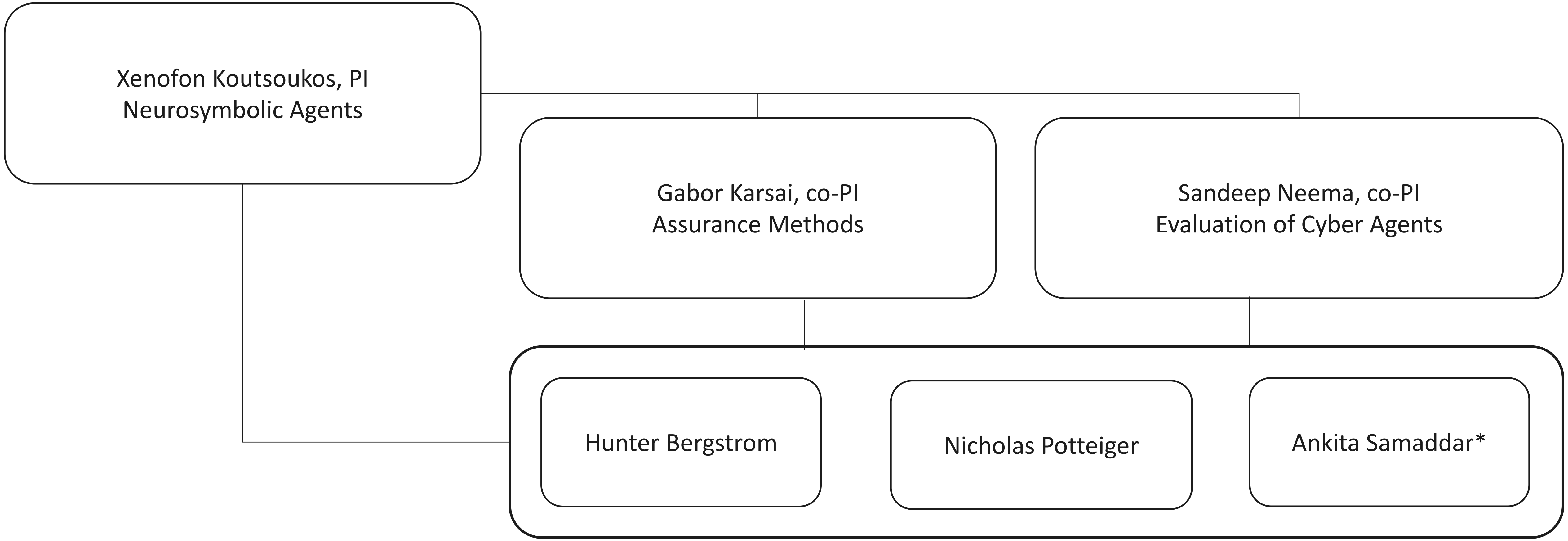
Neurosymbolic Autonomous Agents for Cyber-Defense

Xenofon Koutsoukos
Department of Computer Science
Institute for Software Integrated Systems
Vanderbilt University

