IEEE Robotics & Automation Society CASE 2025

August 17 – 21, 2025
Los Angeles
California, USA

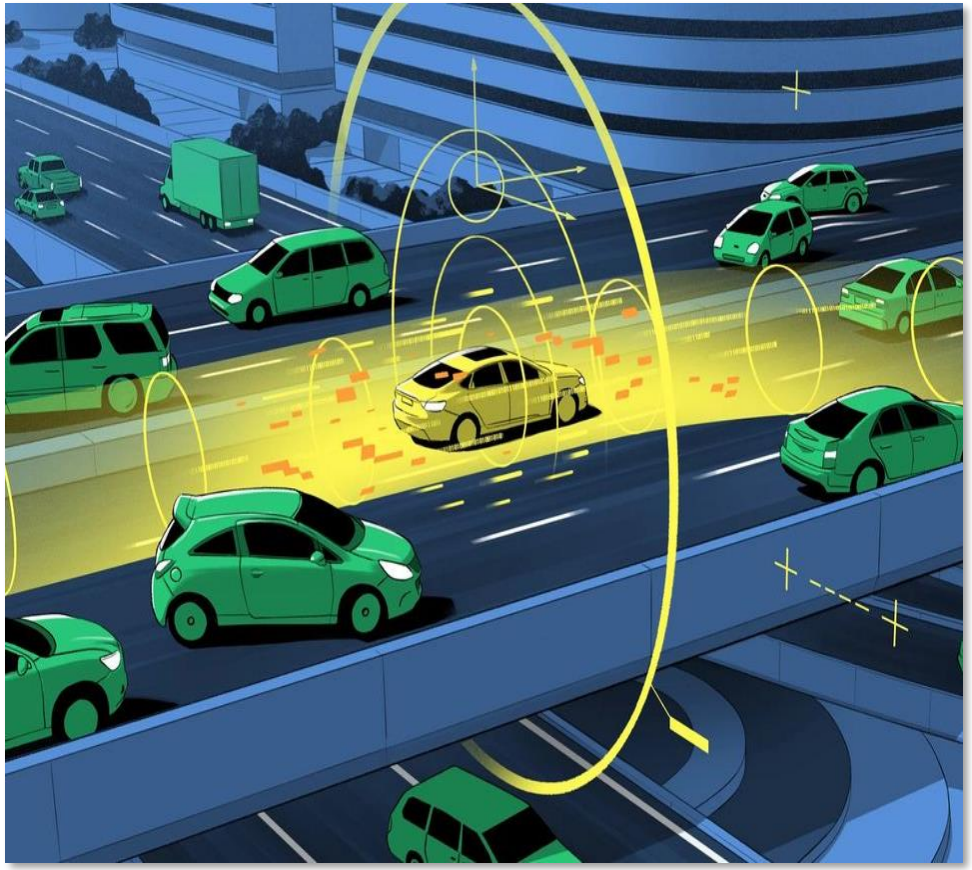
Safety-Guaranteed Policy Composition via Generalized Policy Improvement for Autonomous Vehicles

(Outstanding WiRA Student Paper Award)

Ni Mu, Yao Luan, Qing-Shan Jia
Tsinghua University
mn23@mails.tsinghua.edu.cn

Background of Autonomous Driving

- Promise of Autonomous Driving:
  - ***Safer*** roads: Reduce human error (which causes 94% of accidents) [1].
  - More efficient traffic flow.
  - Accessibility for non-drivers..
- Techniques [2]:
  - Sensors: LiDAR, cameras, radar.
  - Decision-making method: Rule-based, Reinforcement Learning (RL).

