- I. Hasuo, C. Eberhart, J. Haydon et al.: Goal-Aware RSS for Complex Scenarios via Program Logic. *IEEE Trans. Intell. Veh.* 8(4): 3040-3072 (2023)
- C. Eberhart, J. Dubut, J. Haydon and I. Hasuo: Formal Verification of Safety Architectures for Automated Driving, *2023 IEEE Intelligent Vehicles Symposium (IV)*, 2023, pp. 1-8,
- J. Haydon, M. Bondu, C. Eberhart, J. Dubut, I. Hasuo: Formal Verification of Intersection Safety for Automated Driving, *2023 IEEE International Conference on Intelligent Transportation Systems (ITSC)*, 2023.

# Proving Safety of Automated Driving Vehicles

## Formalization of RSS with Program Logic

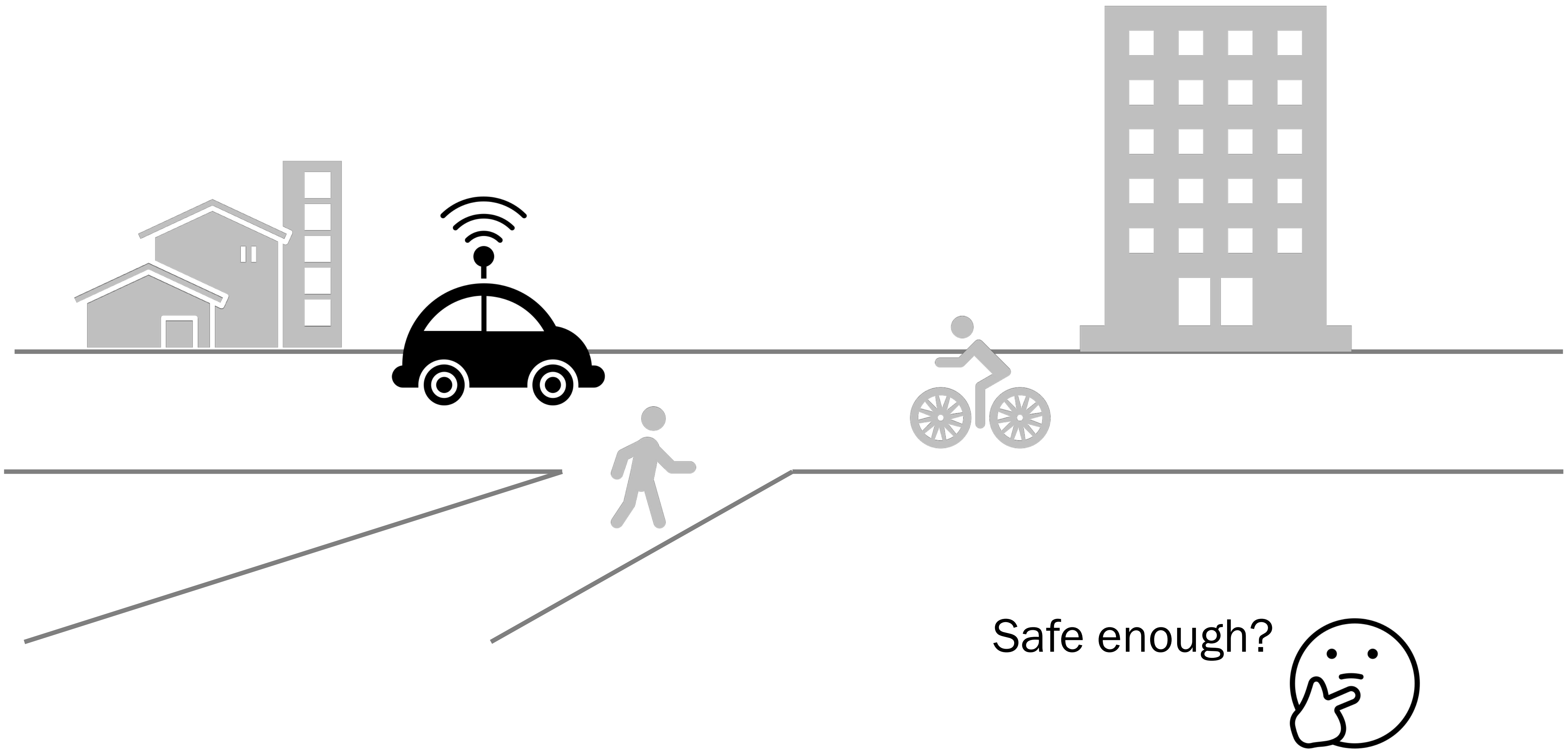

**Ichiro Hasuo**

National Institute of Informatics, Tokyo, Japan
SOKENDAI (The Graduate University for Advanced Studies), Japan

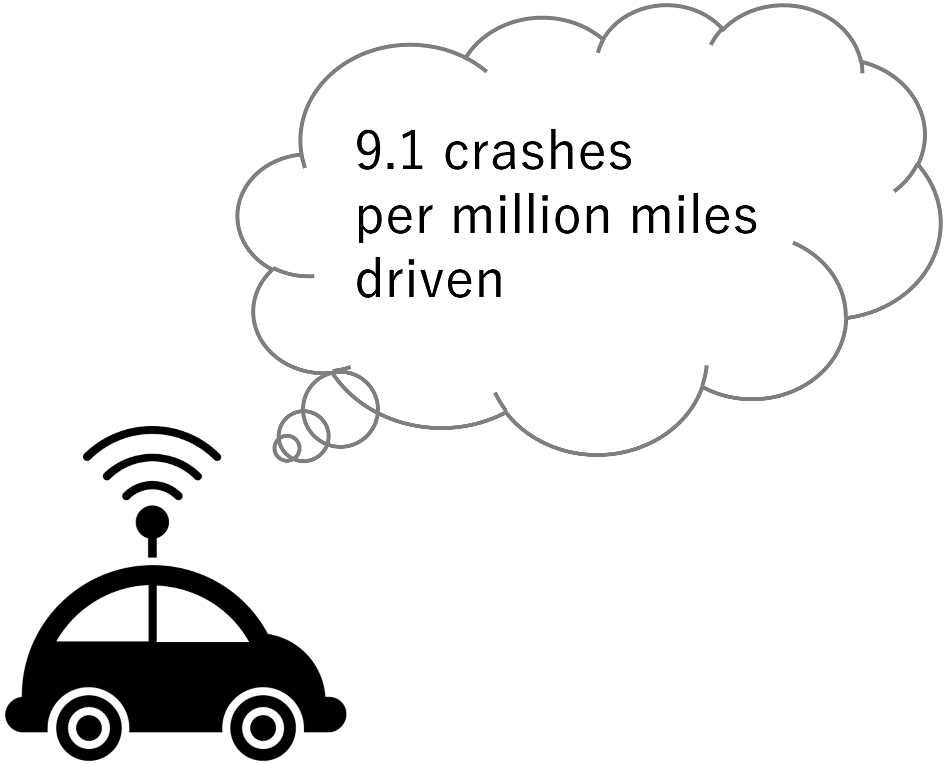

Based on works with **Clovis Eberhart**, **James Haydon**, **Jeremy Dubut**, and many others

# Outline

- A non-technical overview
- Technical contributions: the logic dFHL
- Perspectives, practical & theoretical


Safe enough?

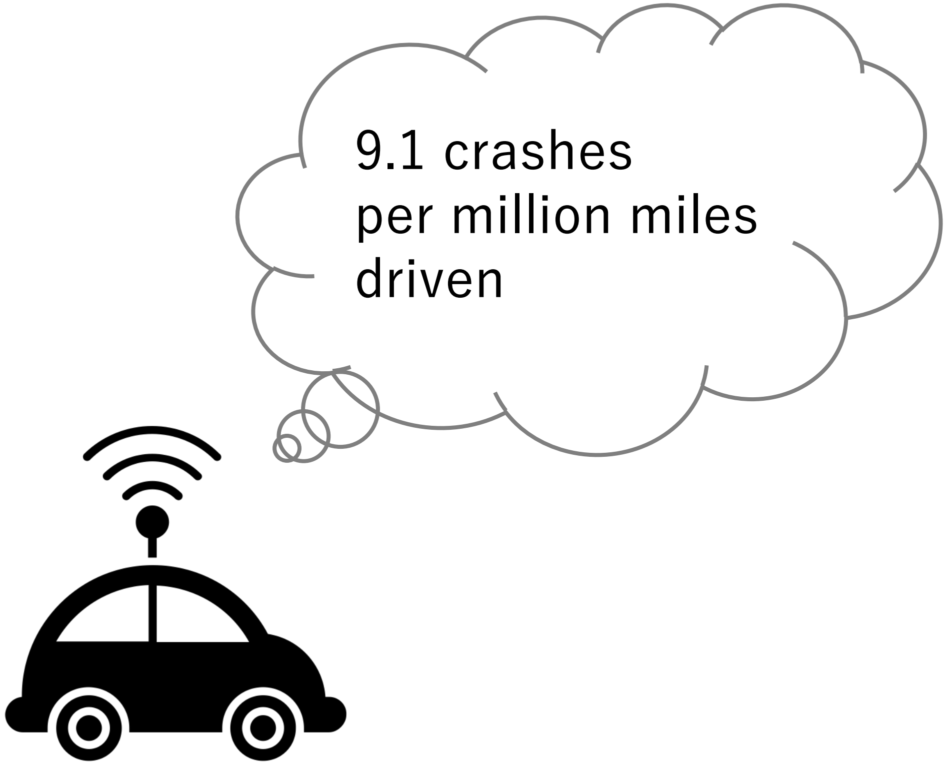
Guarantee
by statistical data




Guarantee
by testing and simulation

Guarantee strong enough?

## Guarantee
## by statistical data




## Guarantee
## by testing and simulation

Explainability?

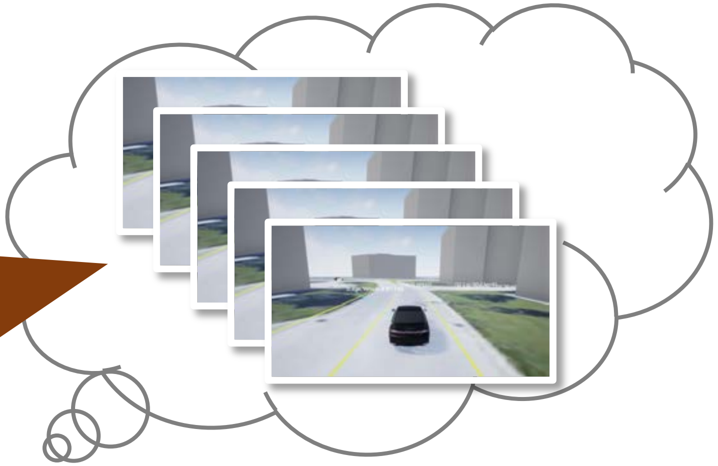
Guarantee strong enough?

Guarantee
by statistical data

Guarantee
by testing and simulation

In particular, on the scenario set:
**how extensive is enough?**

→ setting a **standard** is nontrivial



Explainability?

***Proof.***
We prove the first statement. The rest is shown symm

Let $S \subseteq L$ be an arbitrary subset. We let $S^{\downarrow}$ be th that is,

$$S^{\downarrow} := \{y \in L \mid y \sqsubseteq s \text{ for each } s \in$$

Since $S^{\downarrow} \subseteq L$ is a subset of $L$, it has its supremum | semilattice $(L, \sqsubseteq)$. We claim that $\bigsqcup S^{\downarrow}$ is the infimum

To prove the claim, it suffices to show the two-way acterization in (2.1), that is, we need to show

