Responsibility-Sensitive Safety for Automated Driving

Clovis Eberhart

Software Construction Laboratory, Research Institute of Electric Communications, Tohoku University

"AI-Physical Systems" ASPIRE Kick-off Meeting, March 3-4, 2026

Challenges in Automated Driving

Pros of automated driving

- can reduce travel time
- can reduce fuel consumption
- may become safer than human driving

Cons of automated driving

- little public trust
- no clear regulations
- no clear delimitation of liability
- **safety problem** (twice as many crashes per mile as human driving)

Safety of Automated Driving

Two types of methods for safety of automated driving.

Empirical methods

- statistical methods, simulations, testing...
- **great scalability**
- **works on black-box systems**
- **no formal guarantees**

Formal methods

- logical methods, automata-based methods, **semantic methods**
- **not scalable**
- **need formal definitions**
- **formal guarantee**

Responsibility-Sensitive Safety[1]

Problem with logical methods

Need formal definitions of systems, which are too complex.

Abstract away part of the systems to keep them simple and prove something about them.

$$\begin{cases} \dot{x} = v \\ \dot{v} = a \end{cases} \qquad a \in [-b_{\max}, a_{\max}]$$

Responsibility-Sensitive Safety

- RSS condition: condition under which the system is safe
- RSS response: how to maintain safety
- based on responsibility

[1] Shalev-Shwartz, Shammah, and Shashua, *On a formal model of safe and scalable self-driving cars*, arXiv, 2017.

Responsibility-Sensitive Safety

Driving situation:

Who is responsible in case of a crash?

Strategy (RSS response)

Brake enough to leave at least a distance of

$$\max\left(0, v_r\rho + \frac{a_{\max}\rho^2}{2} + \frac{(v_r + a_{\max})^2}{2b_{\min}} - \frac{v_f^2}{2b_{\max}}\right)$$

between the two vehicles.

Goal of the research[1]

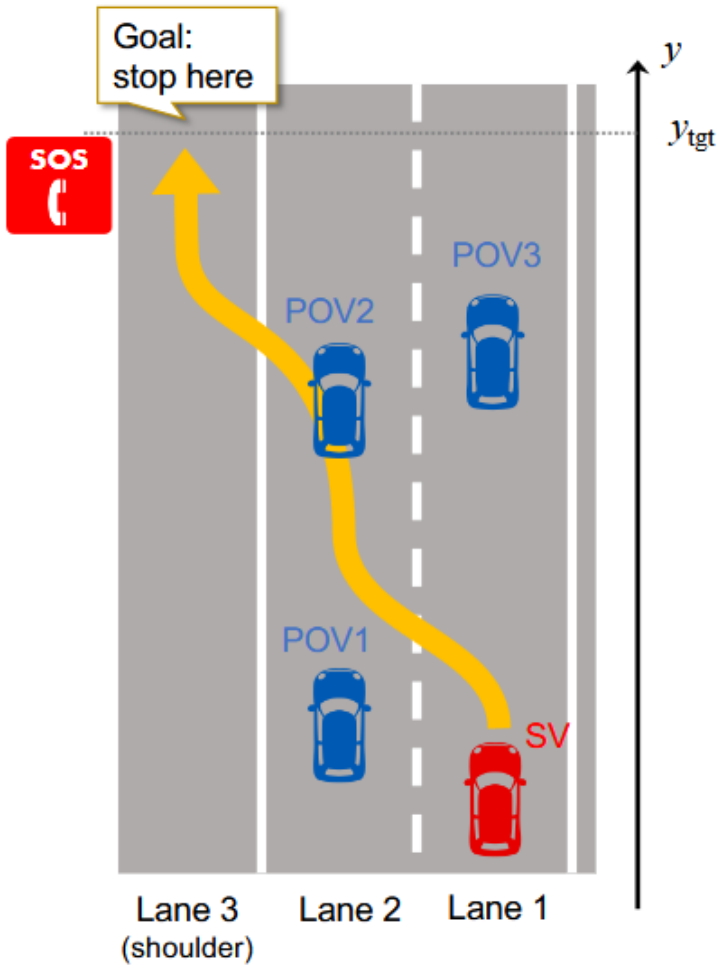
Limitation of RSS

Pen-and-paper proofs: prone to errors, do not scale.