Team

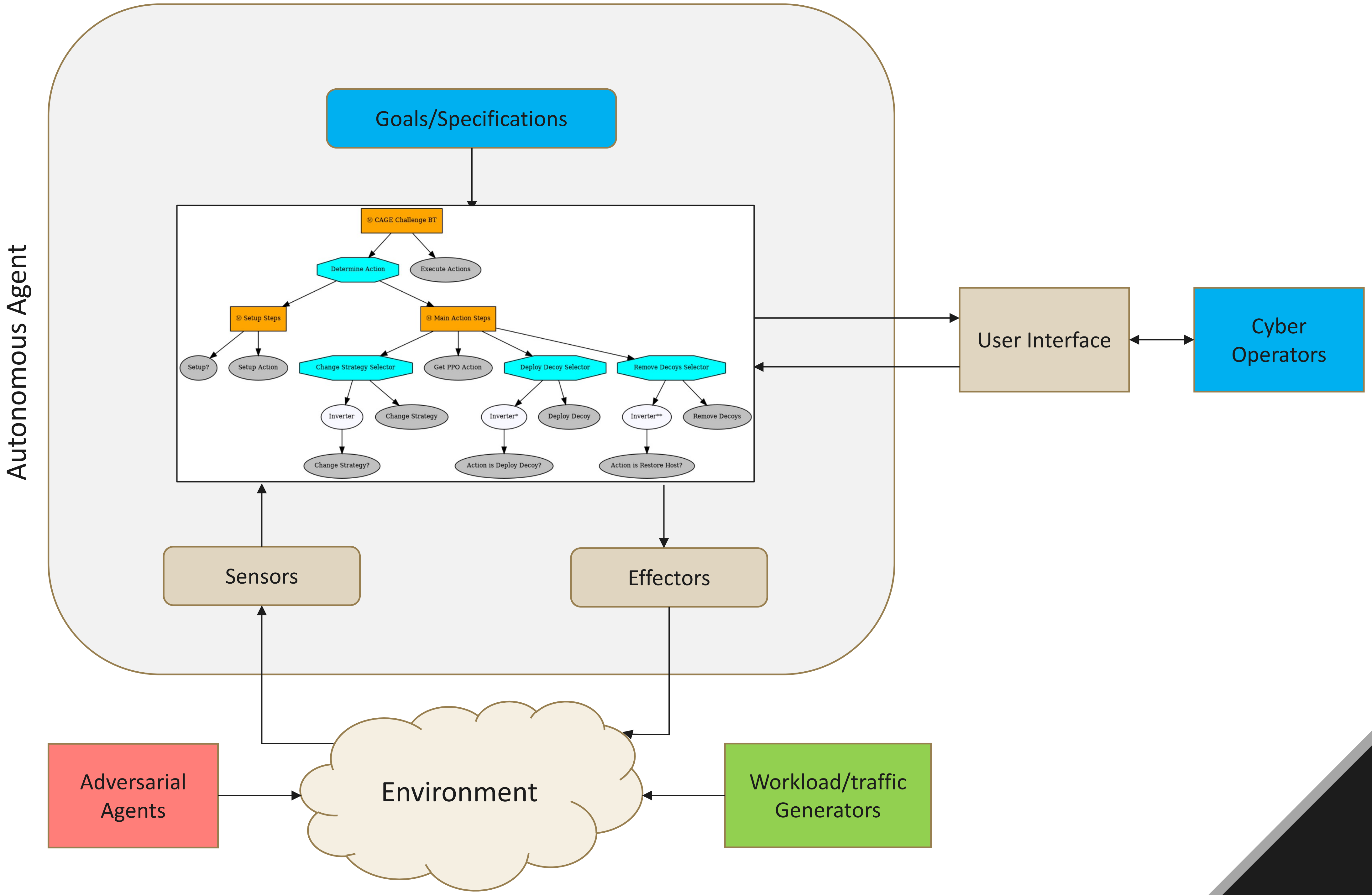


Xenofon Koutsoukos, PI
Neurosymbolic Agents
Gabor Karsai, co-PI
Assurance Methods
Sandeep Neema, co-PI
Evaluation of Cyber Agents
Hunter Bergstrom
Nicholas Potteiger
Ankita Samaddar*