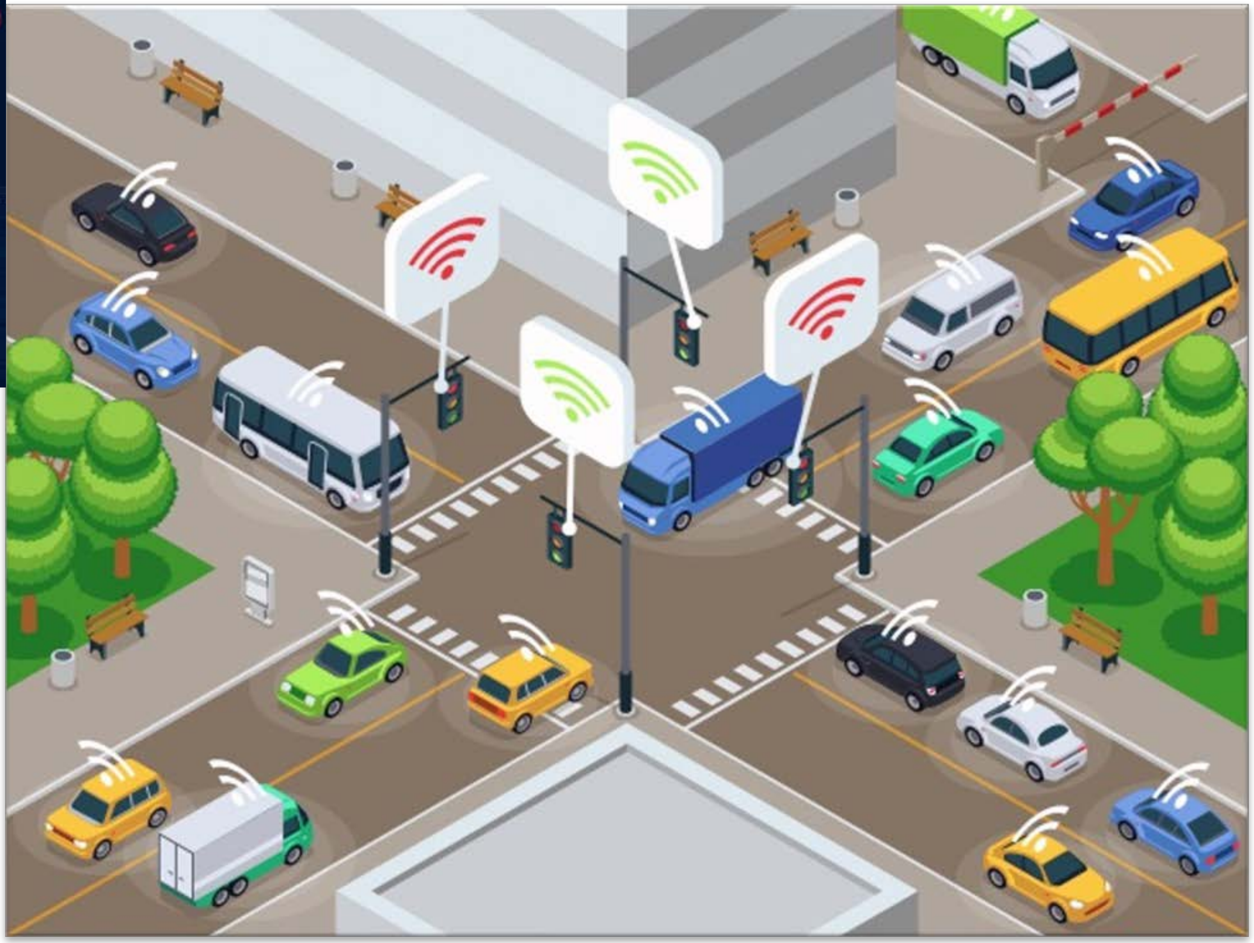
$$\frac{y \sqsubseteq s \quad \text{for each } s \in S}{y \sqsubseteq \bigsqcup S^{\downarrow}}.$$

For the downward implication in **??**,

$$\begin{array}{lll} y \sqsubseteq s \quad \text{for each } s \in S & & \\ \implies \quad y \in S^{\downarrow} & & \text{by def. of } S^{\downarrow} \\ \implies \quad y \sqsubseteq \bigsqcup S^{\downarrow} & & \text{since } \bigsqcup S^{\downarrow} \text{ is an u} \end{array}$$

For the upward implication in **??**, we first observe

$$\bigsqcup S^{\downarrow} \sqsubseteq s \quad \text{for each } s \in S.$$

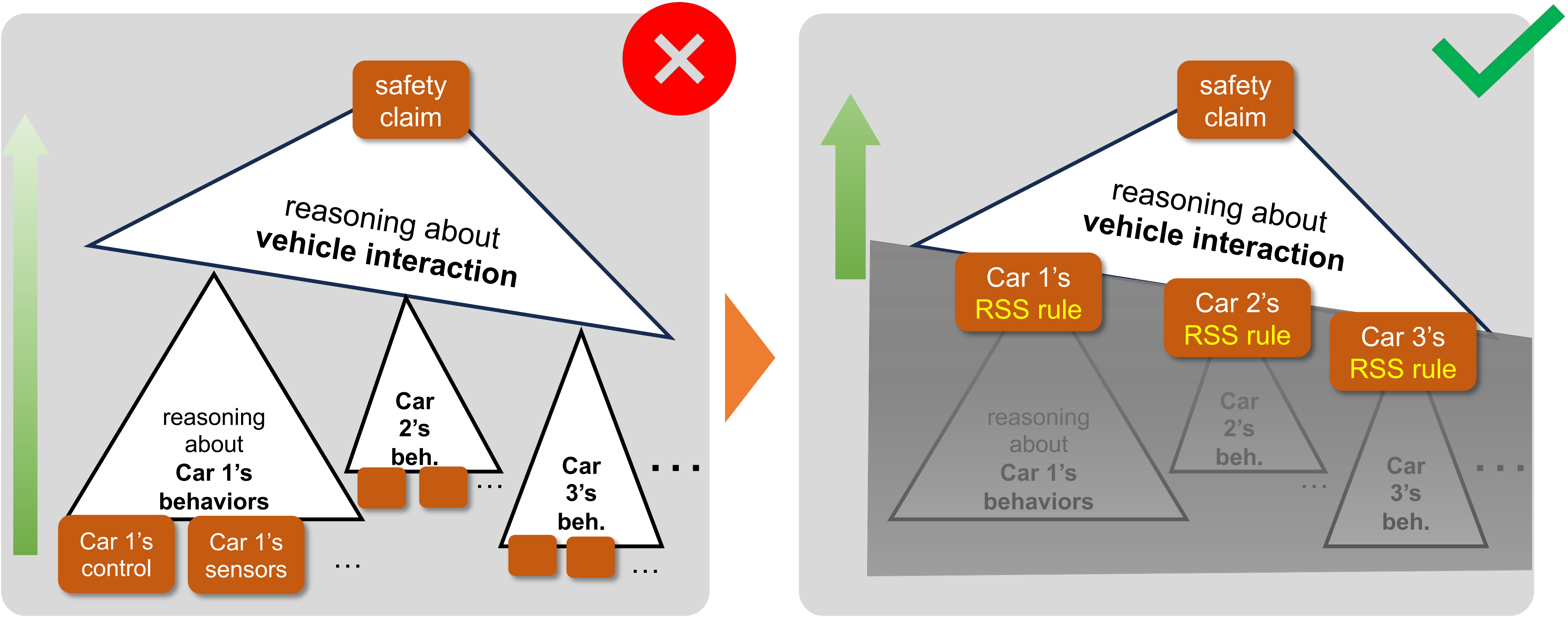



Mathematical safety proofs would certainly be great...

But are they ever feasible?

# Responsibility-Sensitive Safety (RSS)

[Shalev-Shwartz et al., arXiv preprint, 2017]




- **Full safety proofs are infeasible**
  - Lack of white-box models
  - Ultimate safety claim is too far

- Ignore the internal working of individual vehicles
- Instead, impose "behavioral constracts" on them
  - Called **RSS rules**. "Mathematical traffic laws"
- Mathematical proofs assume rule compliance ➔ feasible

# RSS Rule, an Example

[Shalev-Shwartz et al., arXiv preprint, 2017]

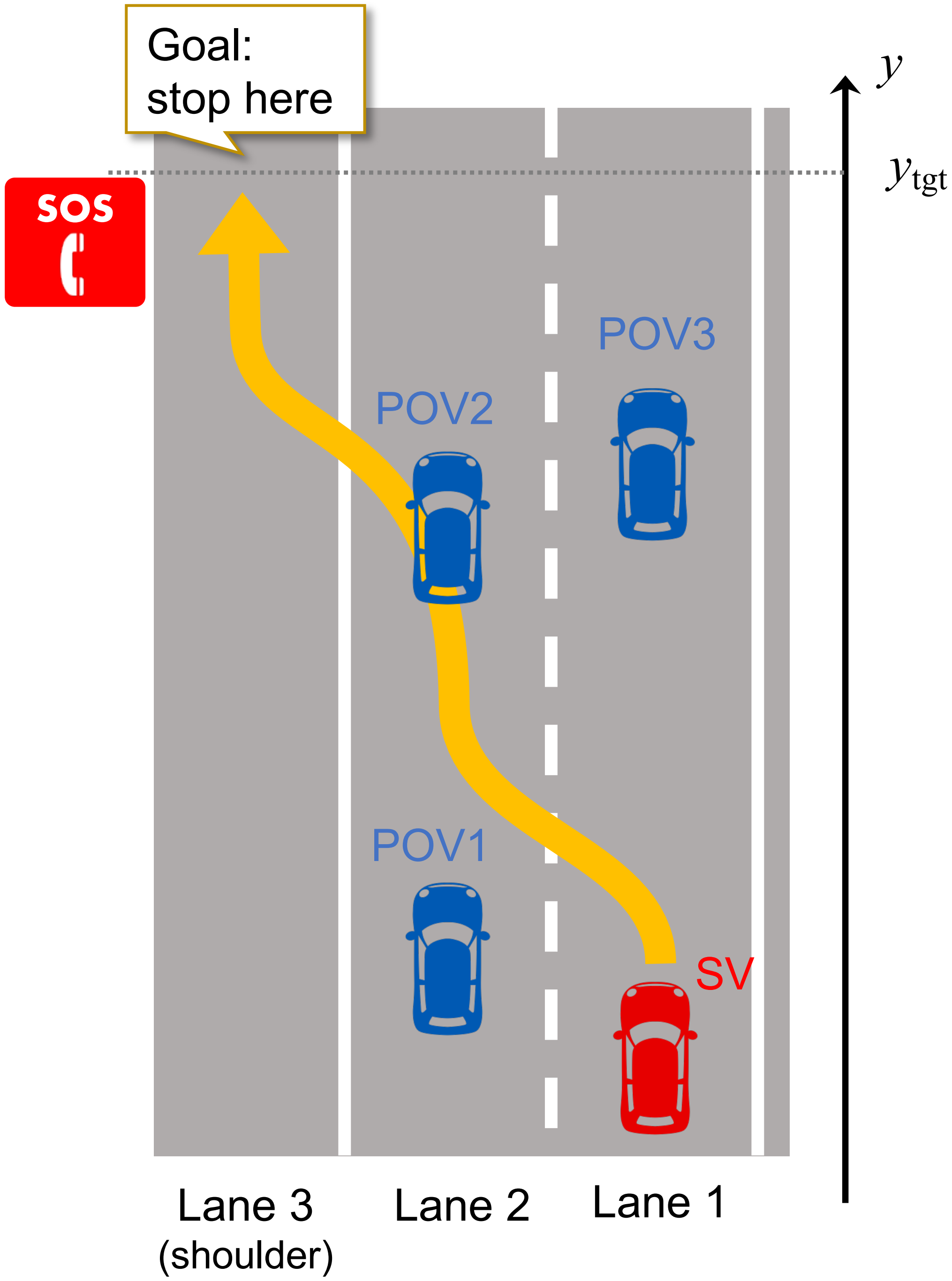

- An RSS rule is a pair (*A*, *α*) of an *RSS condition* *A* and a *proper response* *α*

RSS condition *A*: ("You can still escape if *A* is true")
Maintain an inter-vehicle distance at least

$$d_{\min} = \left[ v_r\,\rho + \frac{1}{2} a_{\max,\text{accel}}\,\rho^2 + \frac{(v_r + \rho\, a_{\max,\text{accel}})^2}{2a_{\min,\text{brake}}} - \frac{v_f^2}{2a_{\max,\text{brake}}} \right]_+$$

Proper response *α*: ("When you escape, use the control strategy α")
Brake at rate $a_{\text{min, brake}}$ within ρ seconds

Conditional safety lemma:
Any execution of *α*, from a state that satisfies *A*, is collision-free.




- Now what about this pull over scenario?
- Essential for eyes-off ADVs to hand the control over to human drivers
- Requires complex decision making
  - Merge before POV1, or after?
  - Accelerate to pass POV1…

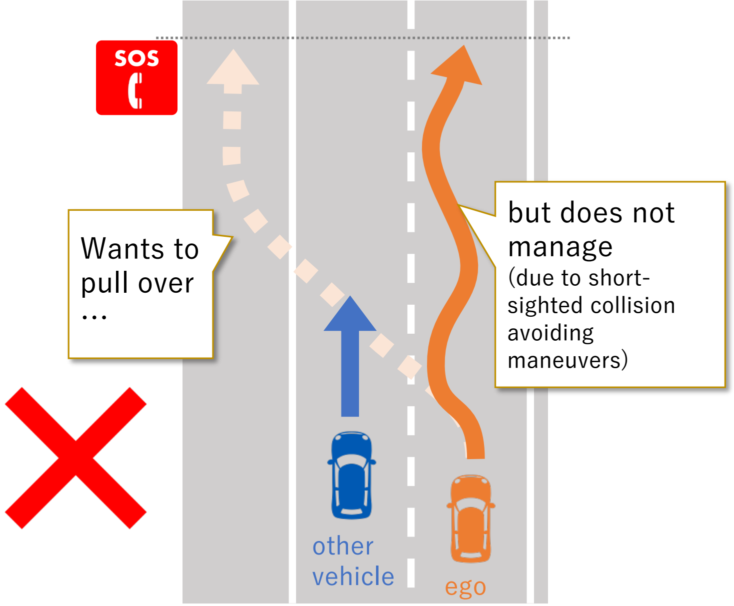
    ➔ Risk of overrun?