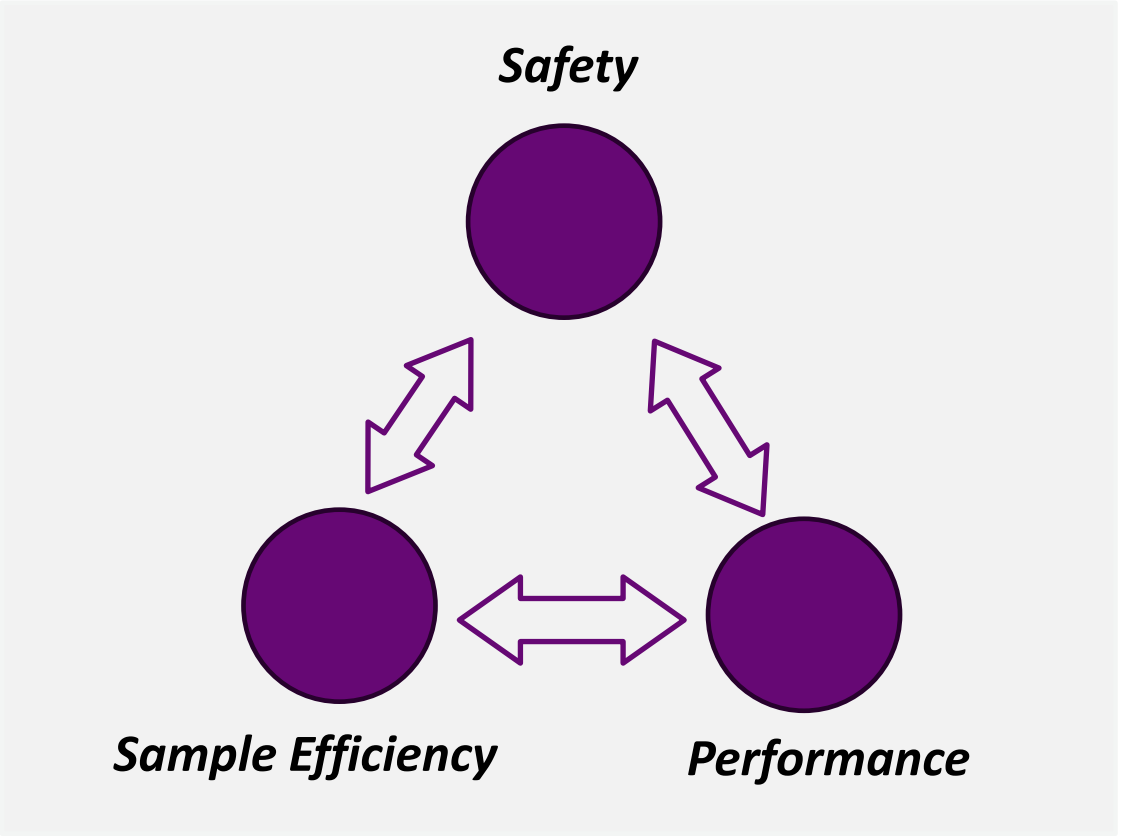


[1] US Department of Transportation National Highway Traffic Safety Administration, Critical Reasons for Crashes Investigated in the National Motor Vehicle Crash Causation Survey, 2015, NHTSA, Washington, DC, USA.
[2] Y. Almalioglu, M. Turan, N. Trigoni, and A. Markham, "Deep learning-based robust positioning for all-weather autonomous driving," Nature machine intelligence, vol. 4, no. 9, pp. 749–760, 2022.

Challenge: The Impossible Triangle

- Existing methods and their trade-offs:
  - **Rule-based**: Safe but with ***poor performance***.
  - **Reinforcement Learning (RL)**: High performance but ***unsafe*** in OOD scenarios.
  - **Hybrid**: Balances both but suffers from ***low sample efficiency***.

Method Type	Sample Efficiency	Safety / Interpretability	Performance
Rule-Based [3], [4]	✓	✓	✗
RL [5], [6]	✗	✗	✓
Hybrid [7 - 9]	✗	✓	✓
Safe-GPI	✓	✓	✓