Goal

Given a “strategy pattern”, get

- an RSS condition (a safety zone)
- an RSS response (a strategy to stay safe)



[1] Hasuo, Eberhart, Haydon, Dubut, Bohrer, Kobayashi, Pruekprasert, Zhang, Pallas, Yamada, Suenaga, Ishikawa, Kamijo, Shinya, and Suetomi. *Goal-Aware RSS for Complex Scenarios via Program Logic*, IEEE T-IV, 2023

Method

Ingredients

- **hybrid programs**: programs with continuous behaviours
- **Hoare quadruples**: properties of hybrid programs
- **dFHL**: logic to prove Hoare quadruples

Proving safety of driving situations

- Turn a “strategy pattern” into a hybrid program
- Turn strategy safety into a Hoare quadruple
- Prove the quadruple with dFHL
- Infer pattern to get a concrete strategy

Scenarios as hybrid programs

Hybrid programs:

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

Example of strategy pattern

$$\begin{array}{rcl} \alpha & = & l := 0; \mathsf{dwhile}\,(P_1)\,\{\,\dot{x} = v, \dot{v} = a_{\max}\,\}\,; \\ & & l := 0.5; \mathsf{dwhile}\,(P_2)\,\{\,\dot{x} = v, \dot{v} = 0\,\}\,; \\ & & l := 1; \ldots \end{array}$$

Safety of hybrid programs

Hoare quadruple

A Hoare quadruple $\{P\}\ \alpha\ \{Q\} : S$ is valid if for all stores $\rho \vDash P$:

- there exists ρ' s.t. $\langle \alpha, \rho \rangle$ converges to $\rho' \vDash Q$,
- for all $\langle \alpha, \rho \rangle \rightarrow^* \langle \alpha', \rho' \rangle$, $\rho' \vDash S$.

Meaning of Hoare quadruple:

- if P holds before α runs
- then Q holds after the run
- and S holds along the whole execution

For us:

- P: RSS condition (to be infered in our case)
- Q: goal condition (progress)
- S: safety (absence of collision)

The logic dFHL

$$\frac{}{\{A\}\ \mathsf{skip}\ \{A\} : A}\ (\textsc{Skip}) \qquad \frac{\{A\}\ \alpha\ \{B\} : S \qquad \{B\}\ \beta\ \{C\} : S}{\{A\}\ \alpha;\beta\ \{C\} : S}\ (\textsc{Seq}) \qquad \frac{}{\{A[e/x]\}\ x := e\ \{A\} : A \vee A[e/x]}\ (\textsc{Assign})$$

$$\frac{\{A \wedge B\}\ \alpha\ \{C\} : S \quad \{\neg A \wedge B\}\ \beta\ \{C\} : S}{\{B\}\ \mathsf{if}\,(A)\,\alpha\,\mathsf{else}\,\beta\ \{C\} : S}\ (\textsc{If}) \qquad \frac{\{A \wedge B \wedge e_{\mathsf{var}} \gtrsim 0 \wedge e_{\mathsf{var}} = x\}\ \alpha\ \{B \wedge e_{\mathsf{var}} \gtrsim 0 \wedge e_{\mathsf{var}} \leq x - 1\} : S}{\{B \wedge e_{\mathsf{var}} \gtrsim 0\}\ \mathsf{while}\,(A)\,\alpha\ \{\neg A \wedge B \wedge e_{\mathsf{var}} \gtrsim 0\} : S}\ (\textsc{Wh})^{\dagger}$$