Project Vision and Research Challenges

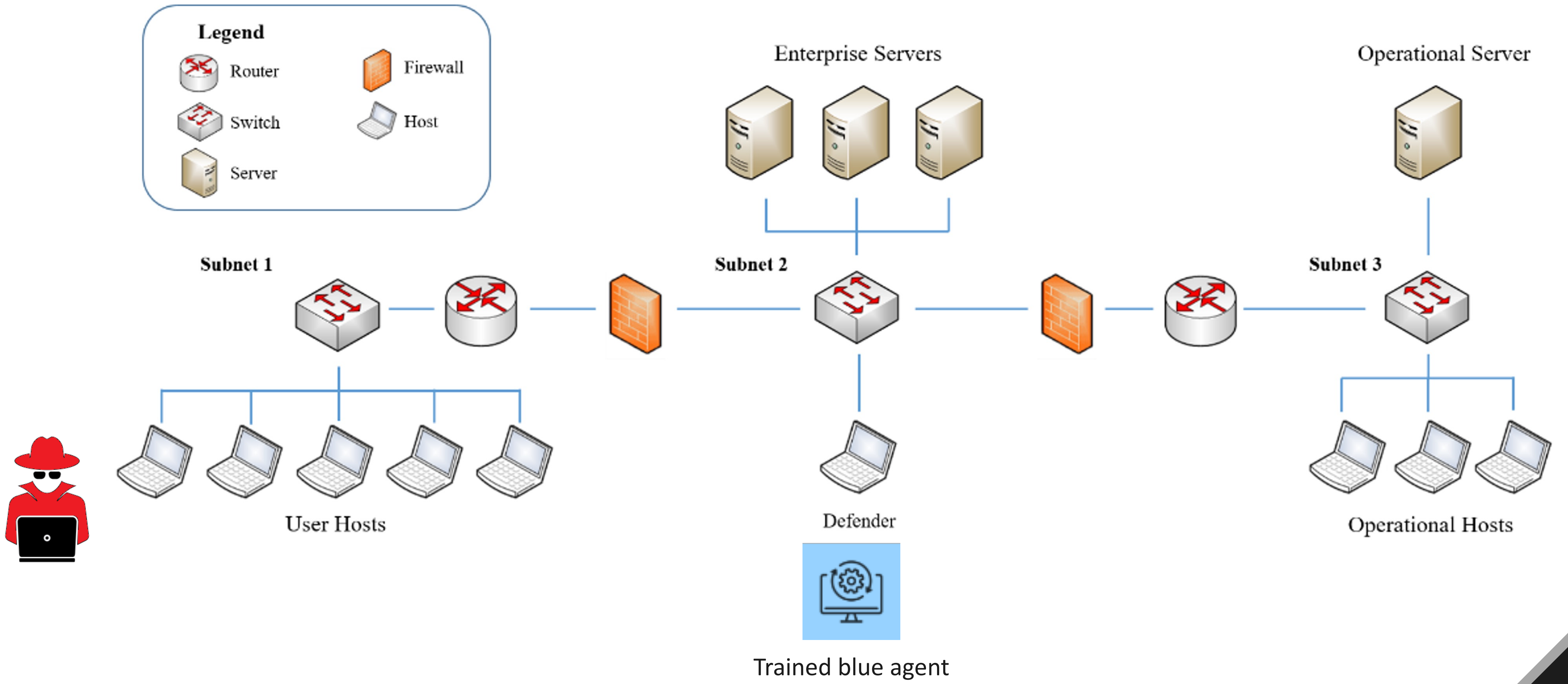
- **Technical Rationale**
  - Autonomous agents for cyber applications need to learn, reason about, and adapt to deploy security mechanisms for defending networked computer systems while maintaining critical operational workflows.
- **Research Challenges**
  - Cyber agents need to complete multiple interdependent tasks over variable length time-intervals.
    - Many tasks can be realized using learning-enabled components (LECs) to handle and uncertainty and variability of the environment.
  - Autonomous cyber agents must continuously explore, improve tasks already learned, learn new tasks, and identify creative ways to synthesize goals, plans, and tasks to increase effectiveness.
  - Robustness and generalizability in new cyber environments is necessary to address novel and fast changing threats.
  - Assurance methods must provide evidence for the correctness of the agents.
  - Interpretability can improve human trust and human-machine teaming.
  - Demonstration and evaluation using a cyber operational environment which is scalable and fast enough to be used in RL training.

