## CSC580 AI II HW6

## ProVE Lab for Ensuring Safety with Cyber Physical Approach

Name : Om Prakash Gunja | Student ID: 2131025

### Introduction to CPS, Need for Safety in AI driven approach

The guest lecture by Dr. Stefan Mitsch, titled "ProVE Lab: ProVE, Validate, Execute Theorem Proving, Runtime Monitoring, and Safe Learning", focused on safety verification for Cyber Physical Systems (CPS).

Cyber Physical Systems (CPS) integrate computational intelligence with real world interactions, playing an important role in fields like autonomous transportation, smart infrastructure and industrial automation.

For AI driven CPS, there is a needs for significant safety concern because CPS must go outside in the real world and interact with actual things, unlike how software operates,In few scenarios AI can take decisions that could end up in a bad situation.

To address these challenges, Dr. Stefan Mitsch’s research focuses on theorem proving, model validation, and runtime monitoring to establish formal safety guarantees in AI driven CPS.

### The Challenge of Safety in Cyber Physical Systems

Ensuring safety in AI driven CPS is a really difficult problem because of mainly three reasons or limitations

* Real world Uncertainties : There are inaccuracies in sensors, unpredictable weather, hard to estimate traffic conditions.
* High stakes applications: There are some applications like Autonomous vehicles, aircrafts, industrial automations where a small mistake can lead to severe damage.
* Limitations of traditional testing: And even in testing, you can't simulate the entire real world. And there are many edge cases that could be missed, which can eventually end up in real world testing or execution and bring us terrible damage.

To overcome these limitations, researchers go for formal verification methods that provide mathematical safety guarantees beyond blind testing. Theorem proving and model validation offer a way to analyze all possible system behaviors before deployment.

### Verification Driven AI: Theorem Proving & Model Validation

#### Theorem Proving for Safety Guarantees

Theorem proving is one of the most effective methods for ensuring safety in CPS systems. which is a formal method that mathematically verifies system correctness. KeYmaera X, a widely used theorem prover for hybrid systems, plays a key role in

* Modeling CPS behavior mathematically.
* Proving safety properties, ensuring the system behaves wrt to constraints
* Detecting potential failures before an actual deployment.

In CPU systems, the hybrid system models represent the interplay between discrete, control logic, and continuous physical processes. So, theorem proving ensures that AI driven controllers operate within strict parameters, not just in expected environments, but also in unexpected situations.

#### Model Validation & Runtime Monitoring

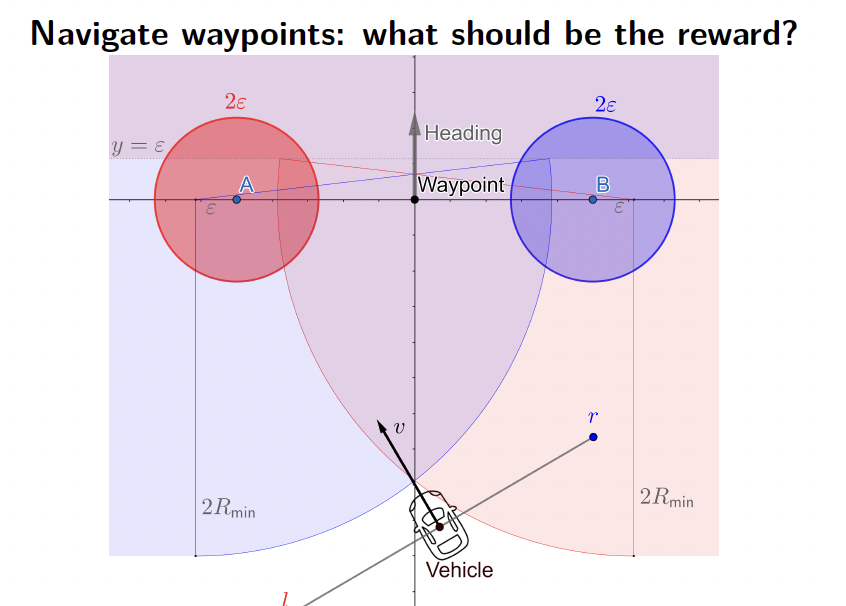
While theorem proving creates formal safety guarantees, the real world environment is unpredictable and changing every time which requires additional validation, like continuous validation. And especially for this, there are a few things like Modelplex and runtime monitors, which provide continuous verification by

* Bridging the gap between formal models and real world execution
* Detecting difference between expected and actual system behavior
* Triggering appropriate or correct actions for specific events to prevent unsafe outcomes

By combining both theorem proving with runtime monitoring, Dr. Mitsch ensures that CPS remain safe throughout their operational lifespan.

### Reinforcement Learning for Control in CPS

So while the above methods provide safety guarantees, reinforcement learning is widely used to optimize decision making, especially in cyber physical systems. RL enables agents to learn patterns using experiences, particularly to control policies through interactions with their specific environment.

If we look at autonomous vehicles, RL assists with high level navigation like selecting waypoints and planning routes, mid level trajectory planning like adapting to road conditions and traffic, and lower level control like steering, acceleration and braking. But reward function design is something that is little more critical to ensure the RL agent is making good decisions. It comes with a new set of challenges, some of them like balancing safety with efficiency, handling sensor noise and ensuring the RL agent generalizes to real world environments.

### Ensuring Safe AI in CPS: Formal methods and Neural Network Verificationpasted-movie.png

#### Provably Safe Neural Network Controllers

Since Ai driven CPS mostly rely on deep neural networks, and their performance is unpredictable at times. Their safety remains challenging, and Dr. Mistch's research integrates the formal methods, which we discussed earlier with neural network verification to create kind of provably safe AI controllers.