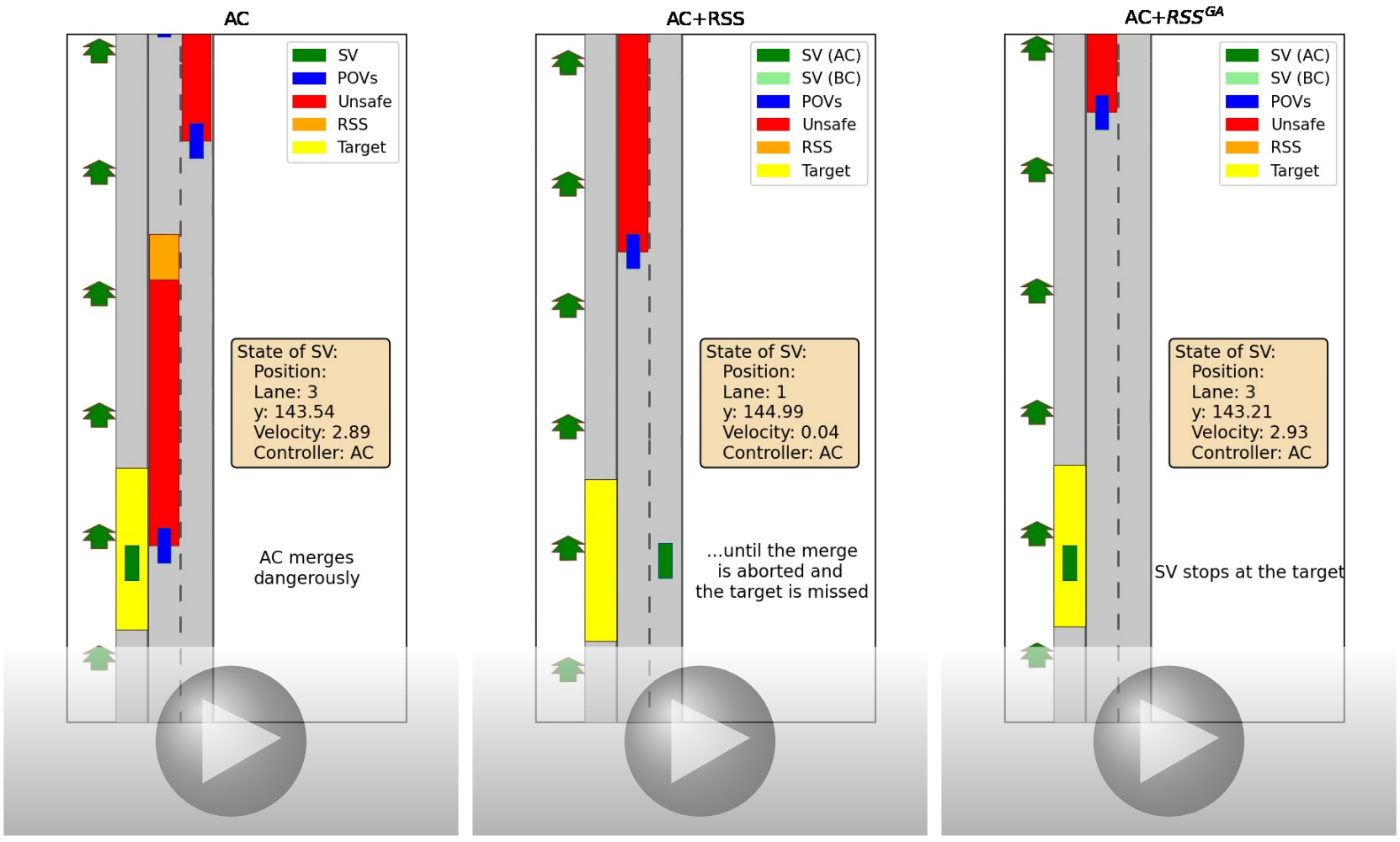
$$\frac{\begin{array}{ll}\mathsf{inv}: & A \Rightarrow e_{\mathsf{inv}} \sim 0 \quad e_{\mathsf{var}} \geq 0 \wedge e_{\mathsf{inv}} \sim 0 \Rightarrow \mathcal{L}_{\dot{\mathbf{x}}=\mathbf{f}}\, e_{\mathsf{inv}} \simeq 0 \\ \mathsf{var}: & A \Rightarrow e_{\mathsf{var}} \geq 0 \quad e_{\mathsf{var}}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

$$\frac{\{A\}\ \alpha\ \{B\} : S \qquad \{A\}\ \alpha\ \{B'\} : S'}{\{A\}\ \alpha\ \{B \wedge B'\} : S \wedge S'}\ (\textsc{Conj}) \qquad \frac{A_0 \Rightarrow (\exists t \geq 0.\ C_{<t} \wedge \neg C_t \wedge B_t \wedge S_{\leq t})}{\{A\}\ \mathsf{dwhile}\,(C)\,\dot{\mathbf{x}} = \mathbf{f}\ \{B\} : S}\ (\textsc{DWh-Sol})$$