Technical Architecture



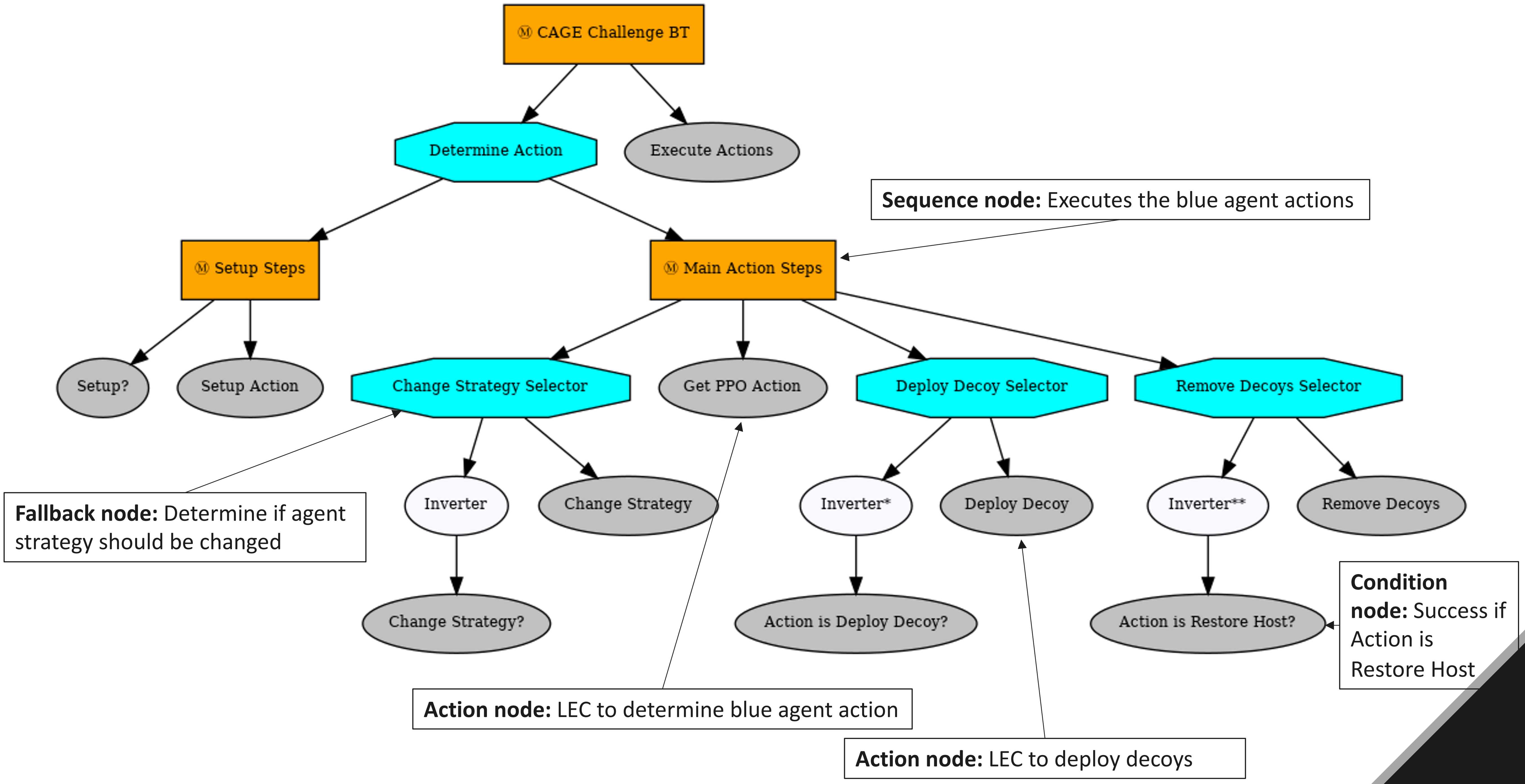
Cyber Operations Research Gym (CybORG)