# Our Contribution: Logical Formalization of RSS → More Scenarios

## RSS

Responsibility-Sensitive Safety, Shalev-Shwartz et al., 2017

- Basic methodology of logical safety rules
- Standardization (IEEE 2846)
- Lack of formal implemention → <u>appl. to complex scenarios is hard</u>
- Guarantees only collision-freedom so far



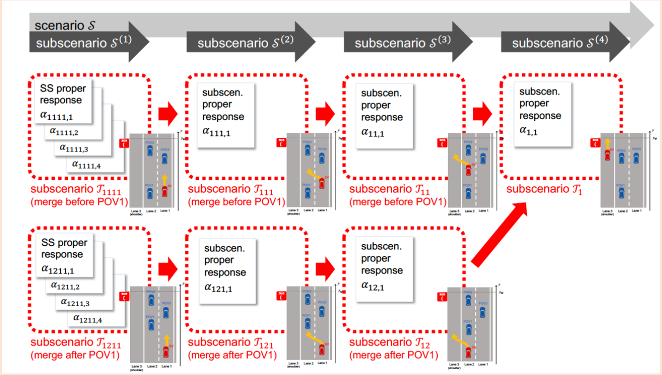

↓ Software science research

## differential program logic dFHL (our contribution)

$$\frac{\begin{array}{lll} \mathsf{inv}: & A \Rightarrow e_{\mathsf{inv}} \sim 0 & e_{\mathsf{var}} \geq 0 \land e_{\mathsf{inv}} \sim 0 \Rightarrow \mathcal{L}_{\dot{\mathbf{x}}=\mathbf{f}}\, e_{\mathsf{inv}} \simeq 0 \\ \mathsf{var}: & A \Rightarrow e_{\mathsf{var}} \geq 0 & e_{\mathsf{var}} \geq 0 \land e_{\mathsf{inv}} \sim 0 \Rightarrow \mathcal{L}_{\dot{\mathbf{x}}=\mathbf{f}}\, e_{\mathsf{var}} \leq e_{\mathsf{ter}} \\ \mathsf{ter}: & A \Rightarrow e_{\mathsf{ter}} < 0 & e_{\mathsf{var}} \geq 0 \land e_{\mathsf{inv}} \sim 0 \Rightarrow \mathcal{L}_{\dot{\mathbf{x}}=\mathbf{f}}\, e_{\mathsf{ter}} \leq 0 \end{array}}{\{A\}\ \mathsf{dwhile}\,(e_{\mathsf{var}} > 0)\,\dot{\mathbf{x}} = \mathbf{f}\ \{e_{\mathsf{var}} = 0 \land e_{\mathsf{inv}} \sim 0\} : e_{\mathsf{inv}} \sim 0 \land e_{\mathsf{var}} \geq 0}\ (\mathrm{DW_H})$$

- A logical system for deriving and proving safety rules

## Compositional rule derivation workflow by dFHL (our contribution)



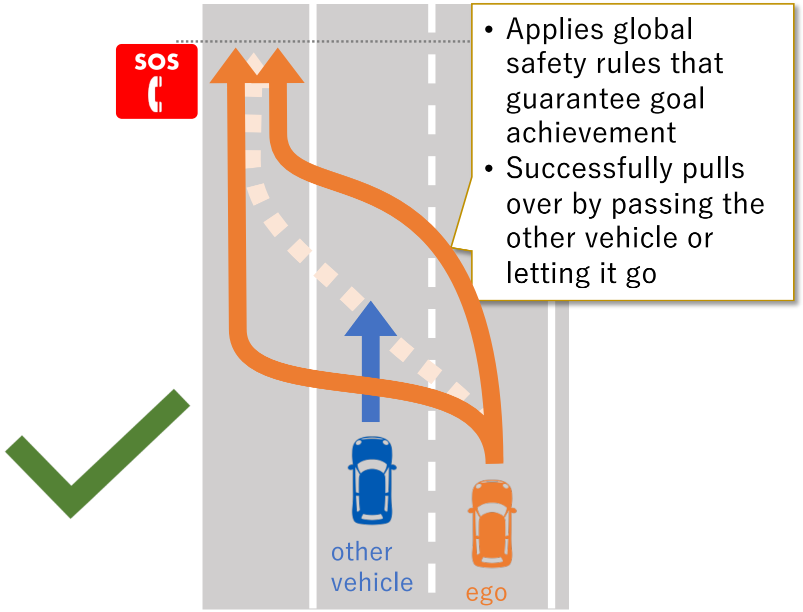

- "Divide and Conquer" complex driving scenarios
- Tool support by autom. reasoning

## GA-RSS (our contribution)

**Goal-Aware**

Responsibility-Sensitive Safety [Hasuo+, IEEE T-IV, 2023]

- Guarantees <u>**goal achievement**</u> (e.g. successful pull over) and collision-freedom
- Global safety rules that combine mult. maneuvers
- Necessary for real-world complex driving scenarios

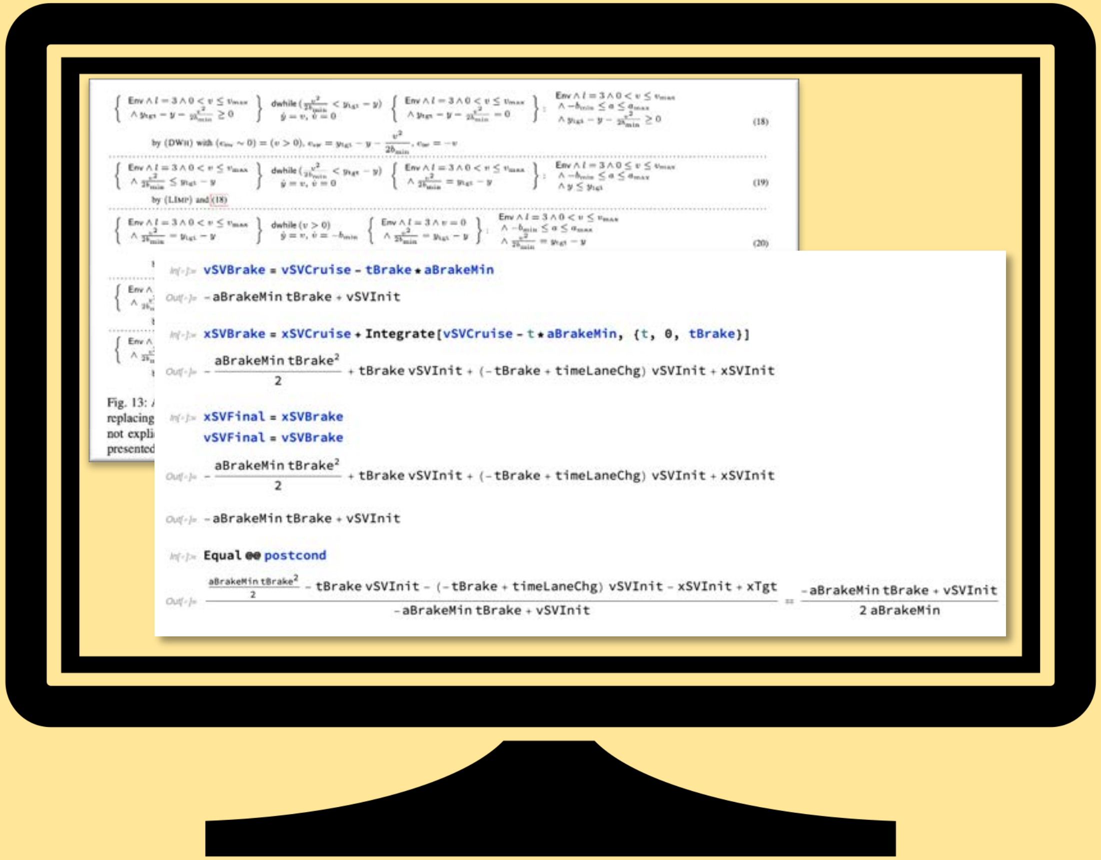


# What is Formalization?

## **Informal**
## pen-and-paper proofs

- Error-prone
- Poor traceability

## **Formal**
## software-assisted proofs




- Symbolic proofs in our formal logical system dFHL
- Software tool checking the validity of each logical step of reasoning

# Outline

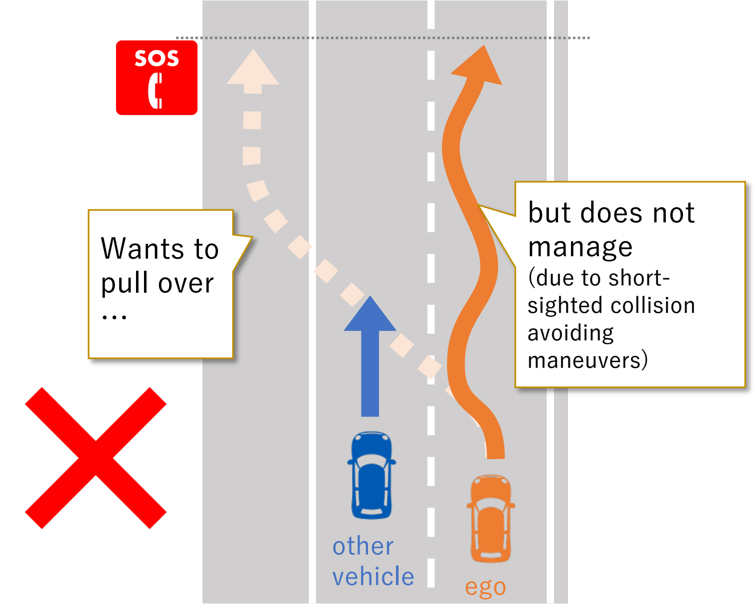
- A non-technical overview
- Technical contributions: the logic dFHL
- Perspectives, practical & theoretical

# Our Contribution: Formal Logic Foundations of RSS → More Scenarios

## RSS

Responsibility-Sensitive Safety, Shalev-Shwartz et al., 2017

- Basic methodology of logical safety rules
- Standardization (IEEE 2846)
- Lack of formal implemantion → **appl. to complex scenarios is hard**
- Guarantees only collision-freedom so far



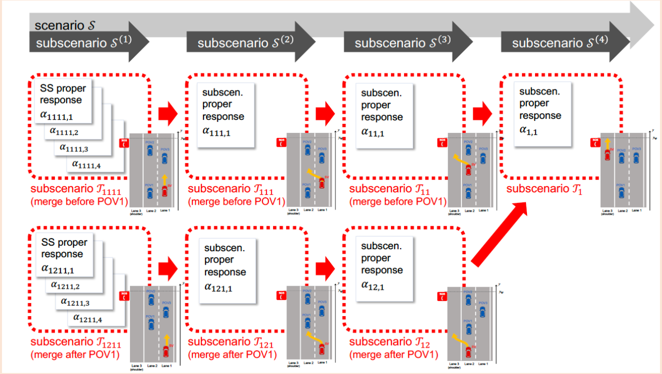

↓ Software science research

## differential program logic dFHL (our contribution)

$$\frac{\begin{array}{lll} \text{inv}: & A \Rightarrow e_{\mathrm{inv}} \sim 0 & e_{\mathrm{var}} \geq 0 \wedge e_{\mathrm{inv}} \sim 0 \Rightarrow \mathcal{L}_{\dot{x}=f}\, e_{\mathrm{inv}} \simeq 0 \\ \text{var}: & A \Rightarrow e_{\mathrm{var}} \geq 0 & e_{\mathrm{var}} \geq 0 \wedge e_{\mathrm{inv}} \sim 0 \Rightarrow \mathcal{L}_{\dot{x}=f}\, e_{\mathrm{var}} \leq e_{\mathrm{ter}} \\ \text{ter}: & A \Rightarrow e_{\mathrm{ter}} < 0 & e_{\mathrm{var}} \geq 0 \wedge e_{\mathrm{inv}} \sim 0 \Rightarrow \mathcal{L}_{\dot{x}=f}\, e_{\mathrm{ter}} \leq 0 \end{array}}{\{A\}\ \mathrm{dwhile}\,(e_{\mathrm{var}} > 0)\, \dot{x} = f\ \{e_{\mathrm{var}} = 0 \wedge e_{\mathrm{inv}} \sim 0\} : e_{\mathrm{inv}} \sim 0 \wedge e_{\mathrm{var}} \geq 0}\ (\mathrm{DWH})$$

- A logical system for deriving and proving safety rules

## Compositional rule derivation workflow by dFHL (our contribution)

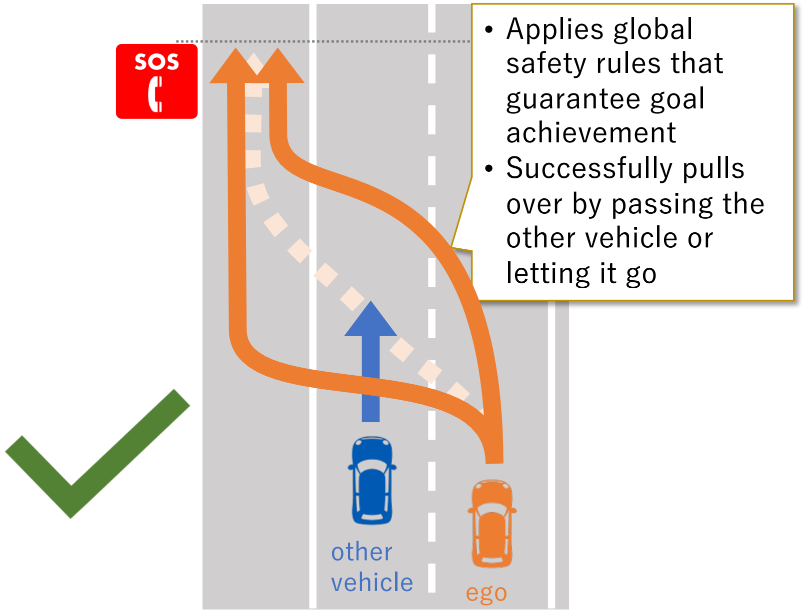



- "Divide and Conquer" complex driving scenarios
- Tool support by autom. reasoning

## GA-RSS (our contribution)

**Goal-Aware**
Responsibility-Sensitive Safety

- Guarantees **goal achievement** (e.g. successful pull over) and collision-freedom
- Global safety rules that combine mult. maneuvers
- Necessary for real-world complex driving scenarios