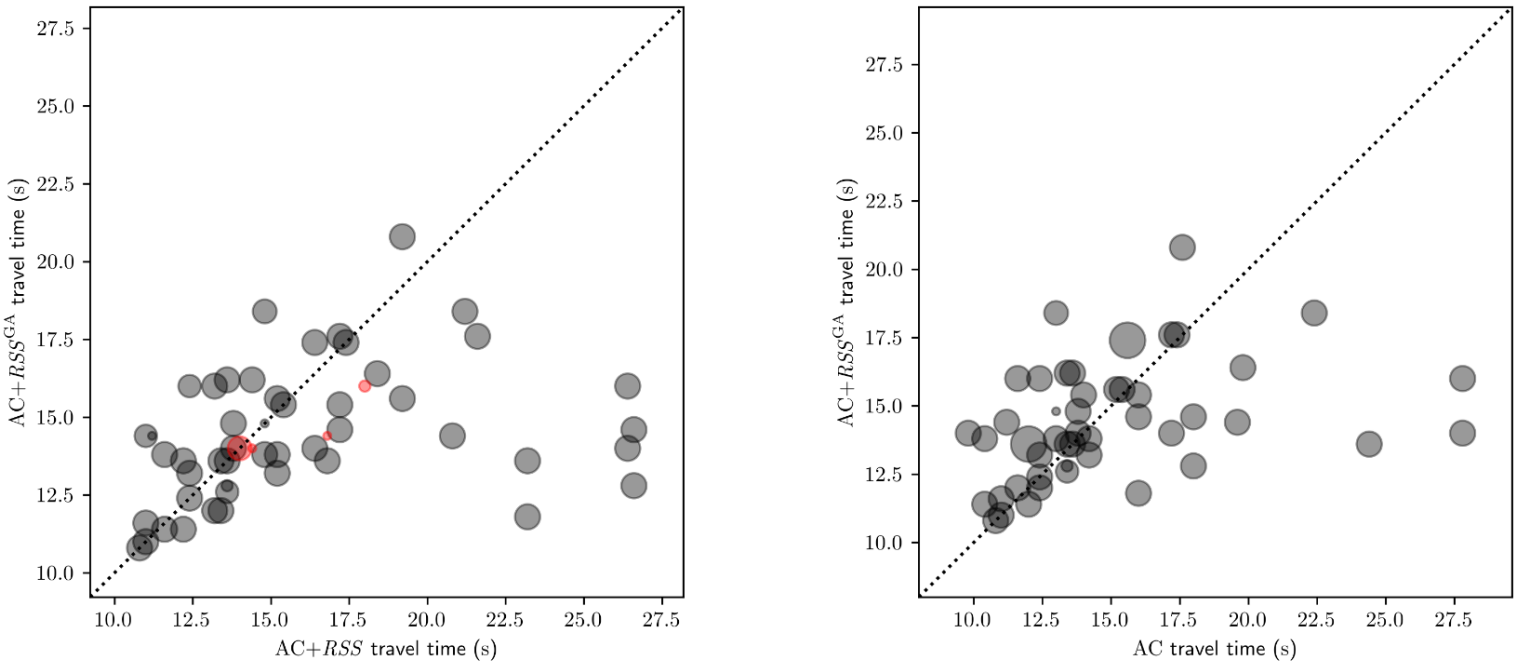
Theorem

Only valid Hoare quadruples can be proved in dHL.

Experimental results



Experimental evaluation



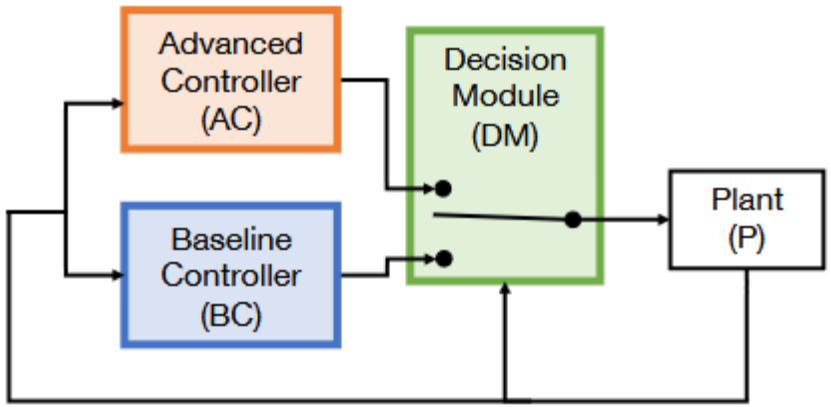
(a) AC+RSS^{CA} vs AC+RSS^{GA}.

(b) AC vs AC+RSS^{GA}.

Fig. 16: Comparison for progress. A red disc indicates that AC+RSS^{CA} failed to achieve the goal in this scenario instance.

Safety architectures[1]

Proving safety of “safety architectures”.



