Software and Embedded Systems Engineering

Technische Universität Berlin

Explainable Self-Learning Self-Adaptive Systems

Verena Klös | ES4CPS 2019

Motivation

uncertainties

safety-critical

dynamic networks

autonomous decisions

physical environment

run-time evolution

embedded devices

TRUST ???

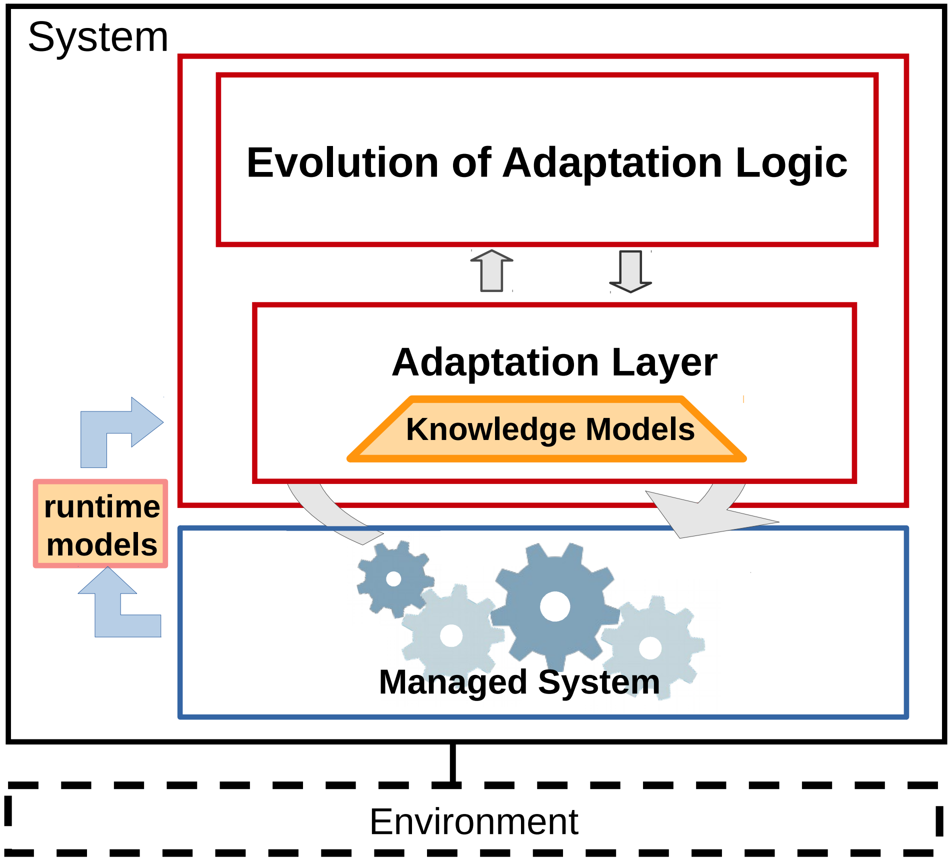
Framework for **Design** and **Verification** of **Self-Learning** Self-Adaptive Systems [SEAA'15, QA4SASO'16, SEAA'17, JSA'18, JSS'18]

Technische Universität Berlin

Main idea:

- run-time **evolution**
- **comprehensibility**
- formal **guarantees**
- **collected data as explanation basis**

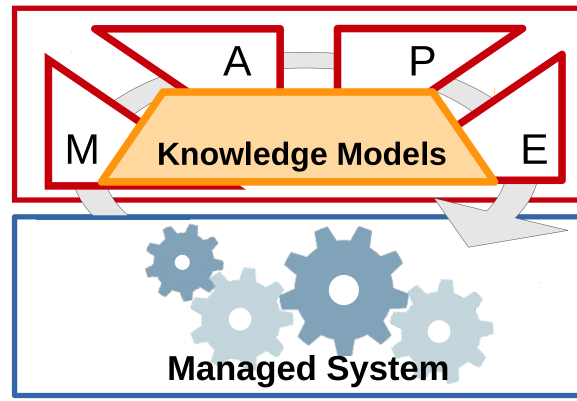
→ **Trust**



SESE

Knowledge Models

- <u>main idea</u>:
  - comprehensible run-time representation of
    - **system** and **environment** state
    - system **goals**
    - **adaptation** options