CYBER AUTONOMY GYM FOR EXPERIMENTATION (CAGE) CHALLENGE 2

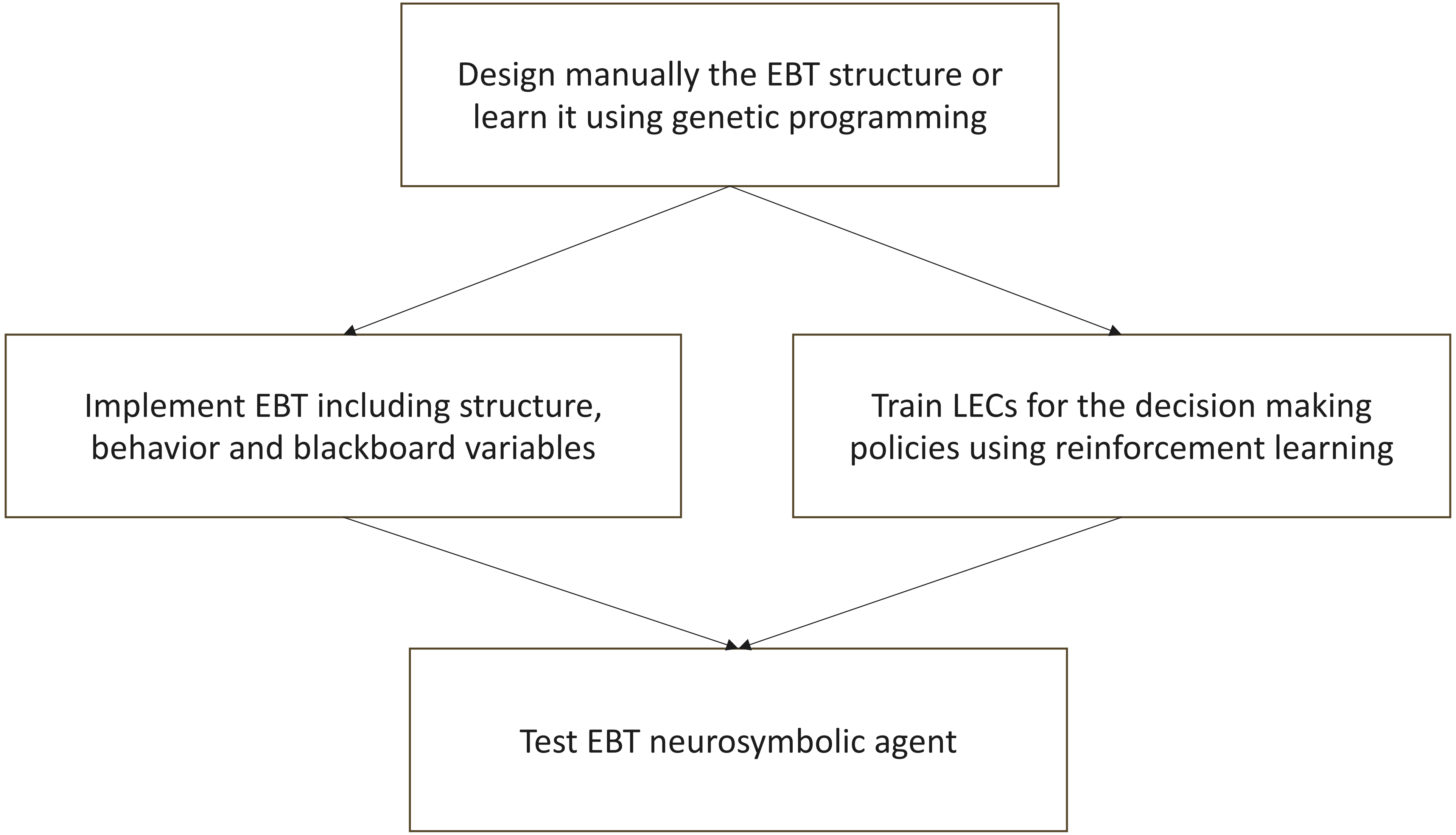


Standen, Maxwell, Martin Lucas, David Bowman, Toby J. Richer, Junae Kim, and Damian Marriott. "Cyborg: A gym for the development of autonomous cyber agents." *arXiv preprint arXiv:2108.09118* (2021).

Evolving Behavior Trees (EBTs)



EBT Design Workflow



Assurance methods for EBTs

- Runtime monitoring algorithms
  - Monitoring deviations of the observed information from the environment and the information that has been used for training the autonomous agent.
  - Integrated in the neurosymbolic model architecture.
- Formally analyze the learning process of the neurosymbolic agents
  - Modeling of the interdependent policies of an agent as interconnected dynamical systems and analyzing the properties using methods from control and system theory.
  - Ensure that agents learn effectively, behave safely, and perform well under various conditions.
- Runtime verification
  - Safety monitors to analyze sequences of sensor readings, state information, and actions.
  - Designed using ML methods.

Demonstration and Evaluation

- CybORG: Cyber Operations Research Gym
  - Configuration scenarios: network topology, operating system, version, services types, applicable CVE's, listening ports, etc.
  - Operational workflows, green and red agents.
  - Other gym environments (e.g., DARPA CASTLE program).

