be resilient, meaning it can adapt to changing circumstances and recover gracefully from disruptions or failures. Resilience ensures that even if a failure occurs, the system can continue to operate safely or enter a safe mode without posing undue risks

Imagine a self-driving car facing a problem with its sensors while crossing a busy intersection. A **resilient** system would act promptly and correctly using another sensor or safely stopping the vehicle.

# Ensuring Safe Autonomous Driving

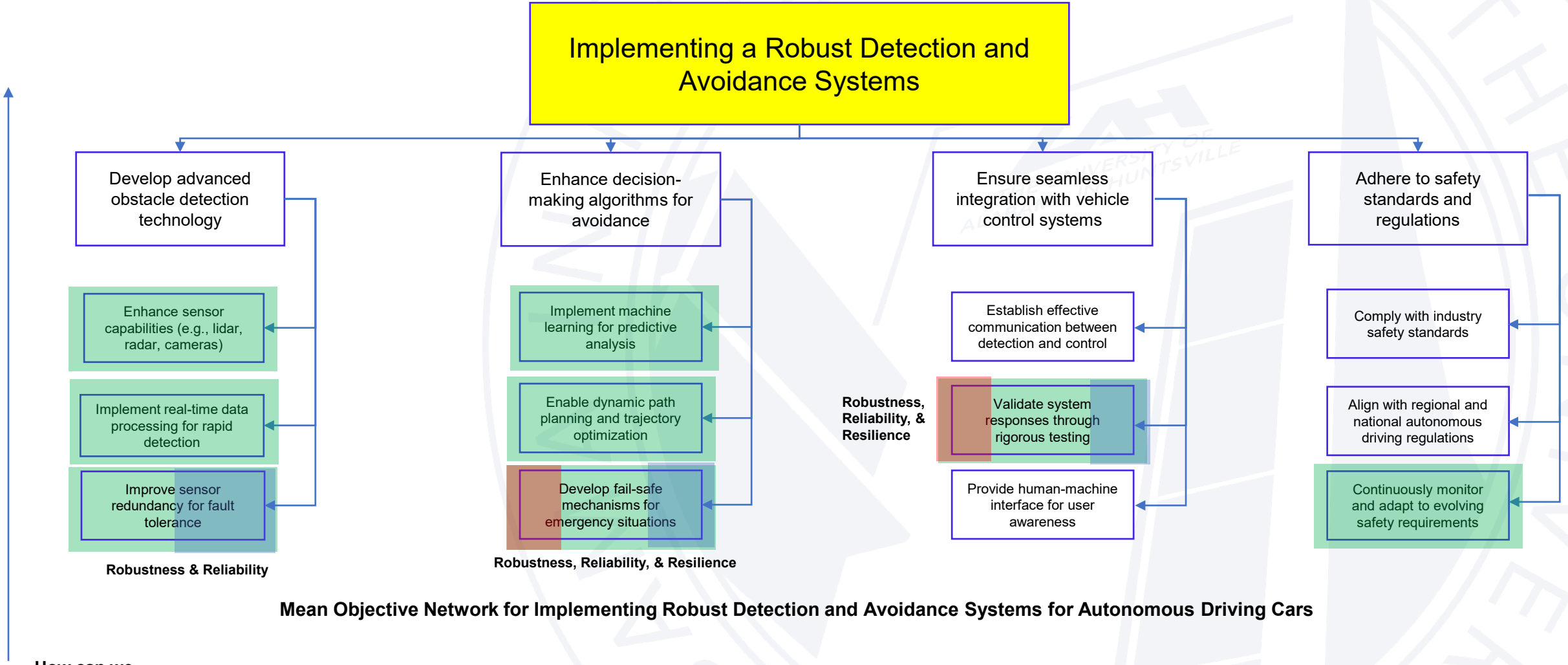



**Fundamental objectives hierarchy for ensuring the safety of autonomous driving**

**Why is this important?**

**Improved by objective:** **Robustness** **Reliability** **Resilience**

# Ensuring Safe Autonomous Driving




**Mean Objective Network for Implementing Robust Detection and Avoidance Systems for Autonomous Driving Cars**

**How can we achieve this?**

**Improved by objective:** **Robustness** **Reliability** **Resilience**

# What's Next?

| Develop rigorous Systems Engineering foundations to effectively address the safety challenges of AI-based systems | Potential Foundations/Methodologies/Tools to address this: |
|---|---|
| **What are the benefits and risks associated with emergent behavior in AI systems, and how can we explore them in-depth?**<br>➢ Dive deeper into the concept of emergent behavior in AI systems. Explore how it can be both beneficial and risky.<br>➢ Develop a framework for identifying, managing, and mitigating emergent behavior for safety. | **Foundation:** Systems theory and complexity theory.<br>**Methodologies**: Agent-based modeling, Decision theory, simulation, and system analysis.<br>**Tools:** Simulation software (ex., NetLogo, AnyLogic), machine learning for anomaly detection. |
| **How can we expand upon the relationship between robustness, reliability, resilience, and safety in AI systems to create an integrated assessment framework?**<br>➢ Expand the relationship between these key concepts and create an integrated framework for assessing AI system safety.<br>➢ Explore practical approaches and methodologies to enhance each aspect of AI systems. | **Foundation:** Ontologies, Systems engineering principles.<br>**Methodologies:** Failure Mode and Effects Analysis (FMEA), Fault Tree Analysis (FTA), System Safety Assessment (SSA).<br>**Tools**: System modeling and analysis tools (ex., SysML, AADL), reliability analysis software (ex., ReliaSoft). |
| **What are the ethical considerations surrounding AI safety, including fairness, privacy, accountability, and bias, and how do these ethical dimensions intersect with safety in AI-based systems?**<br>➢ Delve into the ethical dimensions of AI safety, covering fairness, privacy, accountability, and bias.<br>➢ Examine how ethical considerations intersect with safety in AI-based systems | **Foundation:** Game theory, Formal Logic, Ontologies, Ethical frameworks (ex., utilitarianism, deontology, virtue ethics).<br>**Methodologies:** Ethical impact assessments, algorithmic fairness audits.<br>**Tools:** Ethical AI toolkits (ex., IBM Fairness 360, AI Fairness 360). |