# Differential Program Logic dFHL

MathTrust

- Hoare logic (Tony Hoare, Turing Award 1980)
  + ODEs (dwhile)
  + "safety condition"

$$\{A\}\ \alpha\ \{B\} : S$$

postcondition ↑ (true at the end of α)

"safety condition" ↑ (true throughout α)

- Reasoning about ODEs via differential invariants (barrier cert.) and ranking/Lyapunov functions
- Theoretically not so much different from Platzer's dL. Simplified, aiding proof engineers

**Def.** (dFHL programs)

$$\alpha, \beta \quad ::= \quad \mathsf{skip} \mid \alpha;\beta \mid x := e \mid \mathsf{if}\,(A)\,\alpha\,\mathsf{else}\,\beta \mid \mathsf{while}\,(A)\,\alpha \mid \mathsf{dwhile}\,(A)\,\{\,\dot{\mathbf{x}} = \mathbf{f}\,\}\,.$$

**Def.** (dFHL rules) ⋮

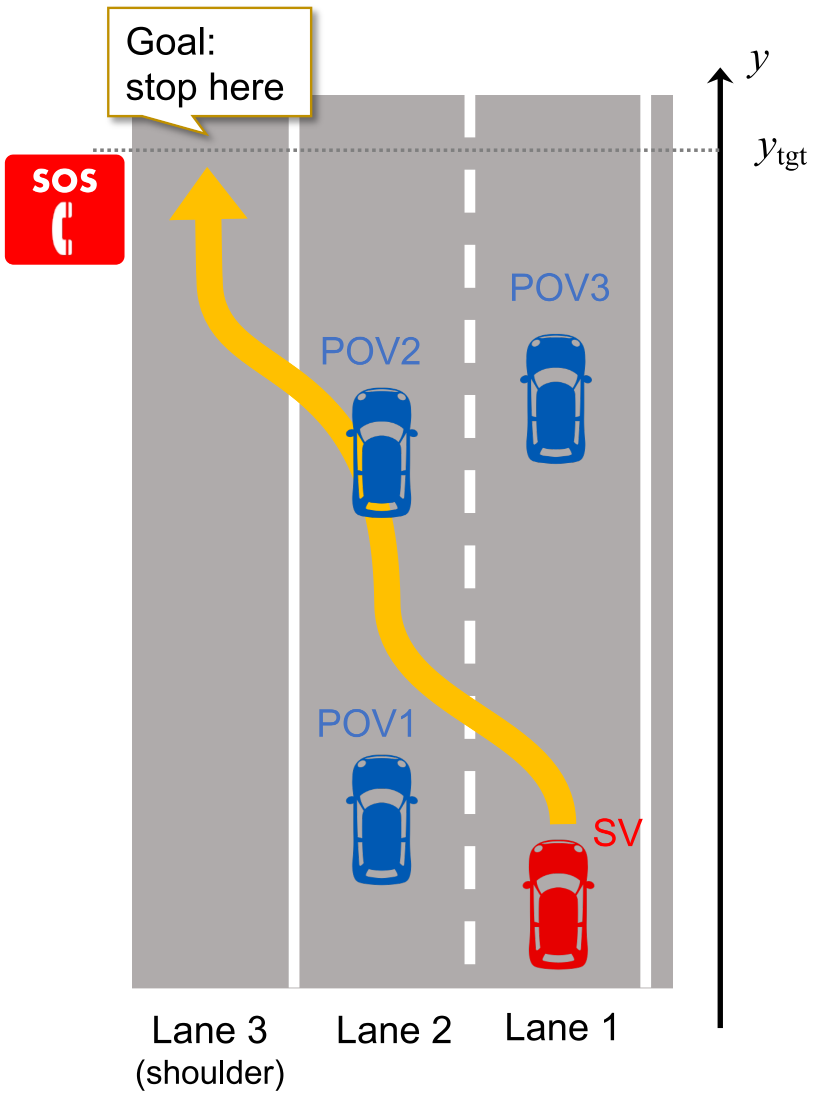
$$\frac{\{A\}\ \alpha\ \{B\} : S \qquad \{B\}\ \beta\ \{C\} : S}{\{A\}\ \alpha;\beta\ \{C\} : S}\ (\textsc{Seq})$$

$$\frac{\{A'\}\ \alpha\ \{B'\} : S' \qquad A \Rightarrow A' \quad S' \wedge B' \Rightarrow B \quad S' \Rightarrow S}{\{A\}\ \alpha\ \{B\} : S}\ (\textsc{LImp})$$

$$\frac{\begin{array}{lll}\mathsf{inv}: & A \Rightarrow e_{\mathsf{inv}} \sim 0 & e_{\mathsf{var}} \geq 0 \wedge e_{\mathsf{inv}} \sim 0 \Rightarrow \mathcal{L}_{\dot{\mathbf{x}}=\mathbf{f}}\, e_{\mathsf{inv}} \simeq 0 \\ \mathsf{var}: & A \Rightarrow e_{\mathsf{var}} \geq 0 & e_{\mathsf{var}} \geq 0 \wedge e_{\mathsf{inv}} \sim 0 \Rightarrow \mathcal{L}_{\dot{\mathbf{x}}=\mathbf{f}}\, e_{\mathsf{var}} \leq e_{\mathsf{ter}} \\ \mathsf{ter}: & A \Rightarrow e_{\mathsf{ter}} < 0 & e_{\mathsf{var}} \geq 0 \wedge e_{\mathsf{inv}} \sim 0 \Rightarrow \mathcal{L}_{\dot{\mathbf{x}}=\mathbf{f}}\, e_{\mathsf{ter}} \leq 0\end{array}}{\{A\}\ \mathsf{dwhile}\,(e_{\mathsf{var}} > 0)\,\dot{\mathbf{x}} = \mathbf{f}\ \{e_{\mathsf{var}} = 0 \wedge e_{\mathsf{inv}} \sim 0\} : e_{\mathsf{inv}} \sim 0 \wedge e_{\mathsf{var}} \geq 0}\ (\textsc{DWh})^{\dagger}$$

⋮

# Compositional Rule Derivation




- We shall derive

$$\{A\} \ \alpha \ \{B\} : S$$

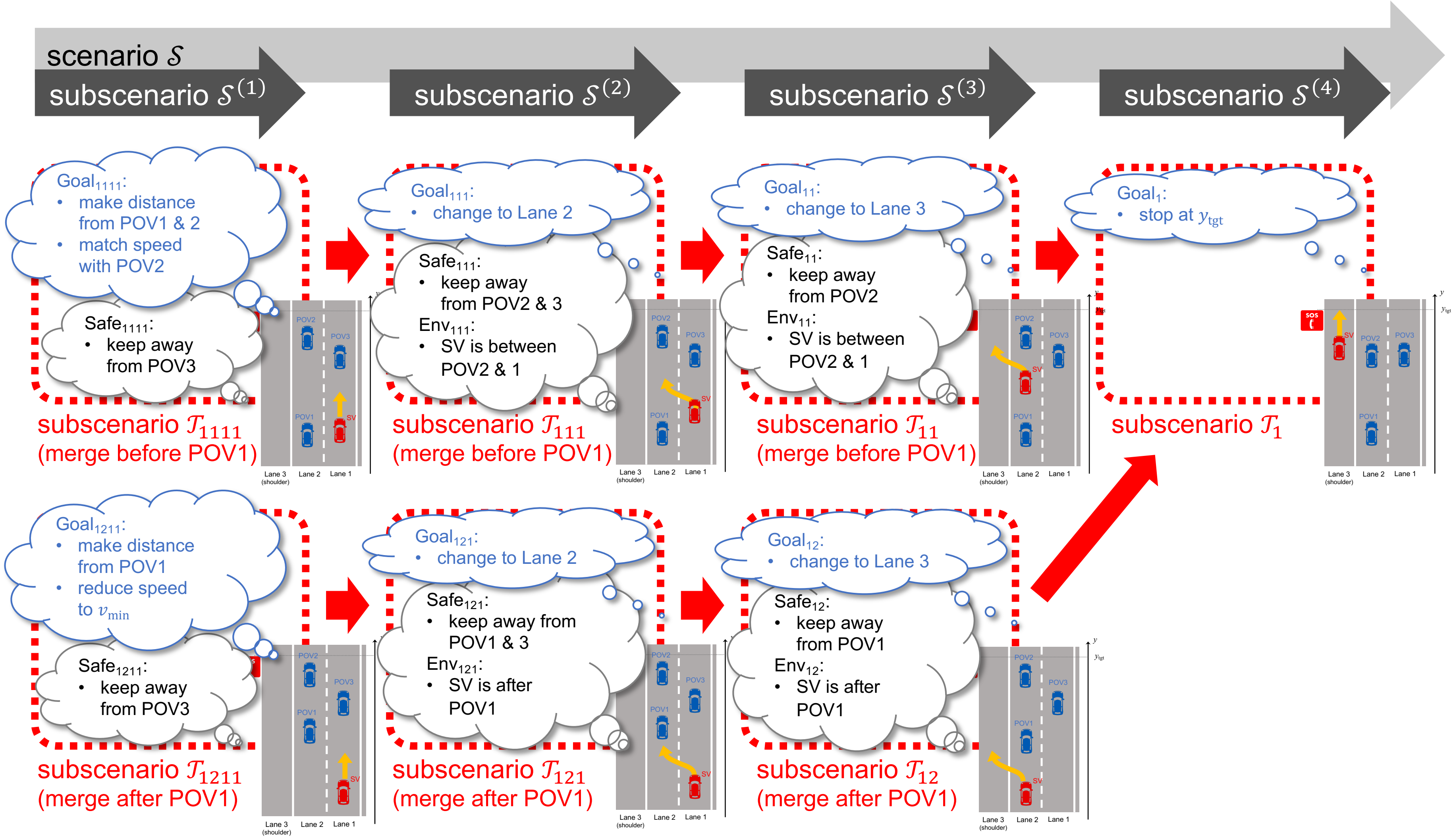
for the following given data

- $\boldsymbol{B}$ is the **goal**: "stoping on the shoulder at $y_{\mathrm{tgt}}$"
- $\boldsymbol{S}$ is the **safety**: "no collision," or better "securing RSS distance from every other car"

- We shall identify
  - α as an **RSS proper response**: "executing α will safely achieve the goal"
  - A as an **RSS condition**: "when A is true, B and S are guaranteed by executing α"

# Compositional Rule Derivation

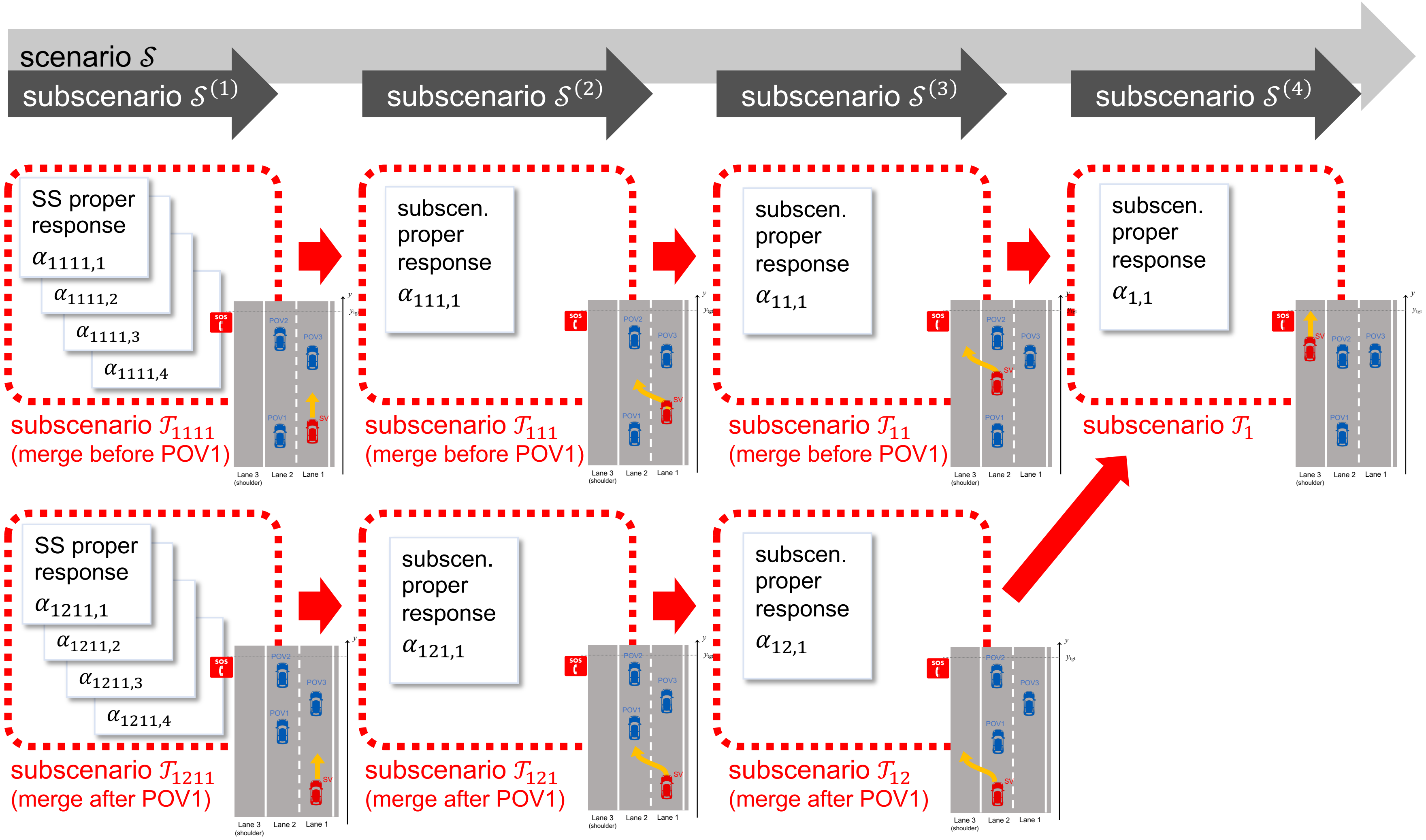
(1) Decompose the scenario into subscenarios,
each of which has clearer focuses and goals