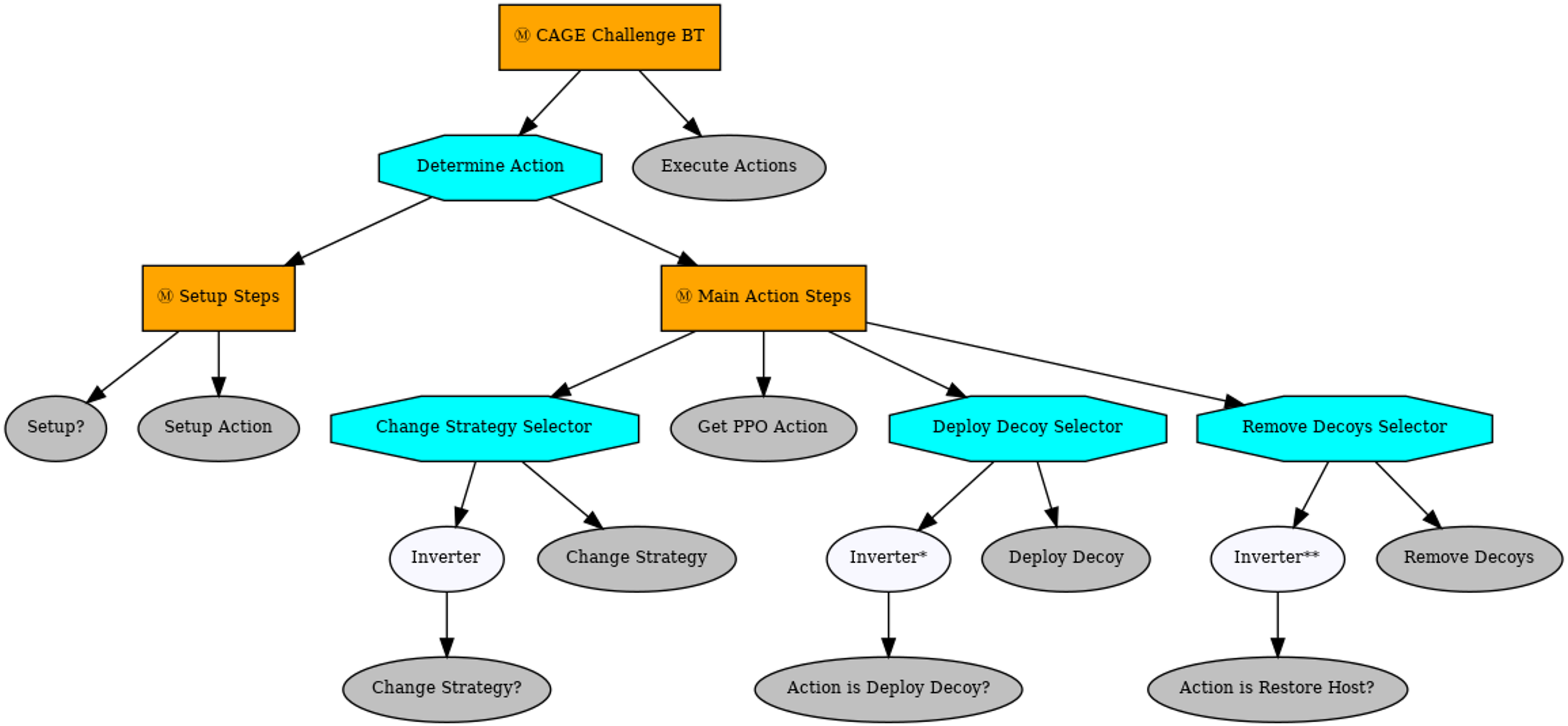
- Evaluation metrics
  - Effectiveness of the training algorithm.
  - Performance of the agents in cyber operations.
  - Interpretability and effectiveness of human-machine teaming.