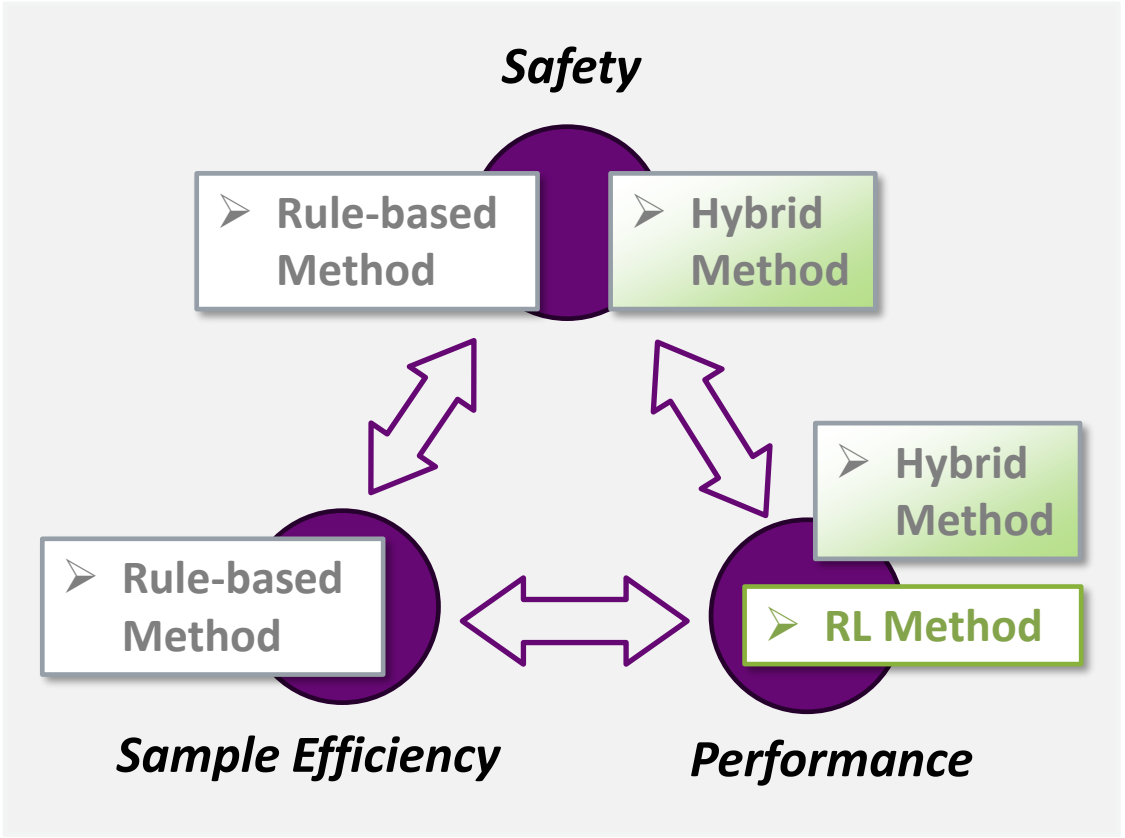


[3] M. Treiber, A. Hennecke, and D. Helbing, "Congested traffic states in empirical observations and microscopic simulations," Physical review E, vol. 62, no. 2, p. 1805, 2000.
[4] A. Kesting, M. Treiber, and D. Helbing, "General lane-changing model mobil for car-following models," Transportation Research Record, vol. 1999, no. 1, pp. 86–94, 2007.
[5] A. E. Sallab, M. Abdou, E. Perot, and S. Yogamani, "Deep reinforcement learning framework for autonomous driving," arXiv preprint arXiv:1704.02532, 2017.
[6] L. Forneris, A. Pighetti, L. Lazzaroni, F. Bellotti, A. Capello, M. Cossu, and R. Berta, "Implementing deep reinforcement learning (drl)-based driving styles for non-player vehicles," International Journal of Serious Games, vol. 10, no. 4, pp. 153–170, 2023.
[7] K. Yang, X. Tang, S. Qiu, S. Jin, Z. Wei, and H. Wang, "Towards robust decision-making for autonomous driving on highway," IEEE Transactions on Vehicular Technology, vol. 72, no. 9, pp. 11 251– 11 263, 2023.
[8] H. Deng, Y. Zhao, Q. Wang, and A.-T. Nguyen, "Deep reinforcement learning based decision-making strategy of autonomous vehicle in highway uncertain driving environments," Automotive Innovation, vol. 6, no. 3, pp. 438–452, 2023.
[9] C.-J. Hoel, K. Driggs-Campbell, K. Wolff, L. Laine, and M. J. Kochenderfer, "Combining planning and deep reinforcement learning in tactical decision making for autonomous driving," IEEE transactions on intelligent vehicles, vol. 5, no. 2, pp. 294–305, 2019.