| **What are the existing and emerging regulatory frameworks related to AI safety across various domains, and what recommendations can be proposed to improve and harmonize these regulations?**<br>➢ Investigate existing and emerging regulatory frameworks related to AI safety across various domains.<br>➢ Propose recommendations for improving and harmonizing these regulations. | **Foundation:**, Legal and policy studies, international standards (ex., ISO 13482 for robots, ISO/IEC 27001 for cybersecurity).<br>**Methodologies:** Comparative legal analysis, regulatory impact assessments.<br>**Tools:** Compliance management software, legal research databases. |

# Conclusion

Thank you for listening. Let's work together to keep people safe by using technology wisely and with care.

# Back-up Slides

Citation and Image References

# References (information)


- International Council on Systems Engineering. (2020). Systems Engineering Handbook: A Guide for System Life Cycle Processes and Activities (4th ed.). Wiley.
- Gabriela, S., Bentler, D., Grote, E.-M., Junker, C., Meyer zu Wendischhoff, D., Bansmann, M., Latos, B., Hobscheidt, D., Kühn, A., & Dumitrescu, R. (2022). Requirements analysis for an intelligent workforce planning system: a socio-technical approach to design AI-based systems. 1-6
- Hyvärinen, J., Vihavainen, J., Ylönen, M., & Valkonen, J. (2022). An overall safety concept for nuclear power plants. Annals of Nuclear Energy, Volume 17, 1-18.
- Shivers, C. H. NASA Space Safety Standards and Procedures for Human Rating Requirements. Publisher: Marshall Space Flight Center, NASA.
- Schenkel, S. (2000). Promoting Patient Safety and Preventing Medical Error in Emergency Departments. From the Department of Emergency Medicine, University of Michigan, Ann Arbor, MI.
- Russell, S. J., Norvig, P., & Davis, E. (2020). Artificial Intelligence: A Modern Approach (4th ed.). Pearson
- Mukhamediev, R. I., Popova, Y., Kuchin, Y., Zaitseva, E., Kalimoldayev, A., Symagulov, A., Levashenko, V., Abdoldina, F., Gopejenko, V., Yakunin, K., Muhamedijeva, E., & Yelis, M. (2022). Review of Artificial Intelligence and Machine Learning Technologies: Classification, Restrictions, Opportunities and Challenges. Academic Editors: Anatoliy Swishchuk and Jakub Nalepa. Published: July 22, 2022. MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.
- Stephen, C. (2020). Impediments to Effective Safety Risk Assessment of Safety Critical Systems: An Insight into SRM Processes and Expert Aggregation. Master's thesis, Virginia Polytechnic Institute and State University, Blacksburg, Virginia.
- Blais, M.-A., & Akhloufi, M. A. (2023). Reinforcement learning for swarm robotics: An overview of applications, algorithms, and simulators. Cognitvie Robotics, volume 3, 226-256
- Muecklich, N., Sikora, I., Paraskevas, A., & Padhra, A. (2023). Safety and reliability in aviation – A systematic scoping review of normal accident theory, high-reliability theory, and resilience engineering in aviation. Safety Science, 162 (106097), 1-16.