Preliminary Results: Modified CAGE Challenge 2

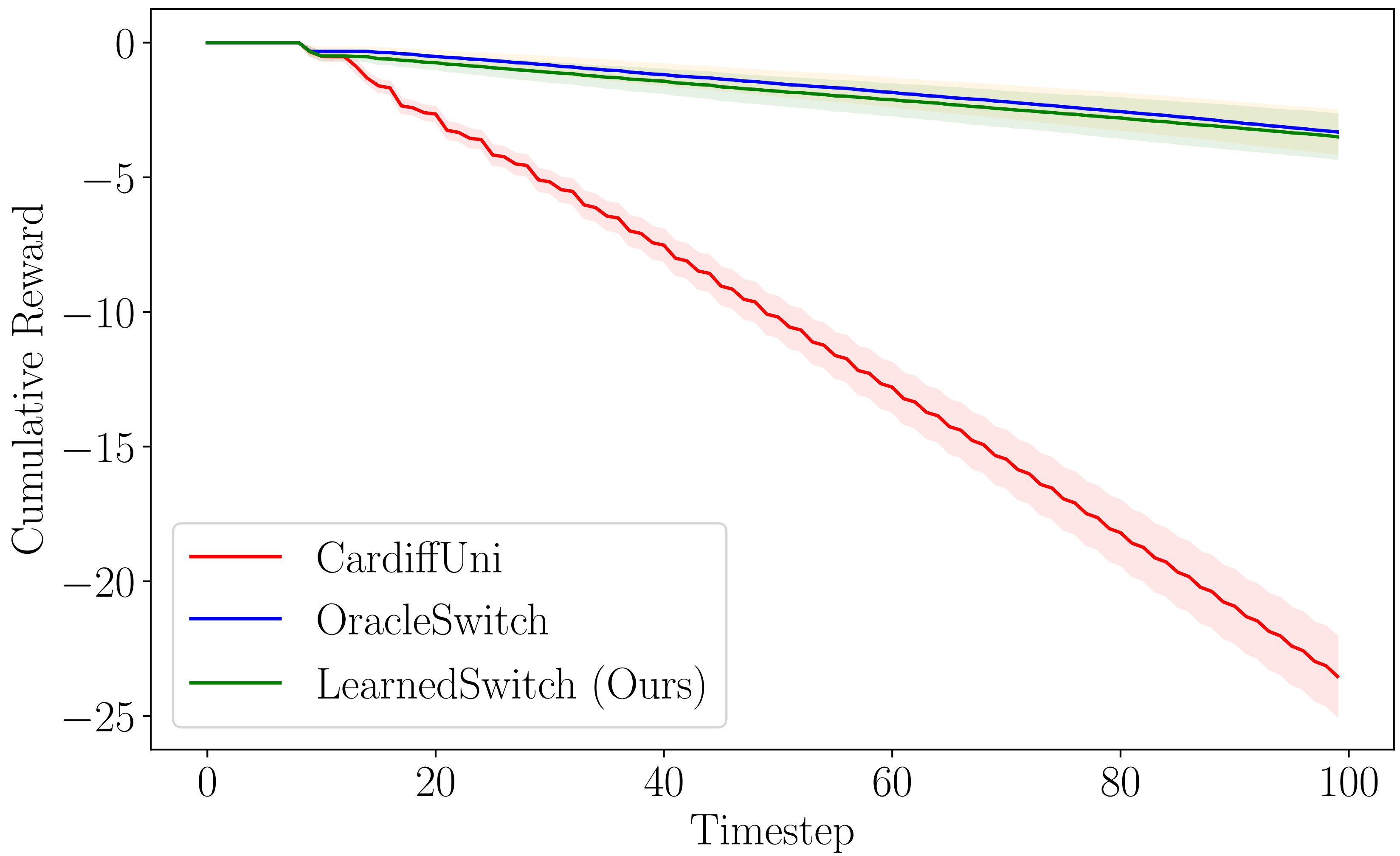
- The red agent executes a Meander (exploration) strategy first.
- Then the red agent switches to a B-line strategy to move directly towards the operational server.
- Existing solutions for blue agents determine their defense strategy in the beginning of the episode.
- The blue agent needs a policy to switch defense strategies during the execution of an episode.

- EBT agent
  - Designed based on the CardiffUni solution
  - Integrate a LEC to detect when the red agent switches strategies
    - LSTM using a sliding window of length 5.
    - Trained using supervised learning.
- Baselines
  - Original CardiffUni solution
  - Switch strategies based on an oracle.

