

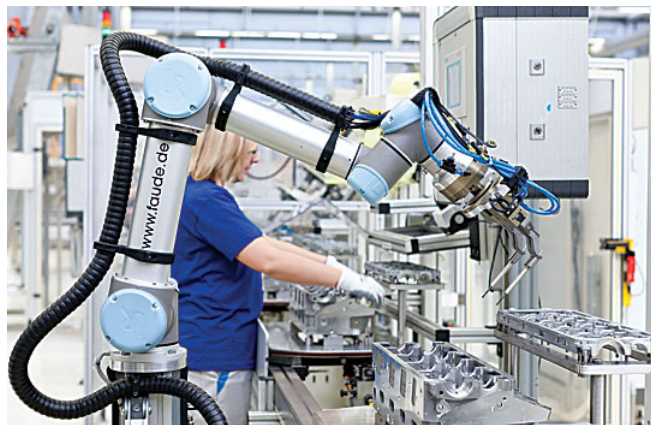
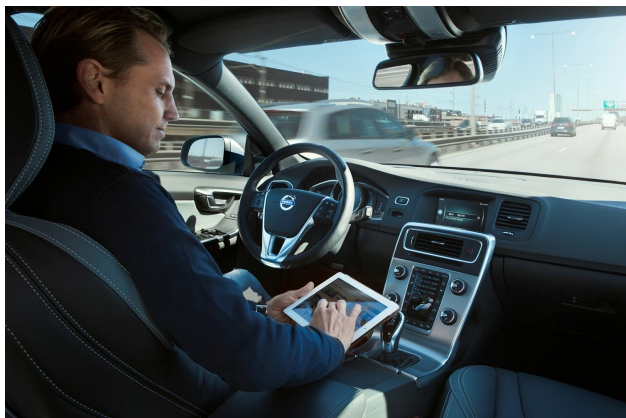
Toward Safe and Accountable AI Systems

Lu Feng

Department of Computer Science

University of Virginia

Growing use of AI/ML in safety-critical systems



Growing concerns on AI safety and accountability

 NBC News

Self-driving Uber car that hit and killed woman did not recognize that pedestrians jaywalk

The automated car lacked "the capability to classify an object as a pedestrian unless that object was near a crosswalk," an NTSB report said...

Nov 9, 2019



 The Independent

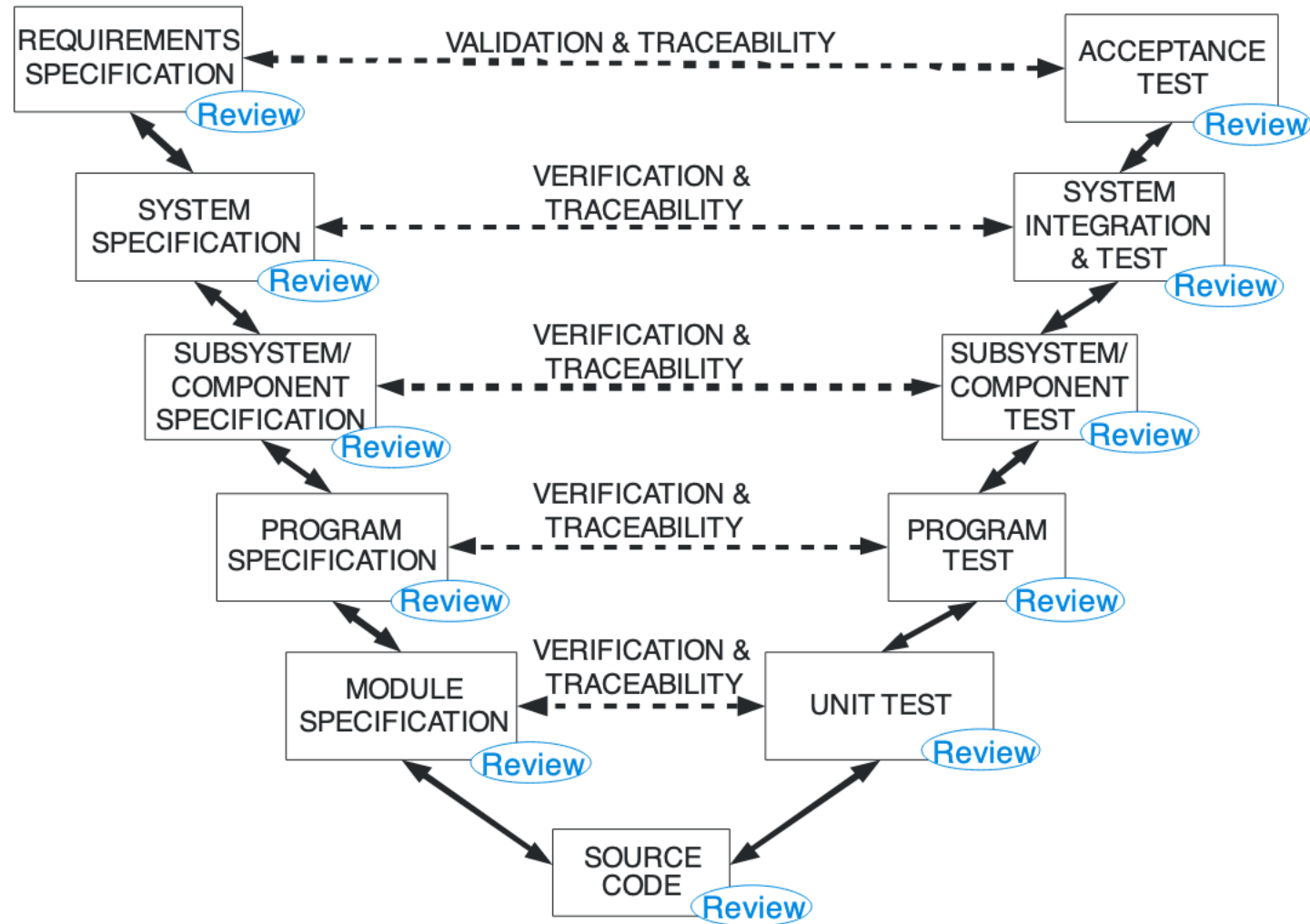
Tesla driver says 'self-driving' tech took control and forced unsafe lane change, causing her to crash