#### Important frameworks include:

VerSAILLE: So in this method, we translate neural network behavior into a theorem proving query or several queries.

Mosaic: Mosaic is a framework that combines symbolic reasoning with open loop neural network verification.

### Further Research

On top of the discussion of formal methods and neural network verification in AI driven CPS, there are a few questions that still remain unanswered or requires further research. this study has demonstrated the power of theorem proving, runtime monitoring, and reinforcement learning in ensuring safety, the integration of all these methods into real complex world applications is still something that I think should be explored.

[1] I’ve reviewed a paper that has comprehensive work by Mr. Mhapsekar, and the paper emphasizes mostly on trust for this kind of AI enabled CPS. This paper not only touch on rigorous formal verification but also on improving transparency and accountability.

This paper focuses on explaining how the existing frameworks provide strong guarantees but also fall short in addressing several issues like biases, ethical concerns, and the black box nature of neural networks. Hence giving some idea of integrating xAI which is also known as Explainable AI into formal safety frameworks. Which enables designers and users to understand the BTS of AI decisions.

[2] In parallel, the literature review by Radanliev presents A different approach or a framework that redefines CPS architectures by including cognitive and economic considerations, and these Ideas talk more about how systems can learn and adapt over time. This also raises a few more questions: how are we going to update our old CPS models to deal with uncertain events or ever changing conditions? Further research could involve developing a new framework that combines modern ideas such as learning and adaptation with traditional safety models.

[3] And also I found this Medium article that I felt is really relevant to the current topic by Megasis Network. This article highlights emerging trends such as human machine collaboration, distributed computing by taking considerations of daily life examples such as energy systems, smart maintenance, and automated vehicles. And suggest that such trends indicate that future research may mainly focus on the interplay between distributed intelligence and centralized formal verification techniques. So now it makes you think about how these new trends or paradigms can be integrated to provide real time responsiveness and also safety guarantees at the same time.

So these works consider various fields, like from agriculture to healthcare, making the field more poised for holistic rethinking of AI driven decision making in CPS systems. Furthermore, the ongoing or new research should not only address technical difficulties to make these systems safer but also consider social impacts and implications of such systems in larger scale.

#### References:

[1] Mhapsekar, R. U., Umrani, M. I., Faizan, M., Ali, O., & Abraham, L. (2024). Building Trust in AI-Driven Decision Making for Cyber-Physical Systems (CPS): A Comprehensive Review.

<https://arxiv.org/html/2405.06347v1>

[2] Radanliev, P., De Roure, D., Van Kleek, M., Santos, O., & Ani, U. (2021). Artificial Intelligence in Cyber-Physical Systems. AI & Society, 36, 783–796.

<https://rdcu.be/ec0Fp>

[3] Megasis Network. (2024, March 13). AI and Cyber-Physical Systems: Integrating the Virtual with the Physical. Medium.

<https://megasisnetwork.medium.com/ai-and-cyber-physical-systems-integrating-the-virtual-with-the-physical-2f81a09cb741>

### Relevance to AI

This guest lecture is very relevant to AI because most of it discusses how they are ensuring that decisions made by AI are safe. The lecture also showed the methods of AI like RL and neural network verifications which are applied in real world CPS. This is very similar to the courses I have learned earlier, but here the new concepts like formal verification techniques are introduced, which are really good for safety and the main goal of these techniques is to guarantee safety in very specific and critical applications.

So the parts that include theorem proving and runtime monitoring go beyond just typical simulation based testing and offer more mathematical methods for guarantees.

### Critical Reflection

So after listening to the guest lecture and also going through the research articles and papers, I had learned several new things. One important takeaway for me is that while the methods described in the presentation, which are formal methods and neural network verification, are a strong layer for safety, there could be several challenges in making these techniques understandable and adaptable to real world changes and there are few solutions. for example, xAI integration would help clear out the doubts that revolve about why and how those systems are making decisions. However, I think there still could be some gaps in balancing safety checks with real time performance.

I also want to mention the concerns regarding the limits because I feel like there could be many practical limits for these methods, and the formal models do not capture every single scenario and Will there be any events that require human in the loop? This reflection drives the need for continuous research in the field.

### Conclusion

In the conclusion, the guest lecture provided some interesting direction for AI in CPS. Professor also mentioned how the combination of RL with rigorous formal verification methods gives us a hint of promising safety In areas like automation, autonomous transportation, smart education, and health care. However, there are some other challenges as well, especially for making Deep learning models or AI more transparent and adaptable which can make good decisions even in unexpected conditions. So moving forward, based on my further research I felt like integrating xAI and rethinking systems would be a beneficial steps or upgrade for the current systems, overall the lecture was really valuable and has provided great insights and many questions for students like me.

#### References:

[1] Mhapsekar, R. U., Umrani, M. I., Faizan, M., Ali, O., & Abraham, L. (2024). Building Trust in AI-Driven Decision Making for Cyber-Physical Systems (CPS): A Comprehensive Review.

<https://arxiv.org/html/2405.06347v1>

[2] Radanliev, P., De Roure, D., Van Kleek, M., Santos, O., & Ani, U. (2021). Artificial Intelligence in Cyber-Physical Systems. AI & Society, 36, 783–796.

<https://rdcu.be/ec0Fp>

[3] Megasis Network. (2024, March 13). AI and Cyber-Physical Systems: Integrating the Virtual with the Physical. Medium.

<https://megasisnetwork.medium.com/ai-and-cyber-physical-systems-integrating-the-virtual-with-the-physical-2f81a09cb741>