- <u>knowledge models</u>:
  - K_{Sys}, K_{Env} : current and past system **parameters** & sensor **values**
  - K_{Goal} : hierarchical **quantitative goal model** (distance)
  - K_{Adapt} : **timed adaptation rules**
  - history: enable **retracing** of adaptation decicions & their context

Timed Adaptation Rules

- <u>main idea</u>: efficient & comprehensible model of adaptation options, their **effect** and **timing**

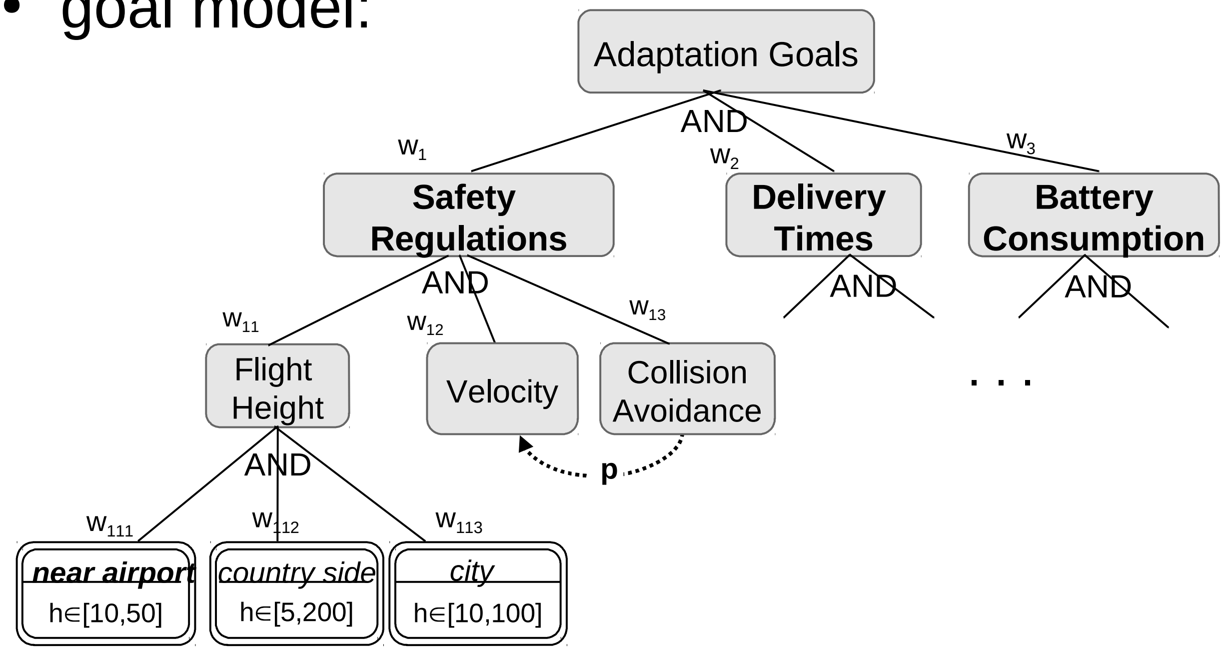
- **timed adaptation rules** as generic mechanism
  - WHEN to apply? *(guard)*
  - WHICH actions? *(commands c_i)*
  - WHAT happens? *(effect)*
  - WHEN? *(time)*

$$r_i : guard_i \,\&\, c_1; c_2; \dots; c_n \rightarrow effect \;\textbf{after}\; time$$

- ➢ formally **encode assumptions** on environment
- ➢ **predictability** & **comprehensibility** of rule-based adaptation