Tesla driver says 'self-driving' tech took control and forced unsafe lane change, causing her to crash. Driver says she tried to take wheel and...

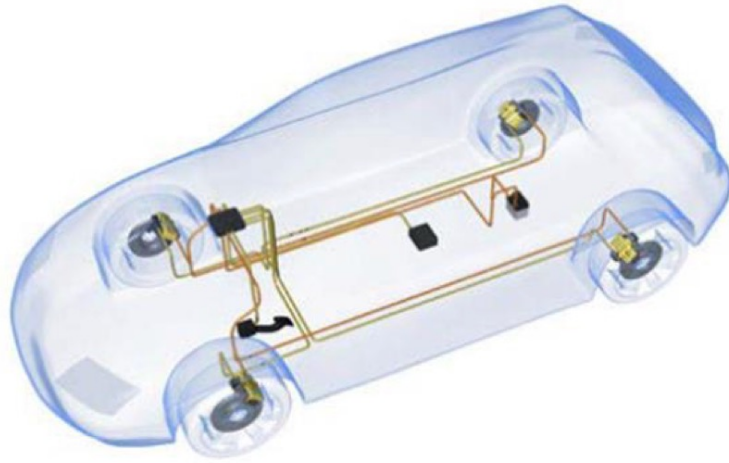
Nov 13, 2021



The “V” software development model

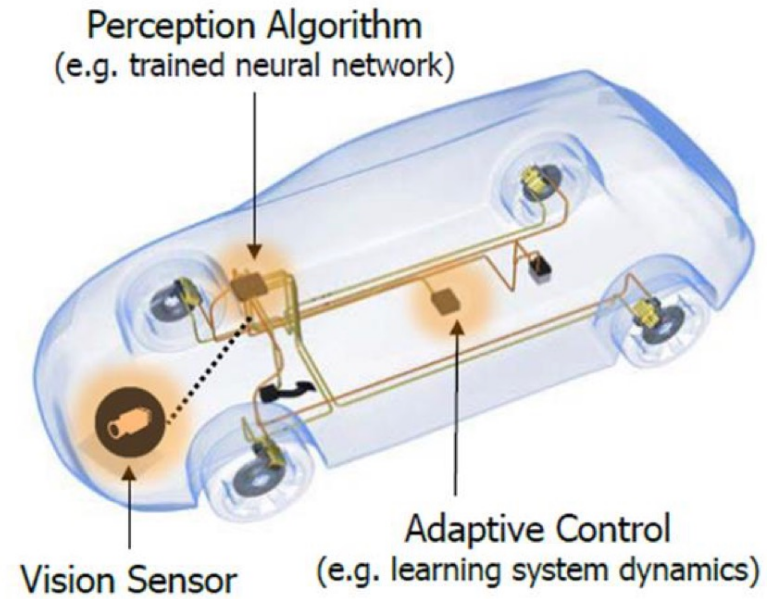


Non-Learning System (e.g. manual brake-by-wire)