# Compositional Rule Derivation

$$\{A\}\ \alpha\ \{B\} : S$$

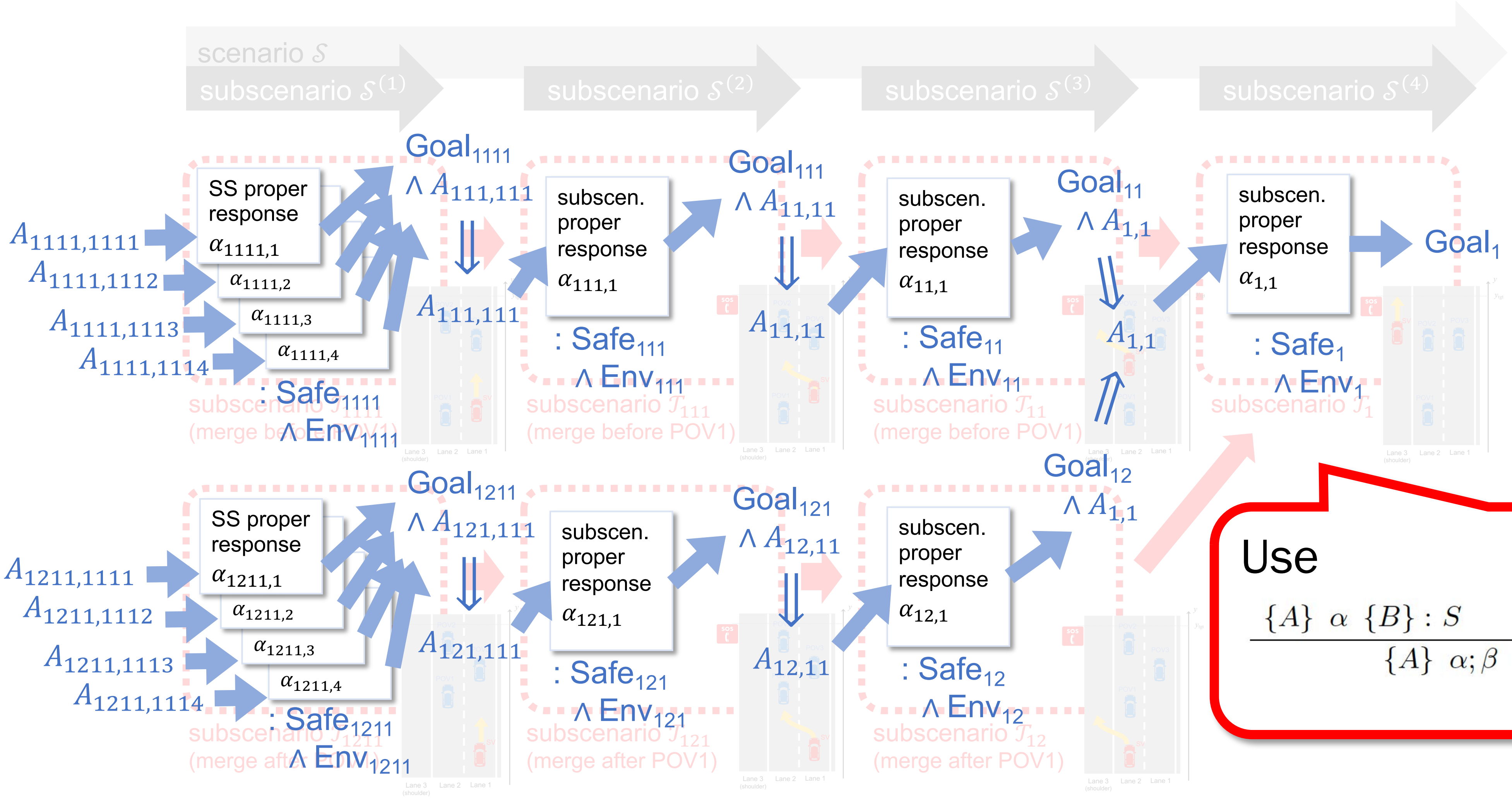
(2) Devise subscenario proper responses for each subscenario



# Compositional Rule Derivation

(3) Backpropagate pre/postconditions,
leading to the scenario-wide precondition

$$\{A\}\ \alpha\ \{B\} : S$$



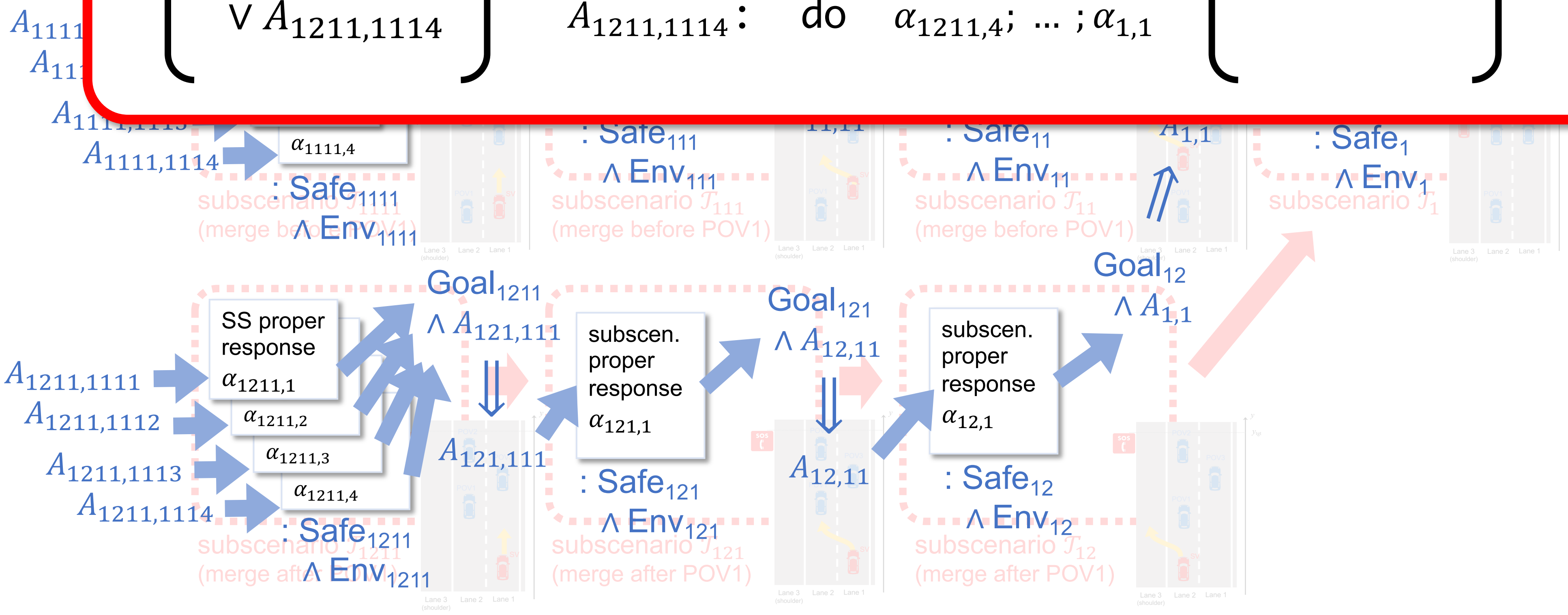

Use

$$\frac{\{A\}\ \alpha\ \{B\} : S \qquad \{B\}\ \beta\ \{C\} : S}{\{A\}\ \alpha;\beta\ \{C\} : S}\ (\textsc{Seq})$$

# Compositional Rule Derivation

(4) Derive a goal-achieving RSS rule

$$
\left\{
\begin{array}{l}
A_{1111,1111} \\
\vee A_{1111,1112} \\
\vee \ \dots \\
\vee A_{1211,1114}
\end{array}
\right\}
\begin{array}{llll}
\texttt{case} & & & \\
A_{1111,1111}: & \texttt{do} & \alpha_{1111,1}; \dots ; \alpha_{1,1} \\
A_{1111,1112}: & \texttt{do} & \alpha_{1111,2}; \dots ; \alpha_{1,1} \\
\quad \dots & & \\
A_{1211,1114}: & \texttt{do} & \alpha_{1211,4}; \dots ; \alpha_{1,1}
\end{array}
\left\{ \texttt{Goal}_1 \right\}
: \quad \texttt{Safe}
$$

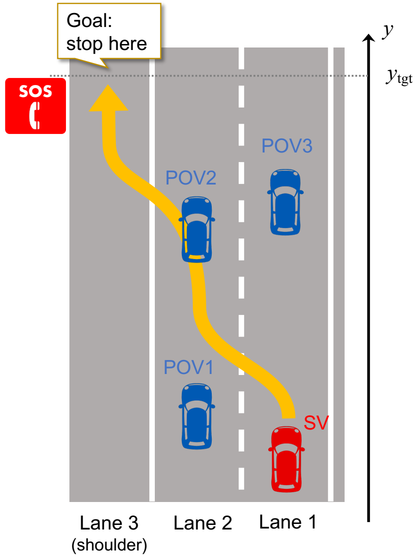


# Compositional Rule Derivation

MathTrust

## (4) Derive a goal-achieving RSS rule

$$\left\{\begin{array}{l} A_{1111,1111} \\ \vee\, A_{1111,1112} \\ \vee\ \ldots \\ \vee\, A_{1211,1114} \end{array}\right\} \begin{array}{llll} \texttt{case} & & & \\ A_{1111,1111}: & \texttt{do} & \alpha_{1111,1};\ \ldots\ ;\alpha_{1,1} \\ A_{1111,1112}: & \texttt{do} & \alpha_{1111,2};\ \ldots\ ;\alpha_{1,1} \\ \ \ \ldots & & \\ A_{1211,1114}: & \texttt{do} & \alpha_{1211,4};\ \ldots\ ;\alpha_{1,1} \end{array} \left\{ \texttt{Goal}_1 \right\} : \texttt{Safe}$$




**Goal-achieving RSS rule**

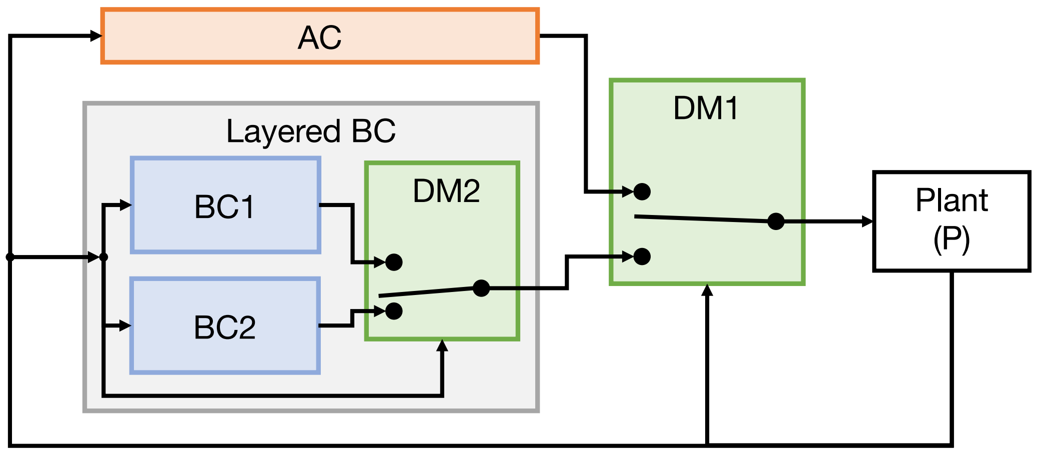
- RSS Condition: ("You can still escape if *A* is true")
  at least one of $A_{1111,1111}$, $A_{1111,1112}$, … , $A_{1211,1114}$ is true
- Proper response: ("When you escape, use this control strategy")

$$\begin{array}{llll} \texttt{case} & & & \\ A_{1111,1111}: & \texttt{do} & \alpha_{1111,1};\ \ldots\ ;\alpha_{1,1} & \leftarrow \text{accelerate and merge in front of POV1} \\ A_{1111,1112}: & \texttt{do} & \alpha_{1111,2};\ \ldots\ ;\alpha_{1,1} & \leftarrow \text{brake, cruise, and merge behind POV1} \\ \ \ \ldots & & & \ldots \\ A_{1211,1114}: & \texttt{do} & \alpha_{1211,4};\ \ldots\ ;\alpha_{1,1} & \end{array}$$