Example – Autonomous Delivery Drones

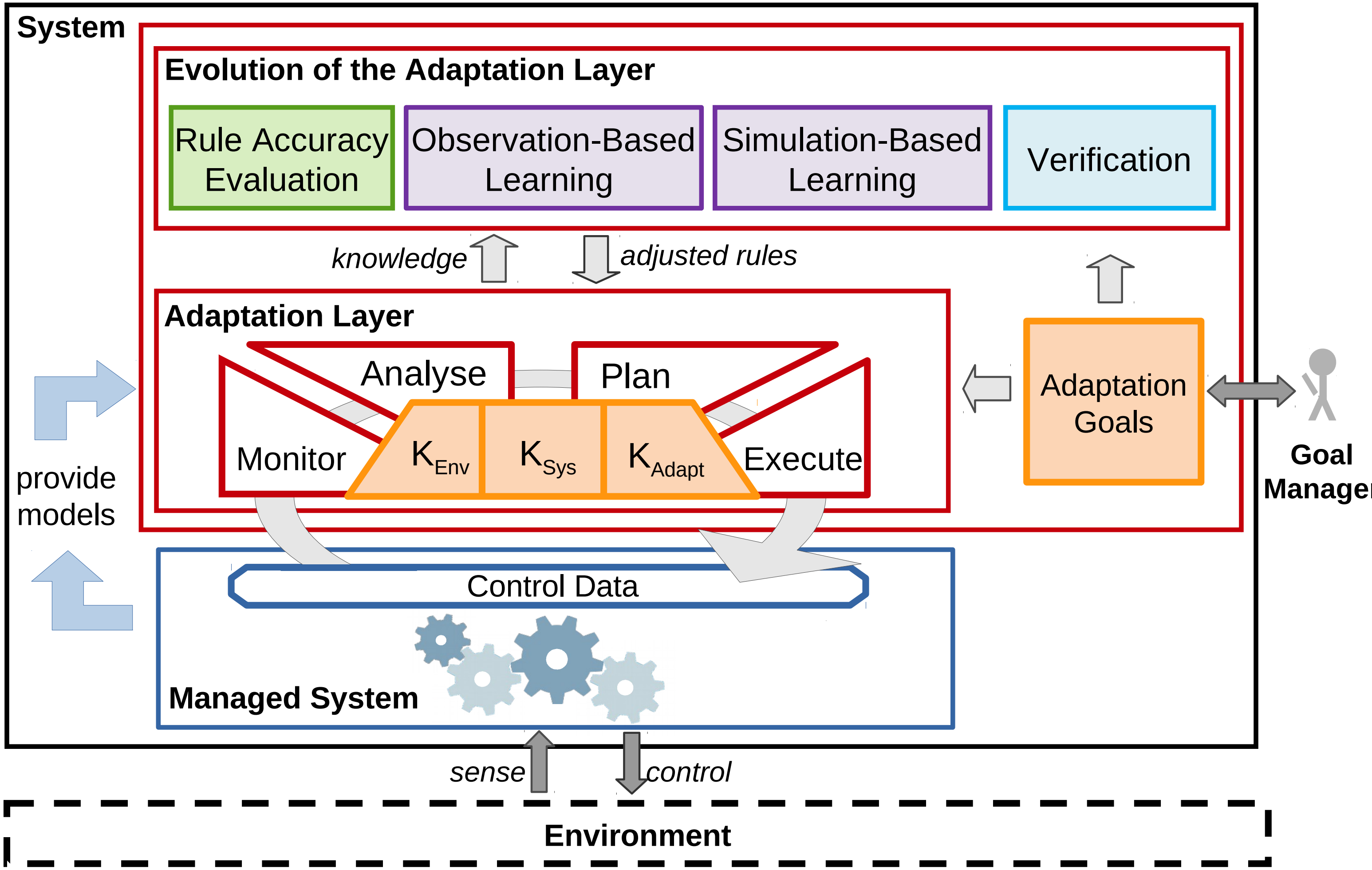
- goal model:



Created by Johndory - Freepik.com

- adaptation rule:
  *remaining_route ≥ battery_range **&&** reduceVelocity()*
  *→ remaining_route < battery_range and estimatedDeliveryTime‘ > estimatedDeliveryTime **after** X seconds*

Evolution of Adaptation Logics



Explainability of **Adaptation** Decisions?

- Example

remaining_route $\geq$ *battery_range* ***&&*** *reduceVelocity()*

$\rightarrow$ *remaining_route* < *battery_range and*

estimatedDeliveryTime‘ > estimatedDeliveryTime

after *X seconds*

- Questions

1) Why was adaptation necessary?

2) Why did you choose THIS option?

3) Which assumptions were made?

4) Which results were achieved?

Explainability of **Adaptation** Decisions?

- Questions

1) Why was adaptation necessary?

2) Why did you choose this option?

3) Which assumptions were made?

4) Which results were achieved?

- Knowledge needed for Answers:

  1) cause (i.e., violated adaptation goals)

  2) context, applicable rules, chosen rule, expected effect/s (in terms of observable changes & goal satisfaction)

  3) chosen rule, expected effect (in terms of observable changes & goal satisfaction)

  4) actual effect

Explainability of **Learning** Decisions?

- Questions

1) Why was learning necessary?

2) How were the results achieved?

3) Which assumptions were made?

4) Which results were achieved?