Challenge: The Impossible Triangle

- Existing methods and their trade-offs:
  - **Rule-based**: Safe but with ***poor performance***.
  - **Reinforcement Learning (RL)**: High performance but ***unsafe*** in OOD scenarios.
  - **Hybrid**: Balances both but suffers from ***low sample efficiency***.

Method Type	Sample Efficiency	Safety / Interpretability	Performance
Rule-Based [3], [4]	✓	✓	✗
RL [5], [6]	✗	✗	✓
Hybrid [7 - 9]	✗	✓	✓
Safe-GPI	✓	✓	✓



[3] M. Treiber, A. Hennecke, and D. Helbing, "Congested traffic states in empirical observations and microscopic simulations," Physical review E, vol. 62, no. 2, p. 1805, 2000.
[4] A. Kesting, M. Treiber, and D. Helbing, "General lane-changing model mobil for car-following models," Transportation Research Record, vol. 1999, no. 1, pp. 86–94, 2007.
[5] A. E. Sallab, M. Abdou, E. Perot, and S. Yogamani, "Deep reinforcement learning framework for autonomous driving," arXiv preprint arXiv:1704.02532, 2017.
[6] L. Forneris, A. Pighetti, L. Lazzaroni, F. Bellotti, A. Capello, M. Cossu, and R. Berta, "Implementing deep reinforcement learning (drl)-based driving styles for non-player vehicles," International Journal of Serious Games, vol. 10, no. 4, pp. 153–170, 2023.
[7] K. Yang, X. Tang, S. Qiu, S. Jin, Z. Wei, and H. Wang, "Towards robust decision-making for autonomous driving on highway," IEEE Transactions on Vehicular Technology, vol. 72, no. 9, pp. 11 251– 11 263, 2023.
[8] H. Deng, Y. Zhao, Q. Wang, and A.-T. Nguyen, "Deep reinforcement learning based decision-making strategy of autonomous vehicle in highway uncertain driving environments," Automotive Innovation, vol. 6, no. 3, pp. 438–452, 2023.
[9] C.-J. Hoel, K. Driggs-Campbell, K. Wolff, L. Laine, and M. J. Kochenderfer, "Combining planning and deep reinforcement learning in tactical decision making for autonomous driving," IEEE transactions on intelligent vehicles, vol. 5, no. 2, pp. 294–305, 2019.

Introduction to GPI

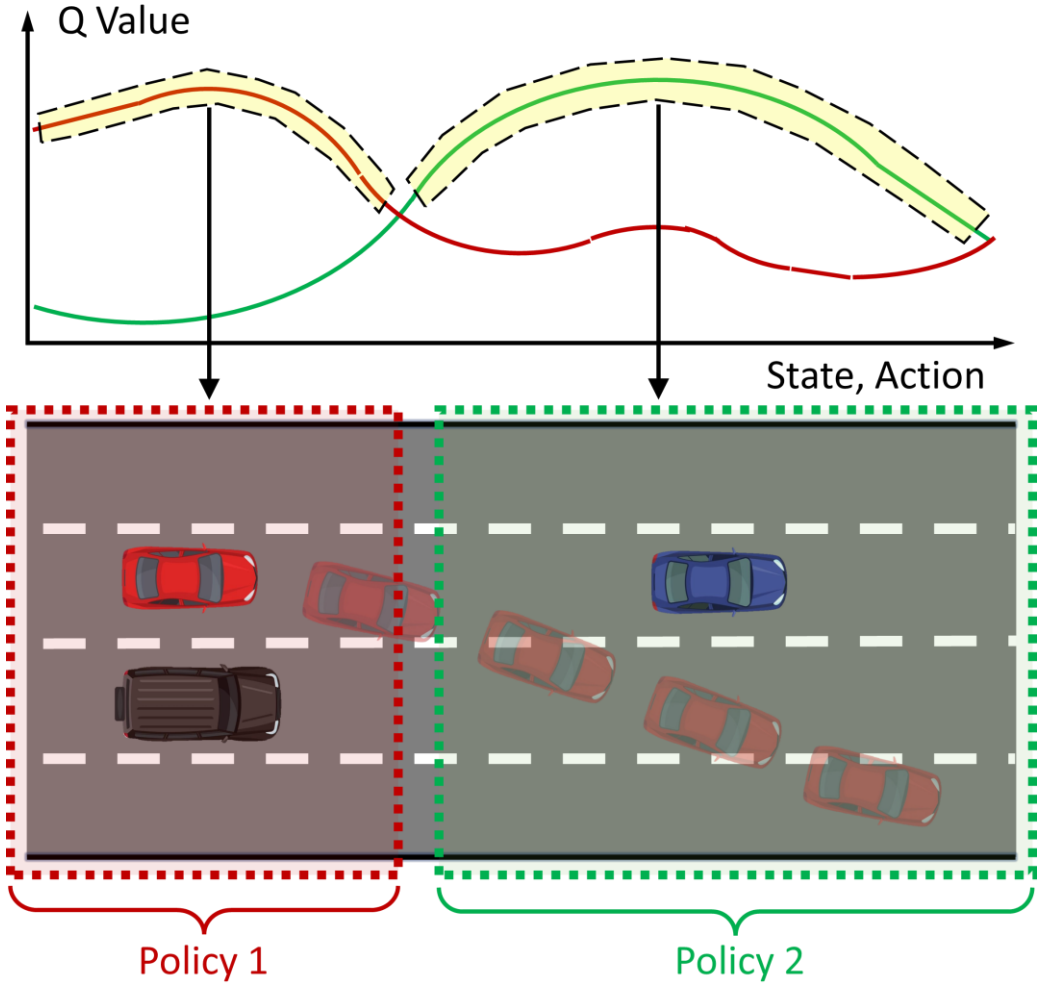
- **Generalized Policy Improvement (GPI):**
  - Combines multiple policies into a single, improved one.
  - GPI picks the action with the highest performance:

$$\pi_{GPI}(s) = \arg\max_a \left[\max_i \tilde{Q}_i(s,a)\right]$$

- **Performance Guarantee**:
  - ***GPI policy is at least as good as the best expert***:

$$Q^{\pi_{GPI}}(s,a) \geq \max_i \tilde{Q}_i(s,a) - \frac{2\gamma\epsilon}{1-\gamma}, \qquad \forall(s,a),$$

  - where $\tilde{Q}_i$ is the approximate Q function, satisfying $\left|\tilde{Q}_i - Q_i\right| \leq \epsilon$, and γ is the discounted factor.



Safe-GPI: Making GPI Safe

- Safe-GPI adds two mechanisms to ensure safety:
  - 1. Always include ***a provable safe policy*** in the GPI's policy library.
  - 2. ***Penalize collision in the reward function*** to discourage unsafe actions.
- Theoretical guarantees of Safe-GPI:
  - 1. **Safety**: Unsafe actions are provably avoided.
  - 2. **Performance**: Preserves GPI's performance bounds.

Theorem 2 (Value Function Bound). *The value function of the composite policy* π_{comp} *satisfies:*

$$V^{\pi_{comp}}(s) \geq \max_i V^{\pi_i}(s) - \frac{2\gamma\epsilon}{1-\gamma} \quad \forall s, \tag{8}$$

where ϵ *bounds the approximation error of Q-functions and* γ *is the discount factor.*

Theorem 3 (Collision Avoidance Guarantee). *Let* ϵ *be the maximum approximation error of Q-functions and* γ *the discount factor. Assume that a collision terminates the episode with reward* $r_{coll} < 0$*, and non-collision transitions yield* $r_{safe} \geq 0$*. If the collision penalty satisfies:*

$$r_{coll} < -\frac{4\epsilon}{1-\gamma}.$$

Then, for any state s, the composite policy will not select actions that lead to a collision.

Experiments: Testing Safe-GPI in Simulated Traffic

- **Environment**:
  - ***Highway-env*** [10] simulation with 2-4 lanes.
  - Realistic traffic: vehicle speeds randomized (23-25 m/s).
- **Baselines**:
  - ***Rule-based***: IDM + MOBIL [3,4].
  - ***Planning method***: Optimistic Planning [11].
  - ***Deep RL***: DDQN [7].
- **Safe-GPI Policy Library**:
  - ***Rule-based***: IDM+MOBIL for safety.
  - ***Planning-based***: Optimistic Planning for performance.
  - *No method that requires training is included.*