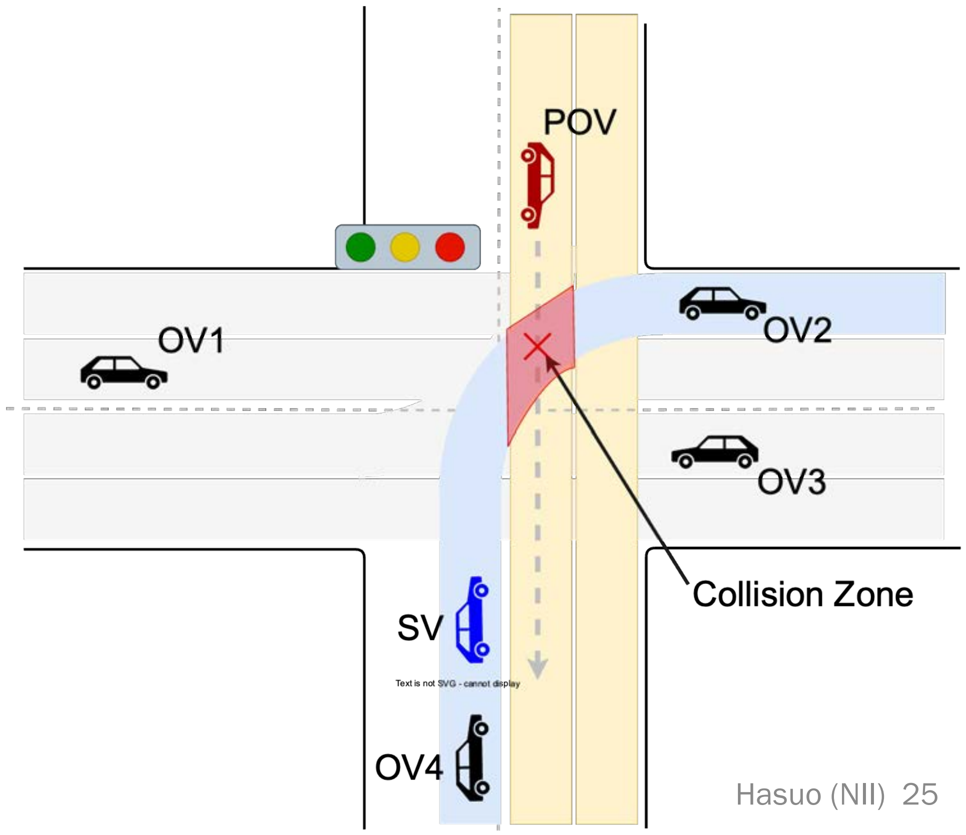
# Further Developments



- Extended logic (4-tuple ➜ 5-tuple) for **multi-layered safety rules** and **graceful degradation**
[Eberhart+, IV'23]




- Reasoning on control-flow graphs for **intersection scenarios**
[Haydon+, ITSC'23]

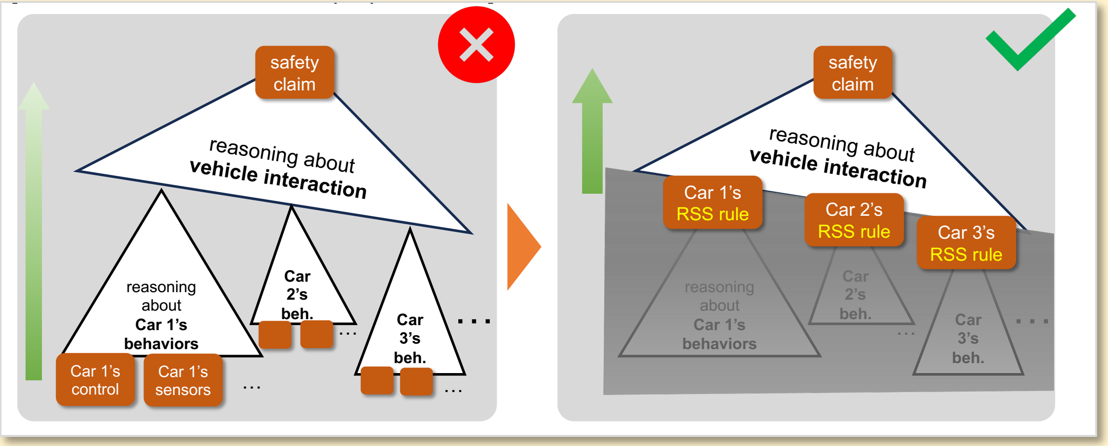


# Outline

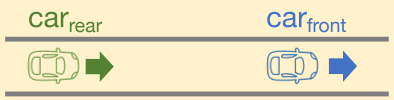
- A non-technical overview
- Technical contributions: the logic dFHL
- Perspectives, practical & theoretical

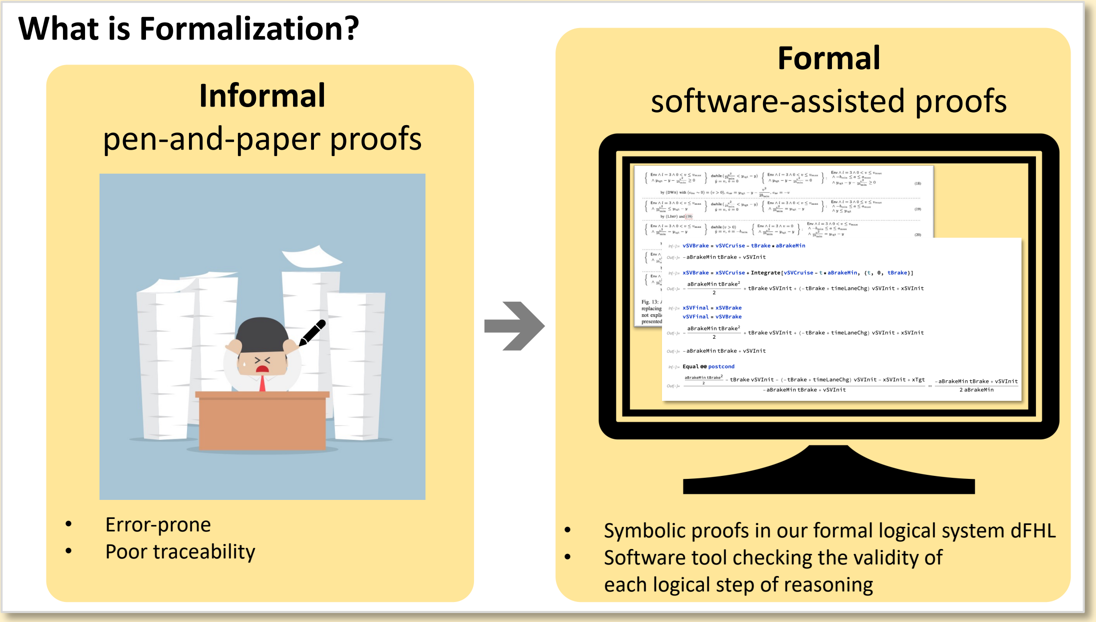
# What We Achieved [Hasuo+, IEEE T-IV 2023]

# Logical Formalization of RSS
# Covering More Scenarios ➜ Real-World Deployment

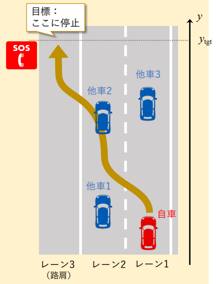



- RSS as in [Shalev-Shwartz et al., arXiv, 2017] is a **methodology**–
  it is sensible and promising,
  but came with no proof technologies
- thus application was limited to simple driving scenarios



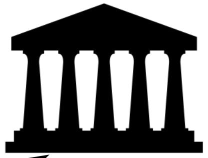




- Our contribution
  [Hasuo+, IEEE T-IV, to appear]:
  **Logical technologies** to prove
  *conditional safety lemmas* for complex scenarios
- Compositional proofs,
  ensuring goal achievements, …
- Much more scenarios proved safety by RSS
  ➜ RSS at work ➜ social acceptance of ADV



# RSS Rules as *Mathematical Traffic Laws*: Proof-Based Ecosystem for Safe Automated Driving

Regulation/Standardization Body



"I'm safe since I respect the safety rules $R_1$, $R_2$, ..."

"I'm safe since I respect the safety rules $R_1$, $R_2$, ..."

"I'm safe since I respect the safety rules $R_1$, $R_2$, ..."

- Decompose **safety** (a complex goal) into **logical safety rules** (explicit, easy to check and enforce)
- "Ultimate assurance" in the form of **mathematical proofs**. Logical explanation by following their reasoning steps
- Safety rules are generic and reusable
  ➜ regulation, standard ➜ social acceptance
- Attribution of liabilities
  (collision ➜ someone must have broken the rules)

$R_1$ $R_2$ $R_3$ ...

## Safety Rule $R_1$

In the *same-lane same-direction* driving scenario,

- Maintain the safety distance

$$d_{\min} = \left[ v_r\,\rho + \frac{1}{2} a_{\max,\mathrm{accel}}\,\rho^2 + \frac{(v_r + \rho\, a_{\max,\mathrm{accel}})^2}{2 a_{\min,\mathrm{brake}}} - \frac{v_f^2}{2 a_{\max,\mathrm{brake}}} \right]_+$$

  from the preceding car
- When that's hard, brake at acceleration $a_{\max,\mathrm{brake}}$

## Theorem (Safety)

There is no collision attributed to the ego vehicle as long as the safety rule $R_1$ is respected

## Proof (of the safety thm.)

⋮

The only non-obvious point is that $e_{\mathrm{inv},2}$ is preserved by the dynamics. We first observe

$$\mathcal{L}_{\delta_f,\delta_r^1}\, e_{\mathrm{inv},2} = \begin{cases} 0 & \text{if } \mathsf{dRSS}_\pm(v_f, v_r, \rho - t) \ge 0 \\ v_f - v_r & \text{otherwise,} \end{cases}$$

where $\mathsf{dRSS}_\pm(v_f, v_r, \rho)$ is given by

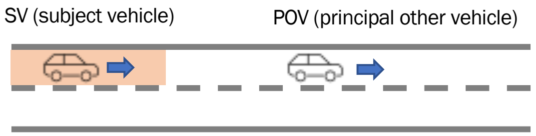
$$\mathsf{dRSS}_\pm(v_f, v_r, \rho) = v_r\rho + \frac{a_{\max}\rho^2}{2} + \frac{(v_r + a_{\max}\rho)^2}{2b_{\min}} - \frac{v_f^2}{2b_{\max}}.$$

Therefore, we can infer as follows.

$$\mathsf{dRSS}_\pm(v_f, v_r, \rho - t) < 0 \iff v_r(\rho - t) + \frac{a_{\max}(\rho - t)^2}{2} + \frac{(v_r + a_{\max}(\rho - t))^2}{2b_{\min}} - \frac{v_f^2}{2b_{\max}} < 0$$

⋮

# RSS Rules as *Mathematical Traffic Laws*: Proof-Based Ecosystem for Safe Automated Driving




Want an RSS rule C …
s.t. compliance with C
guarantees
collision-freedom

A **proof engineer** works on each driving scenario and derives an RSS rule C

- Correctness proof for C, at the same time

RSS rule C

+

Safety proof for C

Checkable by the public

…

proof-search is hard,
proof-check is easy

**Attribution of liabilities**
(collision
➔ ∃ a party who did not comply)

**Safety standards**
(Show compliance before selling cars)
e.g. IEEE P2846

} standardization bodies, gov't, …

**Safety verification**
(the car complies with these rules. Thus safe)

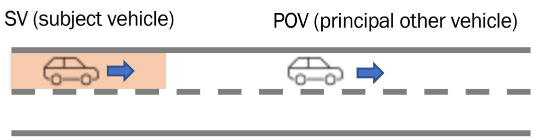
**Runtime monitoring**
(more freedom in ADS design)

} Car makers, ADS vendoes

**Risk analysis in insurance**
(Safety rule compliant ➔ lower risk
➔ cheaper insurance)

} insurance companies

# RSS Rules as *Mathematical Traffic Laws*: Proof-Based Ecosystem for Safe Automated Driving




Want an RSS rule C …
s.t. compliance with C
guarantees
collision-freedom

A **proof engineer** works on each driving scenario and derives an RSS rule C

- Correctness proof for C, at the same time

RSS rule C

+

Safety proof for C

Checkable by the public
…
proof-search is hard, proof-check is easy

**Attribution of liabilities**
(collision
➔ ∃ a party who did not comply)

**Safety standards**
(Show compliance before selling cars)
e.g. IEEE P2846

} standardization bodies, gov't, …

**Safety verification**
(the car complies with these rules. Thus safe)

**Runtime monitoring**
(more freedom in ADS design)

} Car makers, ADS vendoes

**Risk analysis in insurance**
(Safety rule compliant ➔ lower risk
➔ cheaper insurance)

} insurance companies

# Can Be Retrofit to Any ADV Controller
# Monitor & Intervene ➜ Runtime Safety Guarantee

## RSS Rule, an Example

[Shalev-Shwartz et al., arXiv preprint, 2017]


- An RSS rule is a pair ($A$, $\alpha$) of an *RSS condition* $A$ and a *proper response* $\alpha$

RSS condition $A$:
Maintain an inter-vehicle distance at least

$$d_{\min} = \left[ v_r\,\rho + \frac{1}{2} a_{\max,\mathrm{accel}}\,\rho^2 + \frac{(v_r + \rho\, a_{\max,\mathrm{accel}})^2}{2 a_{\min,\mathrm{brake}}} - \frac{v_f^2}{2 a_{\max,\mathrm{brake}}} \right]_+$$

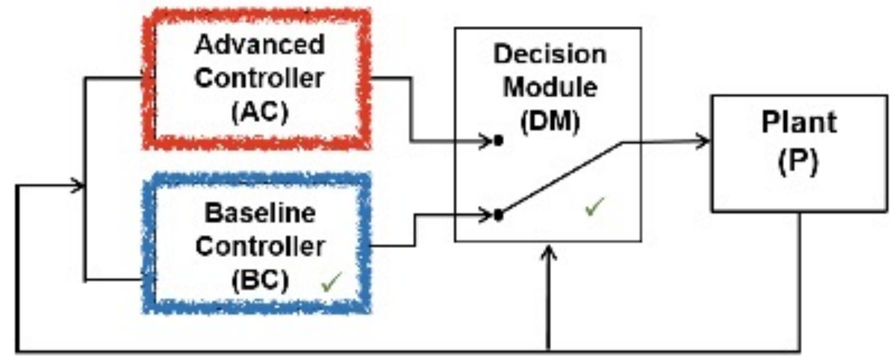
Proper response $\alpha$:
If A is about to be violated, brake at rate $a_{\text{min, brake}}$ within ρ seconds

Conditional safety lemma:
Any execution of $\alpha$, from a state that satisfies $A$, is collision-free.