Compatible with RSS:

- DM: RSS precondition
- BC: RSS response
- AC: unknown (e.g., AI controller)

[1] Eberhart, Dubut, Haydon, and Hasuo. *Formal Verification of Safety Architectures for Automated Driving*, IV 2023

Proving safety of safety architectures

Hoare quintuples

$A : \{P\}\ \alpha\ \{Q\} : S$ where A is an assumption.

We can then prove:

- system safe and makes progress under some strong assumption (e.g., other vehicles at constant speed)
- system safe under a weaker assumption (e.g., other vehicles may change speed)

where AC is an AI controller.

Conclusion

Our approach to RSS:

- is based on semantics of programming languages
- gives **formal** guarantees of safety for automated driving systems
- can **scale** beyond pen-and-paper proofs (needs a proof engineer)
- can be used to prove **safety of AI controllers** (in an architecture)

Operational Semantics

$$\frac{}{\langle \mathsf{skip}; \beta, \rho\rangle \to \langle \beta, \rho\rangle} \qquad \frac{\langle \alpha, \rho\rangle \to \langle \alpha', \rho'\rangle}{\langle \alpha; \beta, \rho\rangle \to \langle \alpha'; \beta, \rho'\rangle} \qquad \frac{}{\langle x := e, \rho\rangle \to \langle \mathsf{skip}, \rho[x \to [\![e]\!]_\rho]\rangle} \qquad \frac{\rho \vDash A}{\langle \mathsf{if}\ (A)\ \alpha\ \mathsf{else}\ \beta, \rho\rangle \to \langle \alpha, \rho\rangle}$$

$$\frac{\rho \nvDash A}{\langle \mathsf{if}\ (A)\ \alpha\ \mathsf{else}\ \beta, \rho\rangle \to \langle \beta, \rho\rangle} \qquad \frac{\rho \nvDash A}{\langle \mathsf{while}\ (A)\ \alpha, \rho\rangle \to \langle \mathsf{skip}, \rho\rangle} \qquad \frac{\rho \vDash A}{\langle \mathsf{while}\ (A)\ \alpha, \rho\rangle \to \langle \alpha; \mathsf{while}\ (A)\ \alpha, \rho\rangle}$$

$$\frac{t \geq 0 \quad \hat{x}(0) = \rho \quad \frac{d\hat{x}}{dt}(t) = [\![\mathbf{f}]\!]_{\hat{x}(t)} \quad \rho' = \hat{x}(t) \quad \forall t' \leq t.\, \hat{x}(t') \vDash A}{\langle \mathsf{dwhile}\ (A)\ \{\, \dot{\mathbf{x}} = \mathbf{f}\, \}, \rho\rangle \to \langle \mathsf{dwhile}\ (A)\ \{\, \dot{\mathbf{x}} = \mathbf{f}\, \}, \rho'\rangle}\ (*)$$

$$\frac{t \geq 0 \quad \hat{x}(0) = \rho \quad \frac{d\hat{x}}{dt}(t) = [\![\mathbf{f}]\!]_{\hat{x}(t)} \quad \rho' = \hat{x}(t) \quad \forall t' < t.\, \hat{x}(t') \vDash A \quad \hat{x}(t) \nvDash A}{\langle \mathsf{dwhile}\ (A)\ \{\, \dot{\mathbf{x}} = \mathbf{f}\, \}, \rho\rangle \to \langle \mathsf{skip}, \rho'\rangle}\ (*)$$

Lemma

Reduction is deterministic: if $\langle \alpha, \rho\rangle \to^* \langle \alpha_1, \rho_1\rangle$ *and* $\langle \alpha, \rho\rangle \to^* \langle \alpha_2, \rho_2\rangle$, *then* $\langle \alpha_1, \rho_1\rangle \to^* \langle \alpha_2, \rho_2\rangle$ *or* $\langle \alpha_2, \rho_2\rangle \to^* \langle \alpha_1, \rho_1\rangle$.

The dwhile rule

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

- makes use of: an invariant (for the safety condition), a variant (to leave condition), and a terminator (to prove termination)
- $\mathcal{L}_{\dot{\mathbf{x}}=\mathbf{f}}\, e$: Lie derivative, how much e varies by following $\dot{\mathbf{x}} = \mathbf{f}$

Advantages of this approach

- diversity of driving situations can be treated by different rules
- explainable safety $\rightsquigarrow$ better public trust
- clearly defined rules $\rightsquigarrow$ can be investigated in case of crash
- clearly defined rules $\rightsquigarrow$ car manufacturers can avoid liability