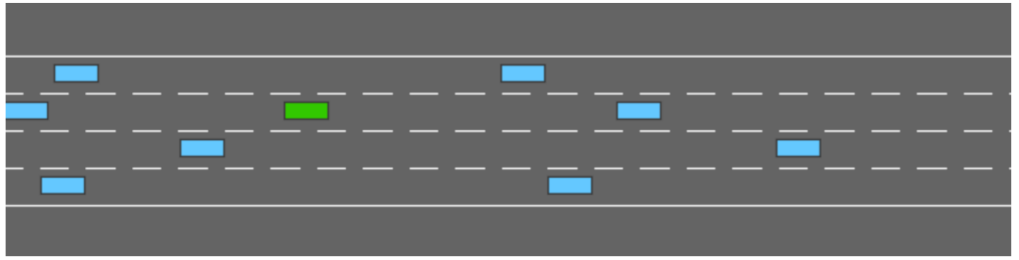


Fig. 2. An illustration of the Highway-env benchmark.

TABLE II

HYPERPARAMETERS USED IN THE IMPLEMENTATION.

Hyperparameter	Value
Network Architecture	[256, 256]
Activation Function	ReLU
Learning Rate	5×10^{-4}
Batch Size	256
Exploration Epsilon	0.1
Collision Reward Coefficient λ_c	-10
Velocity Reward Coefficient λ_v	0.4
Discount Factor γ	0.9

[10] E. Leurent, "An environment for autonomous driving decision-making," https://github.com/eleurent/highway-env, 2018.
[11] J.-F. Hren and R. Munos, "Optimistic planning of deterministic systems," in European Workshop on Reinforcement Learning. Springer, 2008, pp. 151–164.

Experimental Results

- **Performance**: Safe-GPI matches RL methods *(Safe-GPI 9.81 vs. DDQN 8.37)*.
- **Safety**: Collision rate significantly lower than RL methods *(Safe-GPI 0.67 vs. DDQN 1.38)*.
- **Efficiency**: Safe-GPI requires 1/5 training time compared to RL *(Safe-GPI 0.2h vs. DDQN 1.18h)*.

- Safe-GPI balances performance, safety, and efficiency, ***solving the "Impossible Triangle."***

TABLE III

EPISODE RETURNS OF SAFE-GPI AND BASELINES.

	NL = 2	NL = 3	NL = 4
Safe-GPI	1.57 ± 0.51	9.81 ± 1.80	11.11 ± 1.72
DDQN	1.59 ± 0.25	8.37 ± 0.58	10.11 ± 1.20
OP	-5.68 ± 0.59	6.40 ± 1.02	7.78 ± 4.35
IDM+MOBIL	-0.32 ± 0.62	3.31 ± 0.77	4.78 ± 0.40

TABLE IV

COLLISION RATES OF SAFE-GPI AND BASELINES.

	NL = 2	NL = 3	NL = 4
Safe-GPI	0.84 ± 0.56	0.61 ± 0.44	0.67 ± 0.42
DDQN	0.97 ± 0.25	0.77 ± 0.58	1.38 ± 0.28
OP	9.72 ± 1.46	1.91 ± 0.27	1.77 ± 1.39
IDM+MOBIL	0.85 ± 0.20	0.07 ± 0.15	0.00 ± 0.00

TABLE V

THE TRAINING TIME (HOURS) OF SAFE-GPI AND DDQN.

	NL = 2	NL = 3	NL = 4
Safe-GPI	0.08 ± 0.57	0.15 ± 0.09	0.20 ± 0.02
DDQN	0.98 ± 0.03	1.10 ± 0.02	1.18 ± 0.02

Summary: Safe-GPI Solves the Impossible Triangle

- The challenge of ***"Impossible Triangle":***
  - Existing approaches in autonomous driving fail to balance ***1. safety***, ***2. performance***, and ***3. sample efficiency***.
- **Safe-GPI** achieve all three, by dynamically combining multiple policies with ***theoretically provable safety guarantees***.
- Experimental results:
  - ***1. Safety***: Lower collision rates. ✓
  - ***2. Performance***: Matches RL methods. ✓
  - ***3. Efficiency***: Trains 5× faster. ✓

Thank You!

Ni Mu

2025.08.20

mn23@mails.tsinghua.edu.cn