### Structure of an RSS rule

*escape* =
MRM
(minimum risk maneuver)

- RSS Condition A:
  "You can still *escape* if A is true"
- Proper response α:
  "control strategy to *escape*"




Phan et al., ACSD'17

### Simplex architecture

- AC pursues performance and safety
- BC pursues safety (only)
- DM (decision module) switches between them—
  "use BC to escape"

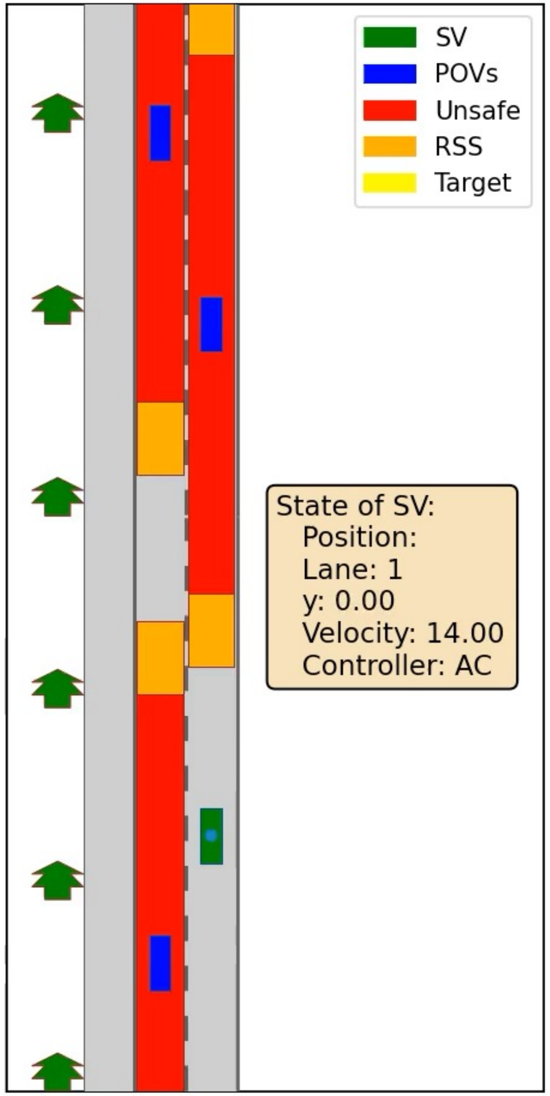
➜ RSS rules fit perfectly!

- AC: existing controller (optimization-based, ML, …)
- BC: executes a proper response
- DM: monitors an RSS condition.
  Violation foreseen ➜ switch to BC

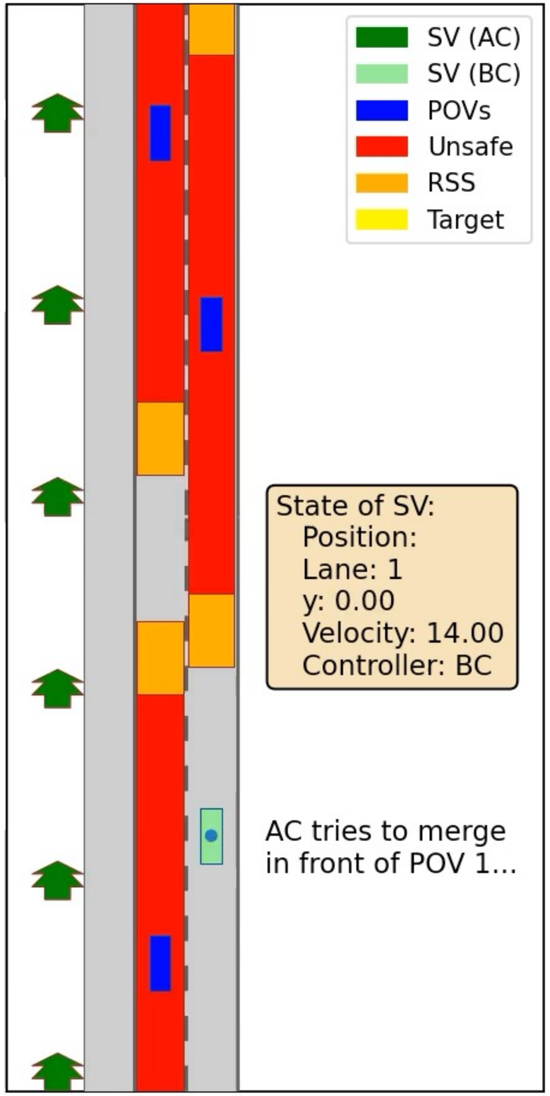
# RSS Safety Envelopes in Action, Scenario I

- AC: no safety envelope
- AC+RSS:
  Original RSS rule [Shalev-Shwartz et al., arXiv, 2017]
  as a safety envelope
  ("short-sighted" collision avoidance)
- AC+RSS$^{GA}$ :
  Our RSS rule [Hasuo+, IEEE T-IV]
  as a safety envelope
  (goal achievement too with longer-term planning)

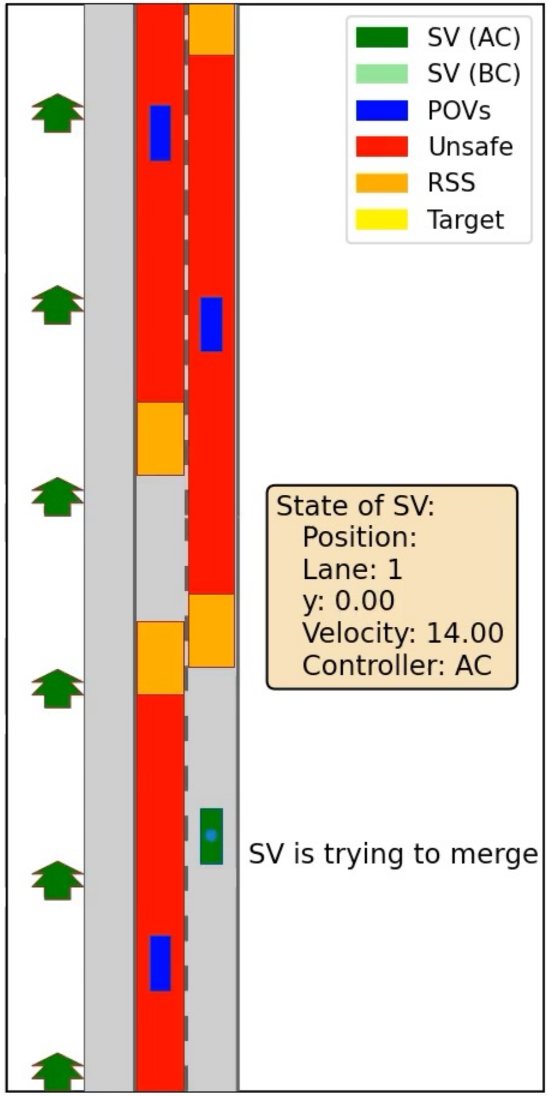
- AC is not safe (hazadous cut-in)
- AC+RSS does not reach the shoulder
- AC+RSS$^{GA}$ successfully deployed the long term strategy of (brake ➜ merge behind).
  Achieved both safety and the goal