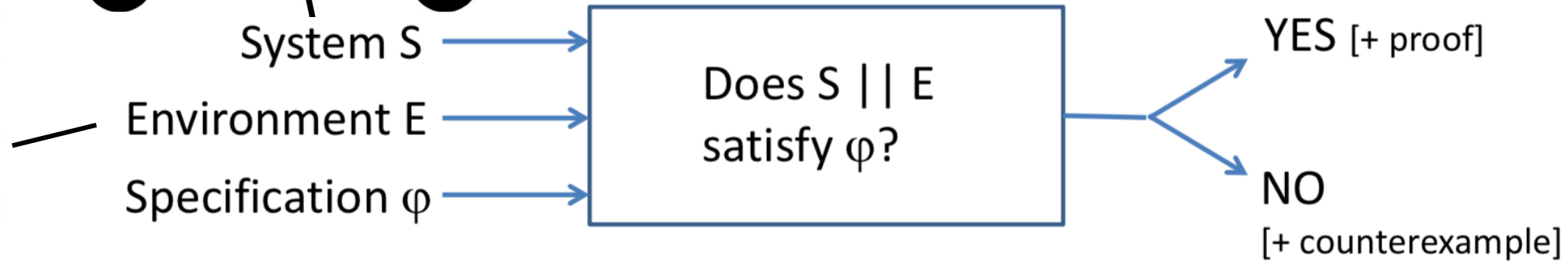
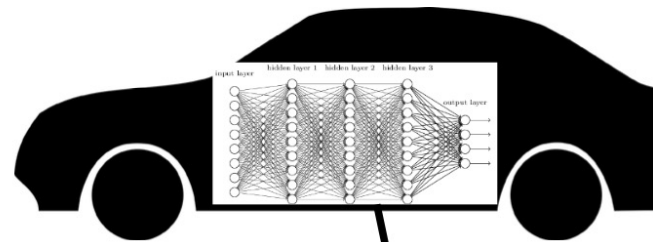
Safety assurance
can be provided

Learning-Enabled Autonomous System (e.g. automated brake-by-wire for collision avoidance)



Safety assurance
can NOT be provided

Challenges for verified AI



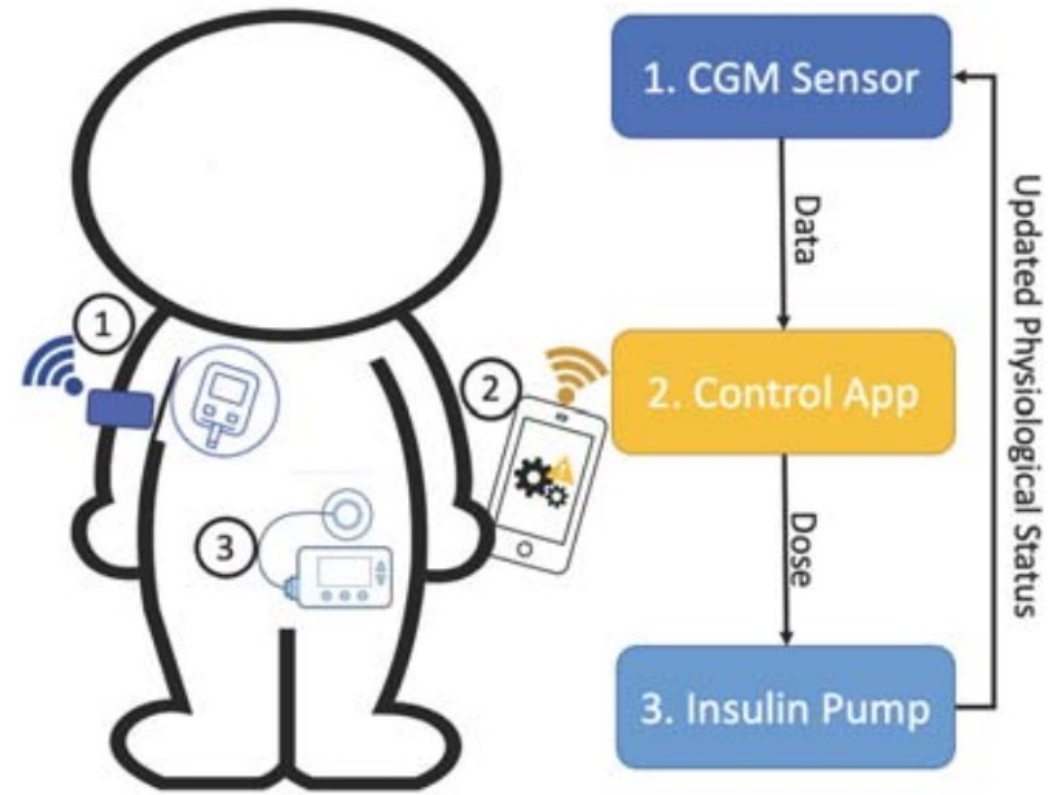
**Need to Search Very High-Dimensional
Input and State Spaces**

AI/ML-based software as a Medical Device



Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan

January 2021




ACM Statement on Principles for Responsible Algorithmic Systems (released on Oct 26, 2022)

It is imperative that algorithmic systems comply fully with established **legal, ethical, and scientific norms** and that the risks of their use be proportional to the specific problems being addressed.

ACM Statement on Principles for Responsible Algorithmic Systems

1. Legitimacy and competency
2. Minimizing harm
3. Security and privacy
4. Transparency
5. Interpretability and explainability
6. Maintainability
7. Contestability and auditability
8. **Accountability and responsibility**
9. Limiting environmental impacts



Public and private bodies should be held accountable for decisions made by algorithms they use, even if it is not feasible to explain in detail how those algorithms produced their results. Such bodies should be responsible for entire systems as deployed in their specific contexts, not just for the individual parts that make up a given system.