# References (information)


- Hollnagel, E. 2016 Resilience Engineering. Available at: https://erikhollnagel.com/ideas/resilience-engineering.html
- Vamplew, P., Foale, C., Dazeley, R., & Bignold, A. (2021). Potential-based multiobjective reinforcement learning approaches to low impact agents for AI safety. Engineering Applications of Artificial Intelligence, 100 (104186), 1-16.
- Ihsan, UJ., Seonggoo, J., Changju, K. (2023). What (de) motivates customers to use AI-powered conversational agents for shopping? Journal of Retailing and Consumer Services, 75 (103440), 1-16.
- Huang, M. H., & Rust, R. T. (2021). A framework for collaborative artificial intelligence in marketing. Journal of Retailing, 4.
- Beerman, J., Beaumont, G., Giabbanelli, P. (2023). A framework for the comparison of errors in agent-based models using machine learning, Journal of Computational Science, 72 (102119), 1-14.
- Dey, S., Lee, S-W. (2021). Multilayered review of safety approaches for machine learning-based systems in the days of AI, The Journal of Systems & Software, 176 (110941), 1-24.
- Diaz-Rodrigueaz, N., Del Ser, J., Coeckelbergh, M., Lopez de Prado, M., Herrera-Viedma, E., Herrera, F., (2023). Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulations, Information Fusion, 99 (101896), 1-24.
- Bachute, M., Subhedar, J., (2021). Autonomous Driving Architectures: Insights of Machine Learning and Deep Learning Algorithms, Machine Learning with Applications, 6 (100164), 1-25.
- Beyza, J., Yusta, J., (2022). Characterising the security of power system topologies through a combined assessment of reliability, robustness, and resilience, Energy Strategy Reviews, 43 (100944), 1-10.
- G. Zissis, The R3 concept: reliability, robustness, and resilience [President's Message], IEEE Ind. Appl. Mag. 25 (1) (2019) 5–6.

# References (information)


- Y. Koç, M. Warnier, P. Van Mieghem, R.E. Kooij, F.M.T. Brazier, The impact of the topology on cascading failures in a power grid model, Phys. A Stat. Mech. its Appl. 402 (2014) 169–179,
- Y. Koc, A. Raman, M. Warnier, T. Kumar, in: Structural Vulnerability Analysis of Electric Power Distribution Grids 31, Jun. 2015, pp. 1–20
- K. Kapur, D. Reed, Integration of Reliability, Robustness and Resilience for Engineered System Motivation, Jun. 2014.
- Felderer M., Ramler R., Quality Assurance for AI-based Systems: Overview and Challenges, 10 Feb 2021
- Pennock M.J., Wade J.P., The Top 10 Illusions of Systems Engineering, Procedia Computer Science 44 ( 2015 ) 147 – 154
- Castellani, B., Complexity and its Implications to the Systems Engineering Process, 2014
- Potts M.W., Sartor P.A., Johnson A., Bullock S., Assaying the importance of system complexity for the systems engineering community, Systems Engineering. 2020;23:579–596.
- Osmundson J.S., Huynh T.V., Langford G.O., Emergent Behavior in Systems of Systems, 2008, INCOSE

# References (images)


- Slide 2:
  - Road - PowerPoint images library
  - Tesla Model S - https://www.usatoday.com/story/money/cars/2021/04/26/tesla-elon-musk-autopilot-model-s-texas-crash/7391539002/
  - Crashed Model S – https://www.wired.co.uk/article/wired-awake-130917
- Slide 3:
  - Brain power - PowerPoint images library
- Slide 4:
  - Space - PowerPoint images library
  - Nuclear Reactors - https://engineering.tamu.edu/news/2021/11/NUEN-how-prolonged-radiation-exposure-damages-nuclear-reactors.html
  - Medical – PowerPoint images library
- Slide 6:
  - Road Trip image - https://timesofindia.indiatimes.com/most-searched-products/combos/road-trip-essentials-24-products-you-have-to-take-on-your-next-car-road-trip/articleshow/80735473.cms
  - AI image - https://jtc1info.org/technology/subcommittees/ai/
  - Blind corner sign - https://www.roadtrafficsigns.com/blind-corner-signsTunnel sign - https://printablesign.net/tunnel-ahead-sign-sample/
  - Gravel road sign - https://www.nationalsafetyproducts.com.au/product/warning-gravel-road/

# References (images)

- Slide 11:
  - PPE - PowerPoint images library
- Slide 12:
  - Two-Face (Batman) - https://batman.fandom.com/wiki/Two-Face