- Knowledge needed for Answers

  1) original adaptation rule + observed deviations, non-existence of applicable rule

  2) kind of learning, underlying observations or used run-time models, fitness function of the learning algorithm

  3) underlying observations or used run-time models + simulation traces

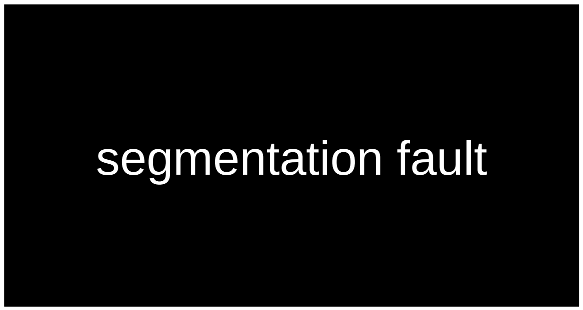
  4) learning result + fitness value

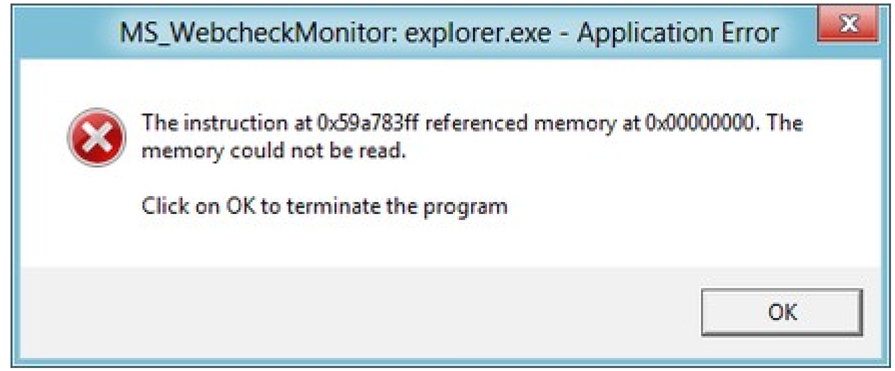
Future Work

- **current** explanations = **all possibly relevant knowledge** for each decision (**explainability = ability to explain**)
  - **only comprehensible for experts**
  - idea: **generate textual explanations** (based on language patterns) for non-experts, **customize explanations** for different target groups (domains, expertise with the system)
- **useful explanations?**

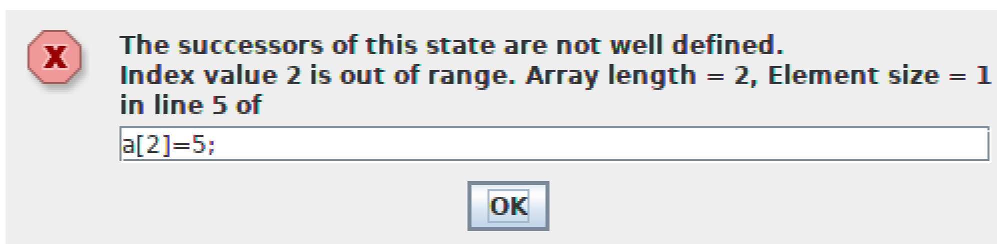
Useful Explanations?

Technische Universität Berlin





vs.



Future Work

- **current** explanations = **all possibly relevant knowledge** for each decision (**explainability = ability to explain**)
  - **only comprehensible for experts**
  - idea: **generate textual explanations** (based on language patterns) for non-experts, **customize explanations** for different target groups (domains, expertise with the system)

- **useful explanations?**
  - provide **filtered information** to **answer questions**
  - cooperate with cognitive science to **investigate human needs and expectations** on explanations
  - use **machine learning on user feedback** to learn characteristics of helpful explanations

References

- [KGG16] Klös, Verena; Göthel, Thomas; Glesner, Sabine: Formal Models for Analysing Dynamic Adaptation Behaviour in Real-Time Systems. In: 3rd Workshop on Quality Assurance for Self-adaptive, Self-organising Systems (QA4SASO). IEEE, pp. 106–111, 2016.
- [KGG18a] V. Klös, T. Göthel, and S. Glesner, “Comprehensible and dependable self-learning self-adaptive systems,” Journal of Systems Architecture, vol. 85-86, pp. 28–42, 2018.
- [KGG18b] V. Klös, T. Göthel, and S. Glesner, “Runtime Management and Quantitative Evaluation of Changing System Goals in Complex Autonomous Systems,” Journal of Systems and Software, vol. 144, pp. 314–327, 2018.