Preliminary Results: EBT Agent

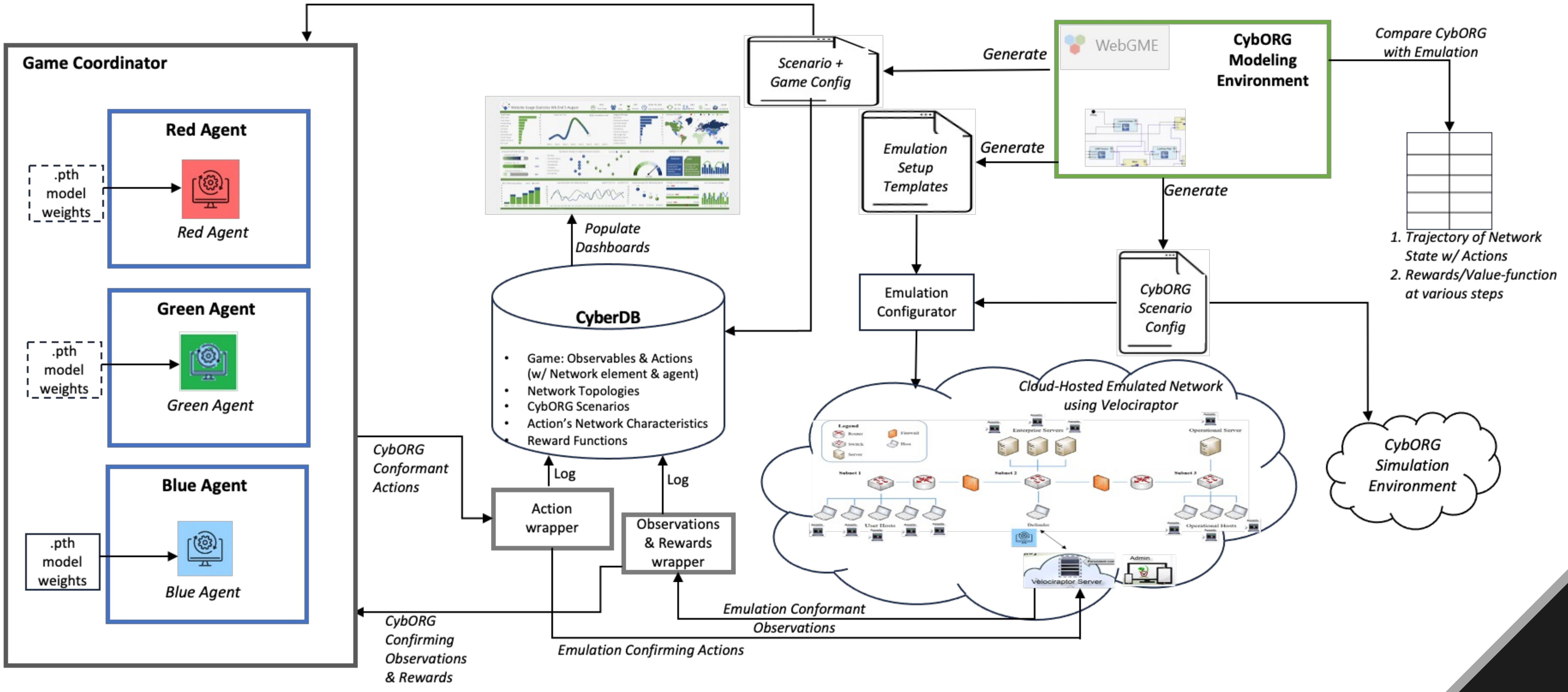


Preliminary Results



- Execute 1000 episodes in CybORG and record the cumulative reward at each timestep.
- Mean and standard deviation over 1000 episodes (smoothing is applied for visual clarity).

Agent Evaluation (DARPA CASTLE)



Explaining EBT Agents using ChatGPT

```
{
  "name": "Deploy Decoy Selector",
  "type": "Fallback",
  "children": [
    {
      "name": "Not Action is Deploy Decoy?",
      "type": "Condition"
    },
    {
      "name": "Deploy Decoy!",
      "type": "Action"
    }
  ]
},
```

- **Prompt:** Can you describe its behavior for the Deploy Decoy Selector?
- **(Partial) Answer:** This setup allows the cyber agent to dynamically decide whether or not to deploy a decoy based on the current strategy or conditions it is facing. The use of a Fallback node here is crucial as it allows for a decision-making process where alternative actions can be considered if conditions are not met.

Extending EBT Agents using ChatGPT

- Threat Detection:
  - Add nodes for real-time threat detection (e.g., anomaly detection).
- Risk Assessment:
  - Implement a behavior for assessing the risk level of different actions or threats.
- Compliance and Policy Enforcement:
  - Include nodes for ensuring compliance with cybersecurity policies and standards.
- Emergency Shutdown:
  - Implement a behavior for emergency shutdown or safe mode activation in case the system detects a severe threat that cannot be mitigated through standard responses.

Conclusions

- Neurosymbolic autonomous agents for cyber defense based on evolving behavior trees
  - Symbolic components captured by the structure of the behavior tree.
  - Neural components are used to realize the various behaviors.
  - Improve robustness and generalization for long-term complex tasks.
  - Improve interpretability and human-machine teaming.
- Assurance methods for neurosymbolic agents
  - Runtime monitoring and verification.
  - Analyze the learning process.
- Demonstration and evaluation
  - CybORG: Cyber operations research gym.
  - Other gym environments (e.g., DARPA CASTLE program).