AC



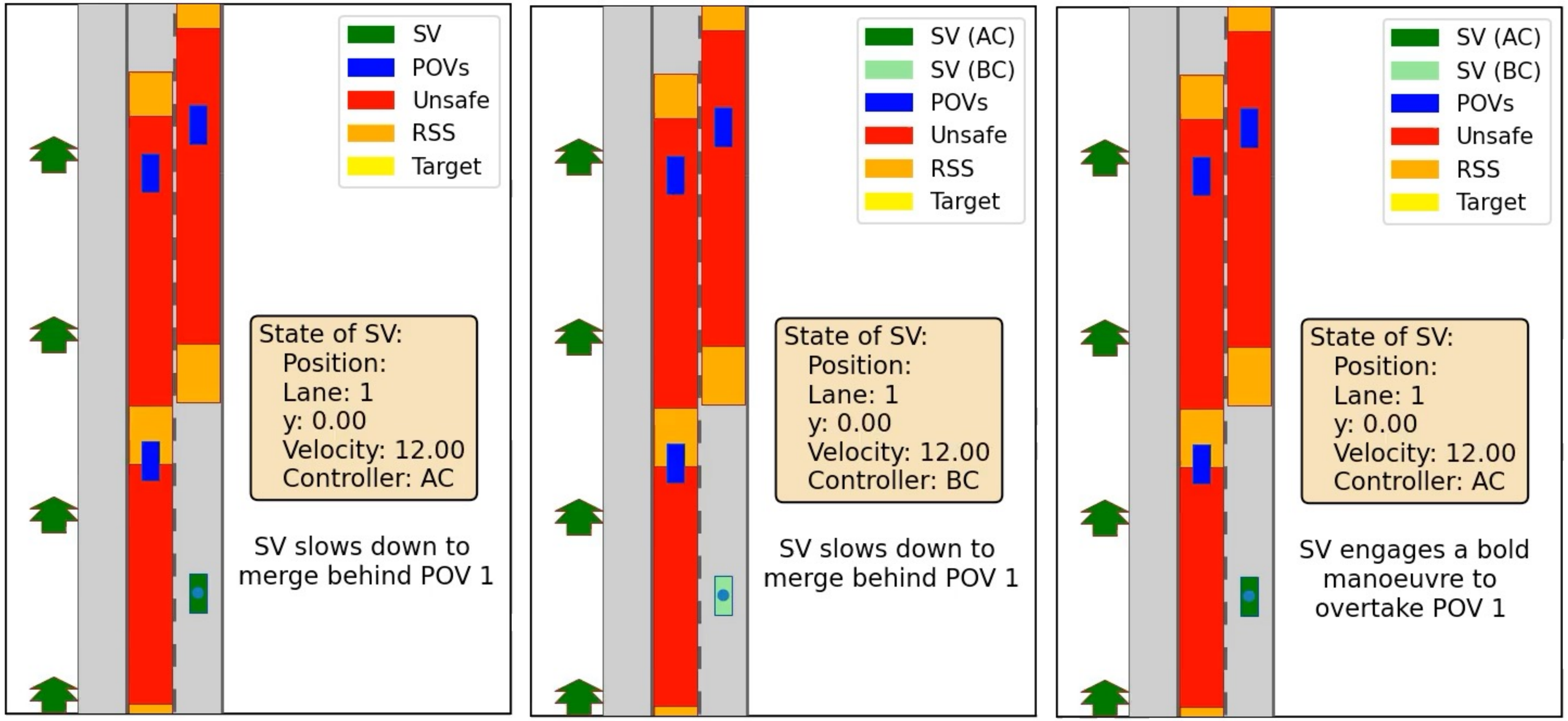

AC+RSS




AC+RSS$^{GA}$

# RSS Safety Envelopes in Action, Scenario II

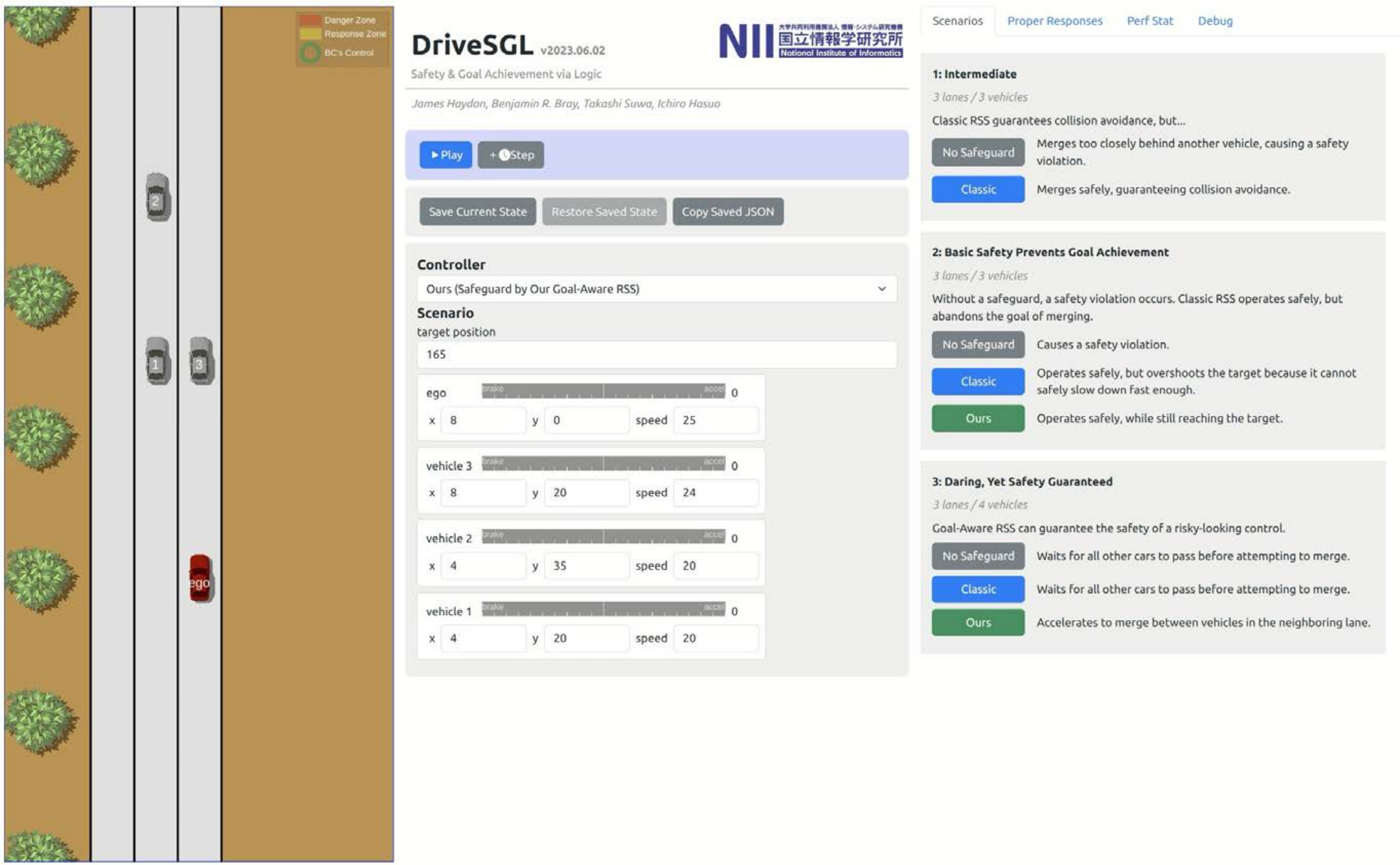
- AC: no safety envelope
- AC+RSS:
  Original RSS rule
  [Shalev-Shwartz et al., arXiv, 2017]
  as a safety envelope
  ("short-sighted" collision avoidance)
- AC+RSS$^{GA}$ :
  Our RSS rule [Hasuo+, IEEE T-IV]
  as a safety envelope
  (goal achievement too
  with longer-term planning)




AC AC+RSS AC+RSS$^{GA}$

- AC & AC+RSS safety achieve the goal, but are **slow**
- AC+RSS$^{GA}$, under mathematical safety guarantee, **boldly** accelerates and merge in front
  - … who says safe ADVs are conservative and boring? ☺

# DriveSGL – Our Live Demo (Under Devel.)



# Two Different Approaches, with Different Business Models

| <br>*Fixed-route* bus, taxi, delivery service | | <br>*Consumer* ADV |
|---|---|---|
| remote | human intervention | on-site (human driver) |
| offers fixed-route mobility and delivery service | business model | sells consumer vehicles with ADV functionality |
| yes<br>(the route is known) | geofencing | no<br>(should drive on all public roads) |
| full ODD<br>(automated driving in the entire route) | ODD<br>operational design domain<br>“Under which condition<br>can the ADV take responsibility?” | partial ODD<br>(automated driving only in prescribed situations, e.g. highway) |

# Two Different Approaches, with Different Business Models

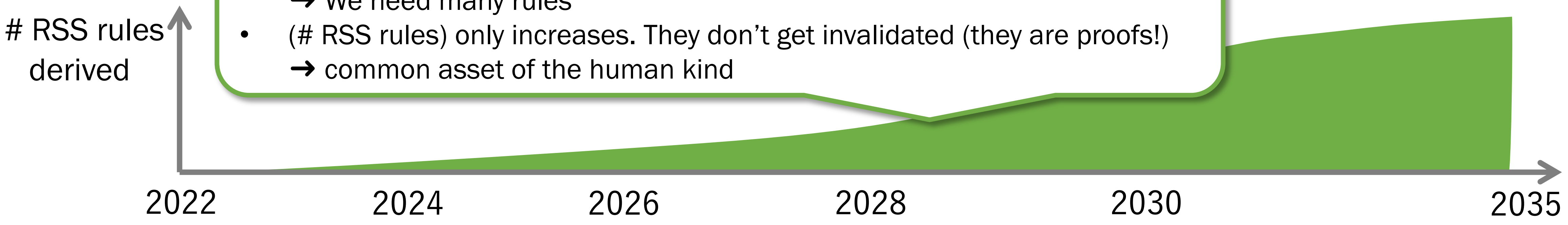
**Either way, to be responsible, we need to know driving scenarios in advance**

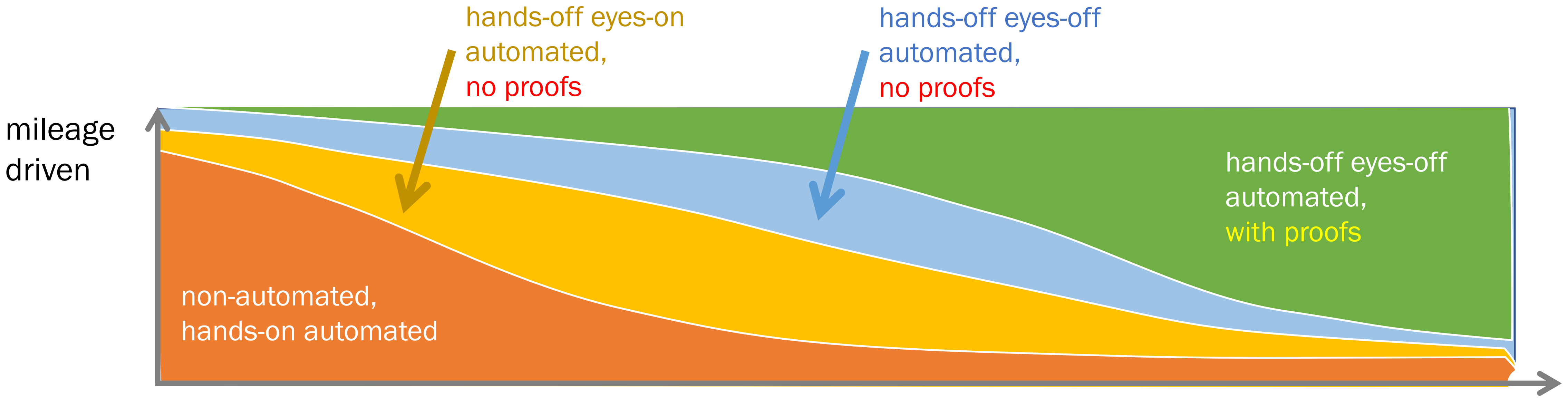
**→ We derive and verify RSS rules for those driving scenarios, and mathematically guarantee safety**

| *Fixed-route* bus, taxi, delivery | | *Consumer* ADV |
|---|---|---|
| remote | uman interventio | on-site (human driver) |
| offers fixed-route mobility and delivery service | business model | sells consumer vehicles with ADV functionality |
| yes (the route is known) | geofencing | no (should drive on all public roads) |
| full ODD (automated driving in the entire route) | ODD operational design domain "Under which condition can the ADV take responsibility?" | partial ODD (automated driving only in prescribed situations, e.g. highway) |

# Incremental Accumulation of RSS Rules, Incremental ODD Expansion of "ADVs with Proofs"

- ~ 1 person.month for each scenario
- Different rules for different traffic laws & customs
  ➜ We need many rules
- (# RSS rules) only increases. They don't get invalidated (they are proofs!)
  ➜ common asset of the human kind



# Two Possible Shapes of ADV Safety. Which is Better?

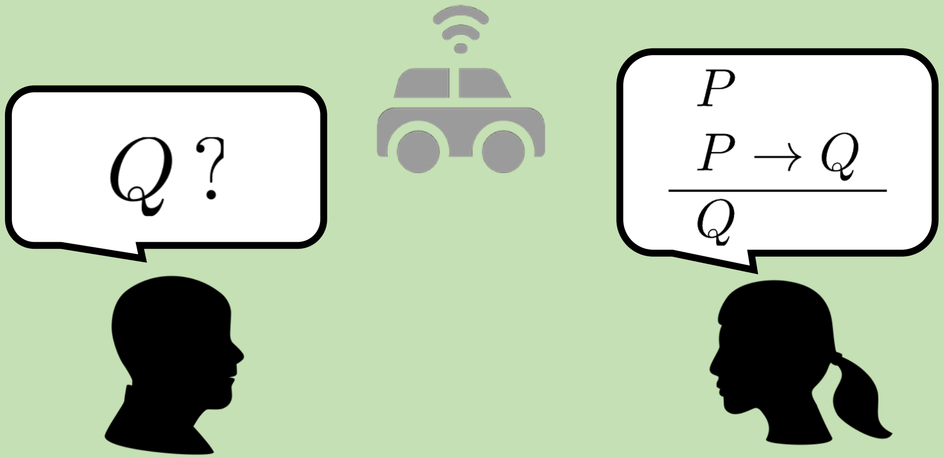
NII
MathTrust

## Blackbox Safety

- Monolithic “safety claims”
- Hard to examine, criticize, or improve

**vs**

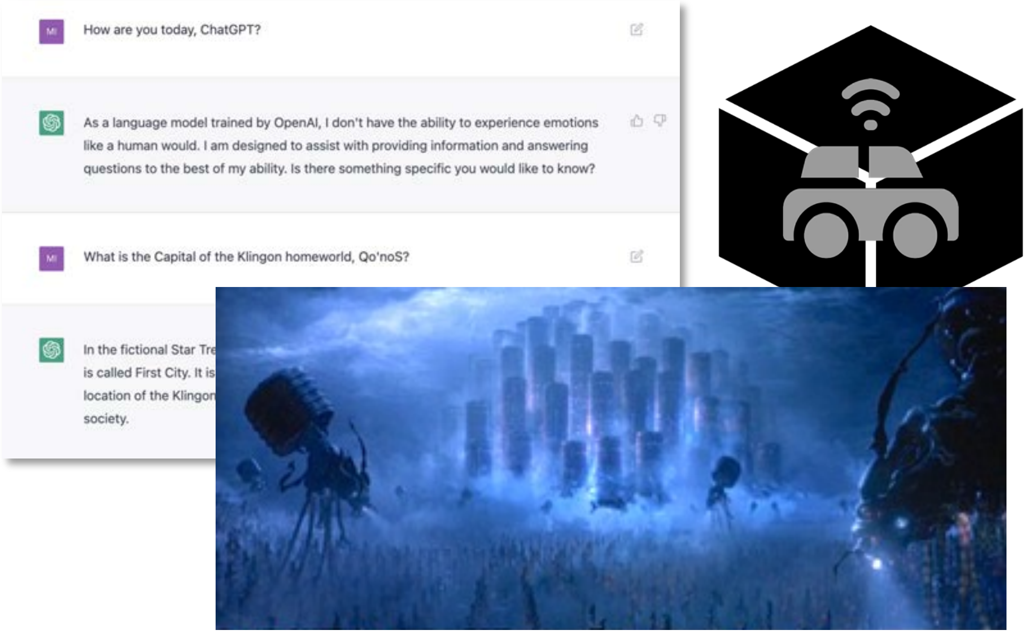
## Accountable Safety

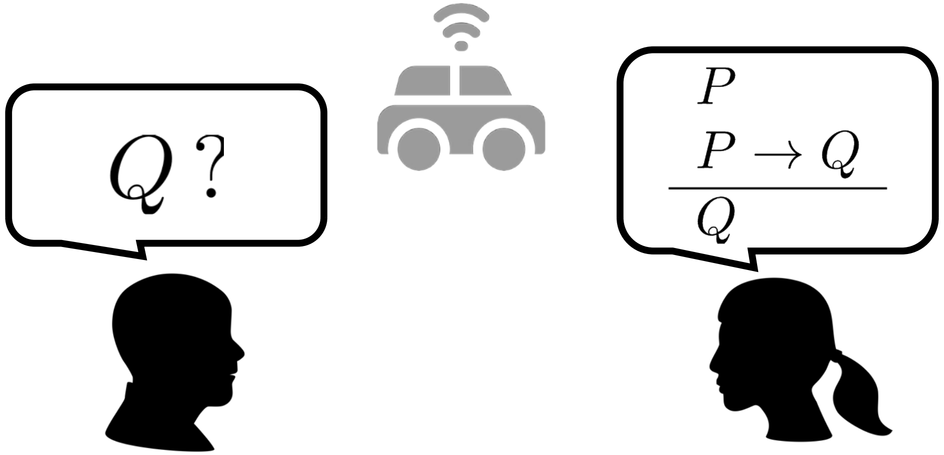



- Explainable and traceable safety cases structured by **logic**
- Supporting society’s collective and endless efforts towards ADV safety
- **The shape that we pursue**

# Logic's Mission in Society

## Safety-Critical Systems Should Never be Blackbox
## Proofs Explicate Assumptions, Contracts, ODDs, and Responsibilities




- Many emerging technologies are statistical and blackbox
- We shouldn't let them operate in safety-critical domains
- (... fight against the "lawyer up" approach towards safety!)

- <u>Conventionally</u>:
  Proofs are for establishing absolute truths
- <u>New</u>: proofs are **communication media** for
  - explicating assumptions and contracts,
  - showing who's responsible for what, and
  - writing and assessing safety cases
- Logiic as a social infrastructure for trust in ICT